

T01

INTEGRITY

Data poisoning

Attacker manipulates data so the model learns the wrong behaviour.

HOW IT WORKS

Malicious or mislabelled records are inserted into training, fine-tuning or retraining data, or into operational data later used for retraining. Effects range from degraded accuracy to hidden backdoor triggers that activate on a chosen input.

EXAMPLE

Fraud-model retraining pipeline ingests disputed-transaction labels from a customer portal; an attacker files crafted disputes so a pattern of their transactions is learnt as legitimate.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Training, fine-tuning and retraining datasets; data labels; feedback loops	Data acquisition and preparation; continuous or iterative learning in operation	Yes

Q Detect

- 01 Dataset hash and row-count drift against the last approved baseline
- 02 Label distribution and per-source contribution anomalies at each ingest
- 03 Holdout and canary evaluation set regression before promotion
- 04 Spectral signature or activation clustering scans on fine-tuning sets
- 05 Unexplained accuracy shift on a stable slice after a retrain

○ Mitigate

- 01 Provenance and lineage record for every dataset (source, owner, licence, hash)
- 02 Allow-list approved sources; quarantine and review user-contributed data before training
- 03 Signed, versioned datasets; write access to training stores limited to the pipeline identity
- 04 Robust training or data sanitisation for high-impact models
- 05 Promotion gate: no model ships without regression on a protected golden set

IF IT HAPPENS

Roll back to last approved model and dataset version · Identify contaminated records through lineage; retrain from clean snapshot · Record as security incident; review source onboarding

27090 MITIGATION CLAUSES 6.2; 5.3; 5.4; 6.11.4; 7.3; 7.4	ISO/IEC 27001:2022 ANNEX A A.5.9; A.5.21; A.8.25; A.8.29; A.8.32; A.8.15	ISO/IEC 42001:2023 ANNEX A A.7.3; A.7.4; A.7.5; A.7.6; A.6.2.4
EU AI ACT ARTICLE 15 15(5) DP; 15(4)	MITRE ATLAS AML.T0020 Poison Training Data	OWASP LLM TOP 10 2026 LLM05 Data and Model Poisoning

T02

INTEGRITY

Evasion

Crafted inputs make a working model give the wrong answer.

HOW IT WORKS

Inputs are altered, often imperceptibly to humans, so the model misbehaves at inference time. Attacks can be white-box (perfect knowledge) or black-box using queries (zero or partial knowledge).

EXAMPLE

Document classification model used for KYC accepts a forged identity document after small pixel perturbations push it over the 'genuine' threshold.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Model inputs at inference; decision thresholds	Operation (inference)	Yes

Q Detect

- 01 Out-of-distribution scoring on inputs (27090 7.7.1)
- 02 Adversarial input detectors, input transformation consistency checks (27090 7.7.2)
- 03 Confidence-margin monitoring: spikes in near-threshold decisions
- 04 Query pattern analysis for iterative probing from one identity

Shield Mitigate

- 01 Adversarial robustness testing in acceptance against a defined perturbation budget
- 02 Adversarial training or ensembles for high-impact models
- 03 Input normalisation and transformation (resize, re-encode, denoise)
- 04 Human review band for decisions near the threshold
- 05 Do not expose raw confidence scores externally

IF IT HAPPENS

Route flagged inputs to manual review · Add confirmed adversarial samples to robustness test suite · Tighten thresholds or add secondary control while retraining

27090 MITIGATION CLAUSES 6.3; 7.7.1; 7.7.2; 7.8; 5.7	ISO/IEC 27001:2022 ANNEX A A.8.29; A.8.16; A.8.26; A.8.8	ISO/IEC 42001:2023 ANNEX A A.6.2.4; A.6.2.6; A.6.2.2
EU AI ACT ARTICLE 15 15(5) AE; 15(1); 15(4)	MITRE ATLAS AML.T0015 Evade AI Model	OWASP LLM TOP 10 2026 Not an LLM Top 10 category (closest: LLM01 for LLM inputs)

T03

CONFIDENTIALITY

Membership inference

Attacker learns whether a specific record was in the training data.

HOW IT WORKS

By querying the model and analysing outputs or confidence, an attacker determines whether a given data sample was used in training. The disclosure itself can be sensitive (for example, presence in a medical dataset).

