ODA3 INSTITUTE

Practitioner Cheat Sheet

AI Security Evidence Collection Guide

DocID: ODA3-2026-07-CHT-SEC-009

Classification: Public — Practitioner Reference

CHT-SEC-009: AI Security Evidence Collection Guide

ODA3 Institute | Practitioner Cheat Sheet

Field	Detail
Classification	Public — Practitioner Reference
Intended Audience	CISOs, Security Architects, AI Governance Leads, GRC Teams, Incident Responders, Internal Audit
Evidence Confidence	Mixed T1–T4 (see inline tags); no proprietary telemetry used
Review Cycle	Annual, or upon material AI platform or regulatory change
Framework Cross-References	UAIF™ v1.0, AI-IRF™ v1.0, GAISSF™ v1.0
Related Publications	CHT-SEC-001 through CHT-SEC-008

Why This Matters

When something goes wrong in an AI system — a model leaks sensitive data, an agent takes an unauthorized action, a jailbreak succeeds — the first 48 hours often determine whether the incident can be reconstructed at all. AI systems generate evidence across a fragmented stack: model weights and versions, prompt/output pairs, retrieval logs, embeddings, agent tool-calls, and vendor-side telemetry the operating organization frequently doesn't control. Much of this evidence is ephemeral by default; many inference platforms don't retain prompts or outputs unless logging is explicitly configured, and multi-turn or agentic context is often lost entirely once a session ends.

Security teams accustomed to traditional log sources (EDR, SIEM, network flow) are often unprepared for what an AI incident specifically requires: which model checkpoint served the request, whether a RAG pipeline retrieved the document in question, whether a tool-calling agent's action was model-initiated or a downstream failure. Absent deliberate evidence architecture, organizations frequently discover — mid-incident, mid-audit, or mid-assessment — that the data needed to classify severity, attribute root cause, or meet disclosure timelines was never captured, or was overwritten before anyone thought to look.

This cheat sheet lays out what to collect, where it flows from, how to prioritize it, and where it's most often lost — before an incident, audit, or assessment forces the question.

1. What Constitutes AI Security Evidence

AI security evidence is any verifiable information that establishes the security posture, operational behavior, configuration, or historical state of an AI system. It should support at least one of: security assessment, incident investigation, root cause analysis, threat hunting, continuous monitoring, or control validation — not indiscriminate data accumulation for its own sake [T4].

Evidence Category	Examples
Configuration	Model configuration, API settings, security policy state
Operational logs	Inference logs, authentication events, API logs
Model artifacts	Model version, checkpoint hash, signing records
Security events	Alerts, detections, anomalies

Evidence Category	Examples
Human activity	Administrative actions, approvals, human-in-the-loop reviews
Testing results	Red team findings, validation/verification reports [T2]
Simulated results	Controlled attack simulations, academic proxy studies [T3]

2. Evidence Architecture — How Evidence Flows Through an AI Platform

Evidence isn't collected at a single point — it's generated (and lost) at every layer an interaction passes through. Mapping the flow makes it obvious where collection gaps typically sit.

USER
↓
APPLICATION
↓
API GATEWAY → request/response metadata, auth context
↓
IDENTITY LAYER → identity, permission scope, session
↓
PROMPT ASSEMBLY → system prompt, injected context, template version
↓
PROMPT GUARD / SAFETY FILTER → filter decisions, policy version
↓
LLM INFERENCE → model version, checkpoint hash, output, tokens
↓
RAG RETRIEVAL → retrieved doc IDs, ranking scores, KB version
↓
VECTOR DB → index version, upsert/query audit
↓
AGENT / ORCHESTRATOR → tool selection, parameters, approvals
↓
EXTERNAL TOOLS → downstream system logs, execution results
↓
RESPONSE → final output, session close
↓
MONITORING / SIEM → alerts, correlation, anomaly detection
↓
EVIDENCE REPOSITORY → retained, hashed, access-controlled artifacts

