



PREVENTIVE

AIC-01 Dataset provenance and integrity

Every training, fine-tuning and retrieval dataset has a provenance record (source, owner, licence, date, hash) that is verified before use.

WHY IT MATTERS

Poisoned or unverified data is the cheapest way to change how a model behaves. Provenance makes every training input traceable, so a change to the data cannot pass unnoticed.

OWNER	APPLIES TO	THREATS ADDRESSED
Data owner with the ML engineering lead	Any system trained, fine-tuned or grounded on your data	T01, T08

≡ Implement

- 01 Register each dataset with source, owner, licence, collection date and SHA-256 hash
- 02 Allow-list approved sources; quarantine user-contributed or scraped data for review
- 03 Re-hash datasets at pipeline start and fail the run on any mismatch
- 04 Restrict write access to training stores to the pipeline identity and log every write
- 05 Keep lineage from raw source to the snapshot used for each model version

📄 Evidence an auditor samples

- 01 Dataset register with hashes, owners and licences
- 02 Pipeline logs showing hash verification
- 03 IAM policy for training and retrieval stores
- 04 Lineage record for a sampled model version

TEST IT	MEASURE
Pick one production model version. Trace its training snapshot to source records, re-hash one dataset and compare with the register.	Percentage of production models with complete dataset lineage

27090 SOURCE CLAUSES
6.11.4; 6.2; 5.3

EU AI ACT
15(5) data poisoning; Art 10

ISO/IEC 27001:2022
A.8.25; A.5.21; A.5.9

TOOLKIT REFERENCES
RT-01A, RT-01B; AT-01; Gap G07, G30

ISO/IEC 42001:2023
A.7.5; A.7.3; A.7.4

SOA TREATMENT
Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-02

PREVENTIVE

Model promotion gate

No model is promoted without regression on a protected golden set and sign-off by the model owner.

WHY IT MATTERS

A gate turns "we test our models" into a control that stops a poisoned or tampered model before it reaches users.

OWNER	APPLIES TO	THREATS ADDRESSED
Model owner	Every model you train, fine-tune or retrain	T01, T06

≡ Implement

- 01 Hold a golden evaluation set apart from training data, writable only by the model owner
- 02 Set pass thresholds per metric, including slices for known high-risk inputs
- 03 Run regression, robustness and canary checks automatically in the release pipeline
- 04 Require named sign-off from the model owner before promotion
- 05 Block deployment of any model that has not passed the gate

📄 Evidence an auditor samples

- 01 Gate configuration and thresholds
- 02 Gate results for sampled releases
- 03 Sign-off records with timestamps
- 04 Change tickets linked to each promotion

TEST IT	MEASURE
Sample three promotions. Confirm each has gate results within thresholds and an approval dated before deployment.	Percentage of promotions that passed through the gate

27090 SOURCE CLAUSES
6.2; 7.3; 7.4

EU AI ACT
15(1); 15(5) data poisoning

ISO/IEC 27001:2022
A.8.29; A.8.32

TOOLKIT REFERENCES
AT-01, AT-18; Gap G08

ISO/IEC 42001:2023
A.6.2.4; A.6.2.5

SOA TREATMENT
Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-03

PREVENTIVE

Signed model artefacts

Model artefacts are signed at build and verified at load and deployment. Unsafe serialisation formats are never loaded from untrusted sources.

WHY IT MATTERS

A model file is code. A tampered weight file or a pickle with an embedded payload runs with the privileges of whatever loads it.

OWNER	APPLIES TO	THREATS ADDRESSED
ML platform lead	Any system that hosts or loads model weights	T06

≡ Implement

- 01 Sign models, adapters and tokenisers at build with a key held in a KMS or HSM
- 02 Verify the signature at load in training, evaluation and inference; fail closed
- 03 Use safe serialisation such as safetensors; block pickle loading from untrusted sources
- 04 Scan third-party model files before intake
- 05 Deploy only from the controlled model registry

📄 Evidence an auditor samples

- 01 Signing configuration and key custody record
- 02 Verification logs from load and deploy
- 03 Registry access policy
- 04 Scan results for third-party models

TEST IT	MEASURE
In staging, attempt to deploy an unsigned artefact. Confirm it is rejected and an alert is raised.	Percentage of production artefacts deployed with a verified signature

27090 SOURCE CLAUSES	ISO/IEC 27001:2022	ISO/IEC 42001:2023
6.7; 7.6; 5.5	A.8.24; A.8.7; A.8.32	A.6.2.5; A.4.4
EU AI ACT	TOOLKIT REFERENCES	SOA TREATMENT
15(5) model poisoning	RT-06A, RT-06B; AT-06	Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-04

PREVENTIVE

AI bill of materials

Each production AI system has an AI-BOM listing base models, adapters, datasets and libraries with versions and sources.