EXAMPLE

A health-risk model trained only on patients from a specialist clinic reveals, through confidence differences, that a named individual was a patient.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Training data (via model outputs)	Operation (inference)	Yes

Q Detect

- 01
- Query volume and repetition patterns per identity against baseline
- 02
- Repeated near-duplicate queries around a single record
- 03
- Membership inference testing in pre-release evaluation

🛡 Mitigate

- 01
- Reduce overfitting (regularisation, early stopping, more data)
- 02
- Differential privacy in training where risk warrants
- 03
- Return labels or coarse scores rather than full probability vectors
- 04
- Data minimisation before training (27090 6.11.2)
- 05
- Rate limiting and authenticated access (27090 7.8)

IF IT HAPPENS

Assess as potential personal data breach; involve DPO · Retrain with privacy-preserving technique if testing confirms exposure

27090 MITIGATION CLAUSES 6.4; 6.11.2; 7.8; 7.5	ISO/IEC 27001:2022 ANNEX A A.5.34; A.8.11; A.8.12; A.8.16; A.5.15	ISO/IEC 42001:2023 ANNEX A A.7.4; A.6.2.4; A.5.4; A.7.6
EU AI ACT ARTICLE 15 15(5) CA	MITRE ATLAS AML.T0024.000 Infer Training Data Membership	OWASP LLM TOP 10 2026 LLM02 Sensitive Information Disclosure

T04

CONFIDENTIALITY

Model exfiltration

Attacker rebuilds a working copy of the model through queries.

HOW IT WORKS

Also called model stealing or extraction. The attacker repeatedly queries the system and trains a substitute with similar functionality. It works through legitimate access, so conventional controls alone may not stop it. The copy can then be used offline to craft evasion attacks.

EXAMPLE

A competitor scripts millions of API calls to a pricing model and trains a substitute that reproduces 95% of decisions.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Model functionality and intellectual property	Operation (inference)	Yes

Q Detect	Shield Mitigate
01 Per-key and per-tenant query volume, diversity and coverage metrics	01 Authenticated access with per-identity quotas (27090 7.8)
02 Queries that systematically sweep the input space	02 Limit output detail: no logits, logprobs or full probability vectors unless needed
03 Watermark or fingerprint checks on suspected substitute models	03 Contractual terms prohibiting extraction; API terms enforcement
	04 Model watermarking for high-value models

IF IT HAPPENS

Throttle or revoke offending keys · Preserve logs as evidence (A.5.28); legal action where warranted

27090 MITIGATION CLAUSES 6.5; 7.8; 7.5; 5.3	ISO/IEC 27001:2022 ANNEX A A.5.15; A.5.32; A.8.16; A.8.15; A.8.6	ISO/IEC 42001:2023 ANNEX A A.6.2.6; A.10.4; A.6.2.8
EU AI ACT ARTICLE 15 15(5) CA	MITRE ATLAS AML.T0024.002 Extract AI Model	OWASP LLM TOP 10 2026 LLM06 Unbounded Consumption (covers extraction via repeated querying)

T05

CONFIDENTIALITY

Model inversion

Attacker reconstructs training data from model outputs.

HOW IT WORKS

The attacker supplies inputs and observes outputs to reconstruct the original training data or significant parts of it, such as representative faces or records.

EXAMPLE

Facial recognition API returning confidence scores lets an attacker iteratively reconstruct a recognisable image of an enrolled person.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Training data (via model outputs)	Operation (inference)	Yes

Q Detect

- 01 Optimisation-style query sequences (small successive input changes)
- 02 High query volume against a single class or identity
- 03 Pre-release inversion testing

Shield Mitigate

- 01 Coarsen or round confidence outputs; return top label only
- 02 Differential privacy for sensitive training sets
- 03 Data minimisation and de-identification before training (27090 6.11.2)
- 04 Throttling and authenticated access (27090 7.8)

IF IT HAPPENS

Treat as potential personal data breach · Reduce output granularity immediately; retrain if needed

27090 MITIGATION CLAUSES
6.6; 6.11.2; 7.8; 7.5

EU AI ACT ARTICLE 15
15(5) CA

ISO/IEC 27001:2022 ANNEX A
A.5.34; A.8.11; A.8.12; A.8.16

MITRE ATLAS
AML.T0024.001 Invert AI Model

ISO/IEC 42001:2023 ANNEX A
A.7.4; A.7.6; A.5.4; A.6.2.4

OWASP LLM TOP 10 2026
LLM02 Sensitive Information Disclosure

T06

INTEGRITY

Direct model poisoning

Model weights or pre-trained components are tampered with directly.