Component	Evidence Produced	Retention	Owner	Integrity Protection	Frequency
API Gateway	Request ID, source, auth context, status	Often 30–90 days	Platform/ SecOps	Hash on export	Continuous
Identity Layer	Auth events, privilege scope, MFA/session	Per IAM policy	IAM/SecOps	Append-only where available	Continuous
Prompt Assembly	System prompt, template version, injected context	Often not logged by default	Application team	Requires deliberate config	Per request

Component	Evidence Produced	Retention	Owner	Integrity Protection	Frequency
Prompt Guard	Filter decision, policy version triggered	Often not logged by default	AI Platform team	Requires deliberate config	Per request
LLM Inference	Model/checkpoint ID, prompt hash, output, tokens	Vendor-dependent, often short	AI Platform team	Hash + version pin	Per request
RAG Retrieval	Retrieved doc IDs, scores, KB version	Frequently discarded	Data Engineering	Snapshot at query time	Per request
Vector DB	Upsert/query audit, index version	Often absent by default	Data Engineering	Immutable audit log recommended	Continuous
Agent/ Orchestrator	Tool calls, parameters, approvals	Varies widely	MLOps/AI Platform	Trace ID correlation	Per action
External Tools	Downstream execution logs	Owned by target system	System owner	Standard IT controls	Per call
Monitoring/ SIEM	Alerts, correlation events	Per SIEM policy	SOC	Standard SIEM controls	Continuous
Evidence Repository	Retained artifacts, hashes	Risk-based (Sec. 15)	Named Evidence Owner	Access control + hashing	On collection

This is a reference architecture, not a claim about any specific vendor's default behavior — actual logging availability varies by platform and must be verified per deployment [T4].

3. Principles of Effective Evidence Collection

Evidence should be complete (enough context to reconstruct events), accurate (reflects actual behavior, unaltered), timely (collected as events occur, not reconstructed after the fact), traceable (linked to specific systems, models, users, timestamps), protected (resistant to unauthorized modification or deletion), and relevant (supports a decision, rather than accumulated indiscriminately) [T4].

4. Evidence Quality Dimensions

Beyond source-level confidence tiers (T1–T4), individual artifacts can be evaluated along several quality dimensions. These are ODA3 Institute's own descriptive framing for evidence quality assessment — not a separate certified or trademarked framework.

Dimension	What It Asks
Completeness	Does the artifact contain everything needed to reconstruct the event, or only a fragment?
Fidelity	Is it the raw artifact, or a summary/derivative that may have lost detail?
Coverage	Does collection span all relevant layers (Section 2), or only some?
Freshness	Was it collected close to the event, or reconstructed later from secondary sources?
Continuity	Is there an unbroken sequence, or gaps where collection lapsed?
Provenance	Can its origin — system, collector, method — be established?

Dimension	What It Asks
Sufficiency	Does it, combined with other evidence, actually support the classification being drawn?
Traceability	Can it be linked to a specific identity, system, and timestamp?

5. Evidence Priority Classification

Not all evidence carries equal investigative weight. Classifying evidence by priority helps teams focus collection and review effort where it matters most.

Priority	Definition	Examples
Critical	Without this, incident classification or root cause cannot be established	Model version, prompt, output, RAG retrieval set, tool-call record
High	Materially strengthens investigation but isn't strictly blocking	Identity/auth context, safety filter decisions, session ID
Supporting	Adds context or corroboration	Latency metrics, rate-limit events, non-security config
Contextual	Useful for broader trend analysis, rarely incident-critical	GPU utilization, general infrastructure metrics

6. Minimum Evidence Package

When an assessor, investigator, or auditor asks “what do I need at minimum,” this is the baseline package every AI security assessment or investigation should be able to produce:

Minimum Evidence Package

- ☐ Model version / checkpoint identifier
- ☐ Full prompt (system + injected context + user input)
- ☐ Full output
- ☐ Retrieval IDs (if RAG is in use)
- ☐ Tool-call records (if agentic)
- ☐ Identity of the invoking user/service/agent
- ☐ Timestamp (synchronized)
- ☐ Deployment configuration
- ☐ Safety/guardrail policy version in effect
- ☐ Deployment/build version

If any of these is unavailable for a given system, that gap should be documented explicitly as a known limitation — not silently treated as “not applicable.”