WHY IT MATTERS

You cannot respond to a compromised model, dataset or library if you do not know where it is used.

OWNER	APPLIES TO	THREATS ADDRESSED
System owner with procurement	Every production AI system, including third-party models	T06, T01

≡ Implement

- 01 Record base model, version, source, licence and hash for each system
- 02 List adapters, datasets, embedding models and ML libraries with versions
- 03 Generate the AI-BOM in the build pipeline rather than by hand
- 04 Link each entry to its supplier record and vulnerability feed
- 05 Update it on every model or dependency change

📁 Evidence an auditor samples

- 01 AI-BOM for each production system
- 02 Build pipeline step that generates it
- 03 Supplier record links
- 04 Change history

TEST IT	MEASURE
For one system, compare the deployed model hash and library versions with its AI-BOM.	Percentage of production systems with a current AI-BOM

27090 SOURCE CLAUSES
5.5; 6.11.4

EU AI ACT
15(5) model poisoning

ISO/IEC 27001:2022
A.5.21; A.5.9

TOOLKIT REFERENCES
Gap G13; AT-16

ISO/IEC 42001:2023
A.4.4; A.10.3

SOA TREATMENT
Organisation-defined control;
source ISO/IEC 27090 under
ISO/IEC 27001 6.1.3



DETECTIVE

AIC-05 Adversarial robustness testing

High-impact predictive models are tested against a defined perturbation budget before release and on change.

WHY IT MATTERS

Evasion works on models that look accurate in normal testing. Only adversarial testing shows how far an input can be pushed before the decision flips.

OWNER	APPLIES TO	THREATS ADDRESSED
Model owner with the red team lead	Predictive models that make or drive automated decisions	T02

≡ Implement

- 01 Define attacker knowledge, query budget and perturbation budget per model
- 02 Run white-box and black-box evasion attacks during acceptance testing
- 03 Set a maximum attack success rate in proportion to impact
- 04 Retest on every retrain and when new attack techniques are published
- 05 Feed successful adversarial samples back into training or detection

📄 Evidence an auditor samples

- 01 Test plan with budgets
- 02 Robustness test reports
- 03 Remediation records or risk acceptance
- 04 Acceptance sign-off

TEST IT	MEASURE
For a high-impact model, confirm the latest robustness report is under 12 months old and dated after the last retrain.	Attack success rate at the agreed perturbation budget

27090 SOURCE CLAUSES
6.3; 5.7; 7.7.2

EU AI ACT
15(5) adversarial examples; 15(1)

ISO/IEC 27001:2022
A.8.29

TOOLKIT REFERENCES
RT-02A, RT-02B; AT-02

ISO/IEC 42001:2023
A.6.2.4

SOA TREATMENT
Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-06

DETECTIVE

Malicious input detection

Out-of-distribution, adversarial and injection detection is applied to inputs and retrieved content where risk warrants.

WHY IT MATTERS

Detection does not stop every attack, but it turns silent manipulation into an alert someone can act on.

OWNER	APPLIES TO	THREATS ADDRESSED
Application owner with the SOC	External-facing systems and anything that reads untrusted content	T02, T11

≡ Implement	📄 Evidence an auditor samples
01 Score inputs for out-of-distribution signals and route outliers to review	01 Detector configuration
02 Run injection classifiers on user input and on retrieved content before the model sees it	02 Alert samples with dispositions
03 Mark the source of every piece of content placed in the model's context	03 Tuning records
04 Log each detection with the input reference and the action taken	04 Red team detection results
05 Tune thresholds monthly against false positives and red team results	

TEST IT	MEASURE
Replay a known injection payload and an out-of-distribution sample in staging. Confirm both are detected and logged.	Detection rate on red team payloads, against false positive rate

27090 SOURCE CLAUSES	ISO/IEC 27001:2022	ISO/IEC 42001:2023
7.7.1; 7.7.2; 6.9	A.8.16; A.8.26	A.6.2.6
EU AI ACT	TOOLKIT REFERENCES	SOA TREATMENT
15(5) adversarial examples; 15(5) first sentence	RT-11A; Gap G36	Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-07

PREVENTIVE

Output granularity limits

External model interfaces return the minimum output detail required. No logits, logprobs or full probability vectors by default.