HOW IT WORKS

The attacker manipulates the model itself during development or operations: altering weights, inserting a backdoor, swapping a model file, or supplying a malicious pre-trained component or adapter. Unsafe serialisation formats can also execute code on load.

EXAMPLE

A downloaded open-weight model in pickle format runs a reverse shell when loaded in the training environment.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Model parameters; pre-trained components; adapters; model registry	Development, deployment, operation; supply chain	Yes

Q Detect

- 01 Hash or signature verification of every model artefact at load and deploy
- 02 Registry change monitoring and unexpected version promotion alerts
- 03 Static scanning of model files for unsafe serialisation or embedded code
- 04 Behavioural regression on golden set after any artefact change

○ Mitigate

- 01 Signed model artefacts; deploy only from the controlled registry
- 02 Safe formats (for example safetensors) and no pickle loading from untrusted sources
- 03 Third-party models treated as third-party code: provenance, licence, scan, approval
- 04 Segregated, access-controlled development environments (27090 7.6)
- 05 Model bill of materials (AI-BOM) for each production model

IF IT HAPPENS

Revoke and replace artefact from known-good signed version · Treat environment as compromised if unsafe load occurred

27090 MITIGATION CLAUSES 6.7; 5.5; 7.6; 6.11.4; 5.4	ISO/IEC 27001:2022 ANNEX A A.5.21; A.8.7; A.8.25; A.8.31; A.8.32; A.8.4; A.8.24	ISO/IEC 42001:2023 ANNEX A A.10.3; A.4.4; A.6.2.5; A.6.2.3
EU AI ACT ARTICLE 15 15(5) MP	MITRE ATLAS AML.T0018 Manipulate AI Model	OWASP LLM TOP 10 2026 LLM04 Supply Chain

T07

CONFIDENTIALITY

Direct model theft

Model parameters are stolen directly from storage or memory.

HOW IT WORKS

The attacker obtains confidential model parameters (weights) directly, for example from a storage bucket, registry, developer laptop, backup or inference host. 27090 5.1 notes that securing the stack against direct theft is often a more efficient first investment than query-based defences.

EXAMPLE

Model weights left in a publicly readable object storage bucket are copied by a scanner within hours.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Model weights and checkpoints at rest, in transit and in use	All stages	Yes

Q Detect

- 01 Data access logging and alerting on model storage locations
- 02 DLP rules for large weight-file transfers
- 03 Cloud posture monitoring for public or cross-account access

Shield Mitigate

- 01 Encrypt weights at rest and in transit; key access separated from storage access
- 02 Least privilege on registries, buckets and inference hosts
- 03 Confidential computing or TEE for high-value models
- 04 Asset classification of models as confidential information

IF IT HAPPENS

Revoke credentials; contain storage; preserve evidence · Assess downstream risk: assume white-box attacks are now possible

27090 MITIGATION CLAUSES 5.1; 5.3; 7.6	ISO/IEC 27001:2022 ANNEX A A.5.9; A.5.12; A.5.15; A.8.2; A.8.3; A.8.12; A.8.24; A.8.15	ISO/IEC 42001:2023 ANNEX A A.4.4; A.4.5; A.6.2.6
EU AI ACT ARTICLE 15 15(5) CA	MITRE ATLAS AML.T0035 AI Artifact Collection	OWASP LLM TOP 10 2026 LLM04 Supply Chain (partially)

T08

CONFIDENTIALITY

Direct training data leak

Training data is exposed directly from where it is stored or processed.

HOW IT WORKS

Compromise of the confidentiality of training data itself, independent of the model: exposed data lakes, feature stores, labelling vendors, notebooks or exported samples.