7. How AI Evidence Differs From Traditional IT Forensics

Traditional forensics assumes deterministic, replayable systems. AI systems introduce non-determinism (sampling, temperature), model drift over time, and a dependency chain — model, retrieval, tools, orchestration — where the same input can legitimately produce different outputs on different days if any

layer changed. Evidence collection must capture system state at the time of the event, not just the event itself [T4].

8. Model & Version Evidence

- Exact model identifier and version/checkpoint hash serving traffic at the time in question
- Model card or system card in effect at that time, if one exists
- Deployment configuration (temperature, system prompt, guardrail settings, safety filter version)
- For fine-tuned models: base model version, fine-tuning dataset reference, fine-tuning date

Many managed API providers don't expose checkpoint-level version identifiers to customers by default — a material collection gap discussed in Section 12 [T2].

9. Prompt, Retrieval & Agent Evidence

Prompt & Output

- Full prompt (system prompt, injected context, user input) and full output, timestamped
- Prompt template version and any sanitization/transformation applied
- Token counts and truncation flags
- Session/conversation ID linking multi-turn exchanges

Logging must be configured deliberately — most LLM API providers do not log prompts/outputs on the customer's behalf by default, and many contractually disclaim retention of prompt content [T2]. Sensitive prompts containing regulated or personal information should be handled per organizational privacy requirements — this cheat sheet does not address jurisdiction-specific privacy obligations (see Notably Absent).

Retrieval (RAG)

- Retrieved document IDs and content snapshots — not just the retrieval query — at time of the event
- Embedding model version and vector store index version
- Relevance/ranking scores returned, if available
- Knowledge base version and search parameters

Retrieval evidence is often the single most useful artifact for root-causing hallucination or data-exposure incidents, and is one of the most commonly missing — most RAG deployments log the query and the final answer but discard the intermediate retrieved-chunk set [T3].

Agentic Action

- Full tool-call request/response pairs (which tool, what parameters, what was returned)
- Whether an action executed autonomously or required human approval
- Downstream system logs for anything the agent touched
- Delegation chain and orchestration-layer decisions [T2]

10. Identity, Access & Human Activity Evidence

- Identity and permission scope of whoever/whatever invoked the system (human user, service account, another agent)
- Evidence of privilege boundaries at time of the event vs. what actually occurred
- MFA/session evidence for human-initiated interactions

- Administrative actions, overrides, and approval records — frequently the missing link in reconstructing why a system behaved a certain way, not just how

11. Evidence Across the AI Lifecycle

Lifecycle Phase	Representative Evidence
Design	Threat models, architecture diagrams, risk assessments
Development	Code reviews, security testing, dependency analysis
Training / fine-tuning	Dataset provenance, preprocessing logs, validation metrics
Deployment	Configuration baselines, deployment approvals
Operations	Inference logs, alerts, monitoring telemetry
Incident Response	Timeline, forensic artifacts, containment records
Retirement	Archive records, model disposal/decommissioning evidence

[T4 — lifecycle framing, not a claim of universal practice]

12. Third-Party & Vendor Evidence Gaps

Organizations using third-party foundation models typically cannot access model-internal evidence (attention patterns, training data influence, internal safety filter logic). Standard practice is to negotiate evidence-access and log-retention commitments into vendor contracts before an incident, since post-incident requests are frequently declined or delayed on proprietary-information grounds [T3]. This is a known, structural gap across the industry rather than a failing specific to any one vendor.

13. Evidence Loss Scenarios

Evidence quality doesn't only fail through mishandling — often, it simply never existed. Common loss scenarios worth pre-emptively addressing:

- Prompt/output logging disabled or never configured
- Log rotation or retention expiry before an investigation begins
- Missing correlation/trace IDs across asynchronous boundaries
- Ephemeral containers or serverless instances destroyed before capture
- Session or agent memory discarded on process restart
- Cloud provider default retention window expired
- Agent internal state not persisted between steps

Each of these should be treated as a known architectural risk to close proactively, not discovered mid-incident [T4].