WHY IT MATTERS

Fine-grained scores are what make extraction, membership inference and inversion efficient. Removing them raises the attacker's cost sharply.

OWNER	APPLIES TO	THREATS ADDRESSED
API product owner	Any model endpoint reachable by users or customers	T03, T04, T05

≡ Implement

- 01 Return labels or coarse bands by default
- 02 Give consumers that need detailed scores separate, authenticated access
- 03 Round or bucket confidence values wherever they must be shown
- 04 Strip debug fields from production responses
- 05 Review the API specification at every release

📄 Evidence an auditor samples

- 01 API specification
- 02 Sampled production responses
- 03 Access list for detailed scores
- 04 Release review record

TEST IT	MEASURE
Call each external endpoint and inspect the response fields against the specification.	External endpoints exposing fine-grained scores (target zero)

27090 SOURCE CLAUSES
6.4; 6.5; 6.6

EU AI ACT
15(5) confidentiality attacks

ISO/IEC 27001:2022
A.8.11; A.8.12

TOOLKIT REFERENCES
RT-03B; AT-03, AT-04, AT-05

ISO/IEC 42001:2023
A.6.2.2

SOA TREATMENT
Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-08

DETECTIVE

Model use throttling

Every model endpoint has per-identity quotas, rate limits and anomaly alerts.

WHY IT MATTERS

Extraction and privacy attacks need volume. Quotas slow them down and anomaly alerts make them visible.

OWNER	APPLIES TO	THREATS ADDRESSED
Platform engineering with the SOC	Every model endpoint, internal or external	T04, T03, T05

≡ Implement

- 01 Authenticate every endpoint; anonymous access only with a risk acceptance
- 02 Set per-identity rate limits and daily quotas from legitimate peak use
- 03 Alert on sweeps: high volume, high input diversity or systematic coverage from one identity
- 04 Cap tokens and cost per request and per tenant
- 05 Run an exception process for heavy legitimate users

📄 Evidence an auditor samples

- 01 Gateway policy and quota settings
- 02 Anomaly alerts with dispositions
- 03 Exception approvals
- 04 Usage reports

TEST IT	MEASURE
Exceed the quota from a test key. Confirm throttling applies and an alert reaches the SOC.	Time to detect an extraction pattern in exercise

27090 SOURCE CLAUSES
7.8; 6.5

EU AI ACT
15(5) confidentiality attacks

ISO/IEC 27001:2022
A.8.6; A.8.16

TOOLKIT REFERENCES
RT-04A, RT-04B; Gap G37

ISO/IEC 42001:2023
A.4.5; A.6.2.6

SOA TREATMENT
Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-09

PREVENTIVE

Privacy-preserving training

Models trained on personal data are evaluated for membership inference and inversion. Differential privacy or an equivalent is applied where exposure exceeds the threshold.

WHY IT MATTERS

A model can leak the people it was trained on even when the training data itself is locked down.

OWNER	APPLIES TO	THREATS ADDRESSED
Model owner with the DPO	Models trained or fine-tuned on personal or confidential data	T03, T05, T08

≡ Implement

- 01 Minimise or pseudonymise personal data before training
- 02 Deduplicate training data to reduce memorisation
- 03 Run membership inference and inversion tests before release
- 04 Apply differential privacy or stronger regularisation where results exceed the threshold
- 05 Record the privacy versus accuracy decision with the risk owner

📄 Evidence an auditor samples

- 01 DPIA reference
- 02 Data preparation records
- 03 Privacy test reports
- 04 Recorded trade-off decision

TEST IT	MEASURE
For a sampled model trained on personal data, confirm a privacy test exists with a threshold and sign-off.	Membership attack advantage on sensitive models

27090 SOURCE CLAUSES 6.4; 6.6; 6.11.2	ISO/IEC 27001:2022 A.5.34; A.8.11	ISO/IEC 42001:2023 A.7.6; A.5.4
EU AI ACT 15(5) confidentiality attacks; Art 10	TOOLKIT REFERENCES RT-03A, RT-05A; AT-24	SOA TREATMENT Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3



AIC-10

PREVENTIVE

Least model privilege

Each agent or AI service runs under its own identity with the minimum tools, data scope and permissions. Irreversible or external actions need human approval.