EXAMPLE

A labelling contractor's workspace holding raw customer call transcripts is breached.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Training datasets, feature stores, labelling environments, exported samples	Data acquisition, preparation, development	Not necessarily (misconfiguration or vendor error)

🔍 Detect

- 01 Access logging on data lakes and feature stores
- 02 DLP on exports from ML environments
- 03 Supplier security monitoring for labelling and data vendors

🛡 Mitigate

- 01 Classify training data; apply the same controls as the source system
- 02 Minimise, pseudonymise or synthesise data before training (27090 6.11.2)
- 03 Supplier agreements covering labelling and data processing
- 04 Retention and deletion rules for training snapshots

IF IT HAPPENS

Personal data breach procedure; notify within regulatory deadlines · Review whether models trained on leaked data need retraining

27090 MITIGATION CLAUSES 6.11.2; 5.5	ISO/IEC 27001:2022 ANNEX A A.5.12; A.5.19; A.5.20; A.5.34; A.8.10; A.8.11; A.8.12; A.8.33	ISO/IEC 42001:2023 ANNEX A A.7.2; A.7.3; A.7.6; A.10.3
EU AI ACT ARTICLE 15 15(5) CA	MITRE ATLAS AML.T0035 AI Artifact Collection	OWASP LLM TOP 10 2026 LLM02 Sensitive Information Disclosure

T09

CONFIDENTIALITY

Model input leak

Data sent to the model during operation is exposed.

HOW IT WORKS

Compromise of the confidentiality of inputs during operation: prompts and documents captured in logs, telemetry, third-party provider retention, caching, or interception between application and model endpoint.

EXAMPLE

Staff paste customer contracts into a public chatbot whose provider retains and reviews prompts.

ASSET HIT

Prompts, uploaded files, inference requests, logs and caches

LIFE CYCLE STAGE

Operation

ATTACKER REQUIRED

Not necessarily (provider retention, logging, shadow AI)

Q Detect

- 01 Shadow AI discovery: web proxy and CASB for unsanctioned AI domains
- 02 DLP on outbound prompts to AI services
- 03 Review of provider data retention and training settings

Shield Mitigate

- 01 Approved AI tool list with enterprise terms (no training on inputs, defined retention)
- 02 Encrypt in transit; private endpoints for model APIs
- 03 Redact or mask sensitive fields before sending
- 04 Log minimisation: do not log full prompts unless justified and protected

IF IT HAPPENS

Request deletion from provider; assess breach notification · Block unsanctioned service; communicate to staff

27090 MITIGATION CLAUSES
5.3; 6.11.2; 7.5

EU AI ACT ARTICLE 15
15(5) CA

ISO/IEC 27001:2022 ANNEX A
A.5.10; A.5.14; A.5.23; A.8.12;
A.8.11; A.8.23; A.8.15

MITRE ATLAS
AML.T0057 LLM Data Leakage

ISO/IEC 42001:2023 ANNEX A
A.9.2; A.9.4; A.10.3; A.8.2

OWASP LLM TOP 10 2026
LLM02 Sensitive Information Disclosure

T10

CONFIDENTIALITY

Sensitive model output

The model's output reveals training data, prompts or other users' inputs.

HOW IT WORKS

Confidentiality is compromised through what the model says: memorised training data, system prompts or hidden context, another user's prompt or retrieved documents the requester should not see.

EXAMPLE

A support assistant answers with fragments of another customer's ticket because retrieval is not scoped to the requesting customer.

ASSET HIT

Model outputs; context window; retrieved documents; system prompt

LIFE CYCLE STAGE

Operation

ATTACKER REQUIRED

Not necessarily (can occur in normal use)

Q Detect

- 01 Output DLP for PII, secrets and canary strings
- 02 Canary tokens planted in system prompts and restricted documents
- 03 Cross-tenant retrieval audit logging

○ Mitigate

- 01 Enforce the requester's permissions at retrieval time, not in the prompt
- 02 Never place secrets or credentials in prompts or context
- 03 Training data deduplication and filtering to reduce memorisation
- 04 Output filtering and redaction before display

IF IT HAPPENS

Disable the affected retrieval path; assess breach · Rotate any exposed secrets

27090 MITIGATION CLAUSES
6.8; 6.11.2; 6.11.3; 5.3

EU AI ACT ARTICLE 15
15(5) CA

ISO/IEC 27001:2022 ANNEX A
A.5.15; A.8.3; A.8.11; A.8.12;
A.5.34; A.8.26

MITRE ATLAS
AML.T0057 LLM Data Leakage

ISO/IEC 42001:2023 ANNEX A
A.6.2.6; A.7.4; A.6.2.2

OWASP LLM TOP 10 2026
LLM02 Sensitive Information Disclosure

T11

INTEGRITY

Prompt injection

Crafted content is treated as an instruction and hijacks the system.