14. Chain of Custody

A defensible chain of custody records, at minimum: evidence identifier, collection timestamp (UTC), collector (person or system), source system, collection method, cryptographic hash where applicable, storage location, and access/transfer history. Where logs are mutable or subject to rolling retention, capture and hash a snapshot immediately upon detection of an incident, assessment trigger, or audit

request. For high-value evidence, protect against unauthorized modification using access controls, cryptographic integrity checks, or write-once storage where practical [T4].

15. Retention & Legal Hold

- Default retention periods for AI platform logs vary widely by vendor and are often shorter than traditional infrastructure logs [T3]
- Legal hold procedures should explicitly name AI-specific log categories (prompts, retrievals, tool-calls) — generic “system logs” hold language may not cover them
- Retention should be risk-based: aligned to investigation, litigation-hold, audit, and regulatory needs. Over-retention introduces its own privacy and security exposure; under-retention hinders investigation
- Coordinate retention requirements with Legal/Compliance before an incident, not during one

16. Evidence Confidence Status

Distinct from source confidence (T1–T4), individual artifacts can also be tagged by their operational state during an active investigation:

Status	Meaning
Verified	Integrity confirmed (hash match, authenticated source)
Partial	Incomplete but usable for some purposes
Corrupted	Integrity uncertain or demonstrably compromised
Missing	Not available at all
Derived	Inferred from other evidence rather than directly captured

This distinction is particularly useful when documenting why a given classification is provisional versus fully supported.

17. Evidence Sufficiency for Classification

Before classifying incident severity or root cause, confirm the evidence available actually supports the classification being made. A common failure mode is classifying an incident based on output alone, without the retrieval or tool-call evidence needed to establish whether the cause was model behavior, a data pipeline defect, or external manipulation (e.g., prompt injection via retrieved content) [T4].

18. Evidence Collection Maturity

Organizations tend to progress through recognizable stages as evidence practices mature. This is offered as descriptive orientation, not a scored or certified maturity model.

Stage	Characteristics
Ad hoc	Manual, inconsistent capture (e.g., screenshots), no defined ownership
Centralized logging	Infrastructure and application logs aggregated, but AI-specific layers largely unlogged

Stage	Characteristics
Correlated AI evidence	Prompt, retrieval, and agent evidence captured and linked via trace/session IDs
Continuous collection	Automated, ongoing capture across the full architecture in Section 2
Assurance-integrated	Evidence collection is designed to directly support assessment and assurance activities, not just incident response

19. Mapping to ODA3 Frameworks

- UAIF™ v1.0 — the evidence categories in Sections 8–10 map directly to UAIF's incident classification inputs; incomplete evidence in a given category constrains which classification tier can be confidently assigned, and this should be documented rather than inferred.
- AI-IRF™ v1.0 — chain-of-custody, evidence confidence status, and quality-dimension practices in Sections 14–16 correspond to AI-IRF's evidentiary-integrity requirements across detection, investigation, containment, and post-incident review.
- GAISSF™ v1.0 — the ability to produce the Minimum Evidence Package (Section 6) and demonstrate maturity progression (Section 18) is one practical input to assurance readiness under GAISSF's Operational and Optimized Certification tiers. Evidence completeness alone does not constitute certification.

19a. External Framework Landscape — Reference, Not Endorsement

Cited for practitioner orientation only. ODA3 Institute does not assert affiliation, certification, or equivalence with any of the following.

Framework	Relevant Evidence-Related Provision
NIST AI Risk Management Framework	Encourages documentation, measurement, and monitoring as part of the Govern/Map/Measure/Manage functions [T4]
ISO/IEC 42001 (AI Management Systems)	Requires documented information, event logging, and operational evidence tied to AI risk treatment [T4]
EU AI Act, Art. 12 (Record-Keeping)	Mandates automatic logging capabilities for high-risk AI systems sufficient to identify risks and substantial modifications [T2]
MITRE ATLAS	Documents adversarial AI tactics/techniques useful for anticipating what evidence a given attack pattern would leave behind
OWASP Top 10 for LLM Applications	Identifies insufficient logging/monitoring of prompts and outputs as a distinct risk category
Cloud Security Alliance (CSA) AI guidance	Publishes practitioner-oriented guidance touching on AI system observability and logging

20. Real-World Incident / Pattern Reference

Illustrative only — public, disclosed patterns, not accusations or technical findings about named parties.

Pattern	Public Reference	Evidence Lesson
Sensitive data pasted into a public LLM chat interface	Samsung employee data-exposure reports, widely covered 2023 [T2]	Input-boundary logging/DLP matters as much as output

Pattern	Public Reference	Evidence Lesson
		monitoring
Chatbot made an unintended policy commitment, later contested	Air Canada chatbot tribunal case, 2024 [T2]	Full prompt/output/session logs were central to establishing what was actually said
Regulator required evidence of data-handling basis for a public LLM service	Italian Garante temporary ChatGPT restriction, 2023 [T2]	Regulatory response required rapid production of logging and data-handling evidence

21. Assessment & Assurance Readiness Checklist

Governance

- ☐ Evidence collection policy documented, with defined ownership
- ☐ Retention requirements established and reviewed with Legal

Architecture & technical coverage

- ☐ Evidence flow mapped against the architecture in Section 2
- ☐ Prompt/output logging enabled and retained in production
- ☐ Model version/checkpoint captured at time of significant interactions
- ☐ RAG retrieval sets (not just final answers) logged
- ☐ Agent tool-calls logged in full, including approvals
- ☐ Identity/access context captured for each invocation

Integrity & classification

- ☐ Timestamp synchronization confirmed across AI-relevant logging sources
- ☐ Evidence protected against unauthorized modification
- ☐ Chain-of-custody procedure documented with a named owner
- ☐ Evidence priority classification (Section 5) applied to key sources

Operational readiness

- ☐ Minimum Evidence Package (Section 6) achievable on demand
- ☐ Known evidence loss scenarios (Section 13) reviewed and mitigated where feasible
- ☐ Vendor contracts reviewed for log-access and retention commitments
- ☐ Legal hold procedures explicitly cover AI-specific log categories
- ☐ Evidence sufficiency reviewed against classification criteria before incident closure

22. Five Priority Actions

1. Map evidence flow against your actual AI architecture and identify which layers in Section 2 aren't currently logging
2. Confirm the Minimum Evidence Package (Section 6) is achievable on demand — not assumed
3. Classify evidence by priority (Section 5) so collection effort matches investigative value
4. Document a chain-of-custody procedure with a named owner, before it's needed

5. Review the evidence loss scenarios in Section 13 against your own environment

Notably Absent

- No proprietary ODA3 telemetry, client incident data, or internal datasets were used in this cheat sheet.
- No claim is made that any named organization or vendor in Section 20 mishandled evidence — references are illustrative of public patterns only.
- No claim is made that following this guide satisfies any specific regulatory requirement (EU AI Act, sectoral regulation, etc.) — organizations should confirm requirements with legal counsel.
- No quantified industry-wide statistics on logging adequacy are asserted, as no verifiable primary dataset supports such a figure at this time.
- The Evidence Quality Dimensions (Section 4) and Evidence Collection Maturity stages (Section 18) are ODA3 Institute's descriptive framing for this guide — not a certified, scored, or externally validated model.
- This guide does not cover jurisdiction-specific privacy or records-management obligations, digital forensic procedures for litigation, or malware analysis methodology — these are separate topics.
- This is not a certification checklist, and completion of the checklist does not constitute GAISSF™ attestation of any kind.

This cheat sheet maps to UAIF™ v1.0, AI-IRF™ v1.0, and GAISSF™ v1.0 under GEL v1.0. It does not constitute certification, compliance confirmation, or endorsement of any organization, product, or vendor.

ODA3 Institute — Where AI Governance meets Operational Reality.