HOW IT WORKS

Content provided as, or incorporated into, an instruction triggers behaviour the creators did not intend. Direct injection comes from the user; indirect injection arrives inside documents, emails, web pages or tool results the system reads. Impact scales with the permissions of any tools or agents connected.

EXAMPLE

An email assistant reads a message containing hidden instructions to forward the mailbox's last 20 emails to an external address, and its send tool has no approval step.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Instructions and context; connected tools, agents and data	Operation	Yes

Q Detect

- 01 Injection classifiers on inputs and retrieved content (27090 7.7)
- 02 Tool-call logging with anomaly rules (unexpected destinations, volumes, tools)
- 03 Canary instructions in test corpora
- 04 Human approval rejections tracked as signal

Shield Mitigate

- 01 Least model privilege: each agent its own identity and minimum tool scope (27090 6.11.3)
- 02 Human approval for irreversible or external actions
- 03 Separate trusted instructions from untrusted content; mark provenance of content
- 04 Egress allow-list for tools and rendered links
- 05 Assume injection will succeed: design blast-radius limits first

IF IT HAPPENS

Kill switch to disable tools or agent · Review tool-call logs for actions taken; reverse where possible

27090 MITIGATION CLAUSES 6.9; 6.11.3; 5.3; 7.7; 7.5; 5.7	ISO/IEC 27001:2022 ANNEX A A.8.26; A.8.28; A.5.15; A.8.2; A.8.16; A.8.15; A.8.29	ISO/IEC 42001:2023 ANNEX A A.6.2.2; A.6.2.4; A.6.2.6; A.6.2.8; A.9.2
EU AI ACT ARTICLE 15 15(5) para 1; 15(5) AE (interpretation)	MITRE ATLAS AML.T0051 LLM Prompt Injection (.000 Direct, .001 Indirect)	OWASP LLM TOP 10 2026 LLM01 Prompt Injection

INTEGRITY

T12

Output contains injection attacks

Model output carries a payload that a downstream system executes.

HOW IT WORKS

The model's output contains code or markup that harms a downstream consumer: script in a web page, SQL or shell commands, markdown images that call an attacker URL, or malformed data that breaks a parser.

EXAMPLE

A chatbot answer containing a markdown image link renders in the browser and sends conversation data to an attacker's server in the URL.

ASSET HIT	LIFE CYCLE STAGE	ATTACKER REQUIRED
Downstream applications, browsers, databases, interpreters	Operation	Yes (often chained with T11)

Q Detect

- 01
- Output scanning for script, markup and command patterns
- 02
- Content Security Policy violation reports
- 03
- Database and interpreter error and anomaly monitoring

Shield Mitigate

- 01
- Treat model output as untrusted input to every consumer
- 02
- Context-appropriate output encoding; parameterised queries
- 03
- Block or proxy external images and links in rendered output
- 04
- Strict CSP; sandbox any execution of generated code
- 05
- Schema validation for structured outputs

IF IT HAPPENS

Disable rendering path; patch encoding · Assess data exposed through rendered requests

27090 MITIGATION CLAUSES 6.10; 5.4; 7.7	ISO/IEC 27001:2022 ANNEX A A.8.26; A.8.28; A.8.29; A.8.8; A.8.23	ISO/IEC 42001:2023 ANNEX A A.6.2.2; A.6.2.4; A.6.2.5
EU AI ACT ARTICLE 15 15(5) para 1; 15(5) MF (interpretation)	MITRE ATLAS AML.T0051 (delivery)	OWASP LLM TOP 10 2026 LLM10 Improper Output Handling