WHY IT MATTERS

Prompt injection cannot be filtered out reliably. What you can control is how much damage an injected instruction is able to do.

OWNER	APPLIES TO	THREATS ADDRESSED
System owner with IAM	Any assistant or agent that calls tools or changes data	T11, T12

Implement	Evidence an auditor samples
01 Give each agent or AI service its own identity, never a shared or user token	01 Agent inventory with tools and scopes
02 Grant the minimum tools, data scope and permissions the task needs	02 IAM policies for agent identities
03 Require human approval for irreversible, external or financial actions	03 Approval logs
04 Allow-list egress destinations for tools and rendered links	04 Access review records
05 Review agent permissions on the privileged access review cycle	

TEST IT	MEASURE
In test, attempt one out-of-scope action and one unapproved irreversible action. Both must be blocked and logged.	Percentage of agents with approval gates on irreversible actions

27090 SOURCE CLAUSES
6.11.3; 5.3

EU AI ACT
15(5) first sentence

ISO/IEC 27001:2022
A.5.15; A.8.2; A.5.18

TOOLKIT REFERENCES
RT-11A; AT-11, AT-23

ISO/IEC 42001:2023
A.9.3; A.6.2.2

SOA TREATMENT
Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-11

PREVENTIVE

Untrusted output handling

Model output is encoded, validated or sandboxed for each downstream consumer. External links and images in rendered output are blocked or proxied.

WHY IT MATTERS

If output is trusted, an attacker who can influence the model can reach every system that consumes its output.

OWNER	APPLIES TO	THREATS ADDRESSED
Application owner	Any system whose output is rendered, stored or executed	T12

≡ Implement	📄 Evidence an auditor samples
01 List every consumer of model output: UI, database, interpreter, API, email	01 Data flow diagram showing output consumers
02 Encode output for each context and parameterise every query	02 Secure coding standard section on AI output
03 Block or proxy external images and links; apply a strict Content Security Policy	03 CSP configuration
04 Validate structured output against a schema before use	04 Regression test cases
05 Sandbox any execution of generated code	

TEST IT	MEASURE
In test, induce output containing an XSS payload, a markdown image exfiltration link and SQL. Each consumer must neutralise it.	Percentage of output consumers with documented handling

27090 SOURCE CLAUSES	ISO/IEC 27001:2022	ISO/IEC 42001:2023
6.10	A.8.28; A.8.26	A.6.2.2; A.9.4
EU AI ACT	TOOLKIT REFERENCES	SOA TREATMENT
15(5) first sentence	RT-12A, RT-12B; AT-12	Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-12

PREVENTIVE

Retrieval authorisation

Retrieval enforces the requesting user's permissions at query time. Secrets never appear in prompts or context.

WHY IT MATTERS

Most cross-customer leakage comes from retrieval that can see everything and relies on the prompt to behave.

OWNER	APPLIES TO	THREATS ADDRESSED
Application owner with the data owner	Systems that use RAG, search or shared context	T10

≡ Implement

- 01 Enforce the requester's permissions in the retrieval query, at the data layer
- 02 Tag each indexed document with its access control list at ingestion
- 03 Separate tenants in vector stores by index or an enforced filter
- 04 Keep secrets and credentials out of prompts and context
- 05 Plant canary records and alert if they surface in another tenant's output

📄 Evidence an auditor samples

- 01 Retrieval access design
- 02 ACL tagging at ingestion
- 03 Cross-tenant test transcripts
- 04 Canary alert configuration

TEST IT	MEASURE
From two tenants, query for content that exists only in the other. Nothing must come back.	Cross-tenant test failures (target zero)

27090 SOURCE CLAUSES
6.8; 5.3

EU AI ACT
15(5) confidentiality attacks

ISO/IEC 27001:2022
A.8.3; A.5.15

TOOLKIT REFERENCES
RT-10A, RT-10B; AT-10

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

SOA TREATMENT
Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-13

DETECTIVE

AI red teaming

AI systems are red teamed before launch and on significant change, with findings tracked to closure.

WHY IT MATTERS

Infrastructure penetration tests rarely test the model itself. Red teaming is the only control here that tells you whether the others actually work.

OWNER	APPLIES TO	THREATS ADDRESSED
Red team lead	High-impact, external and agentic AI systems	All

≡ Implement

- 01 Agree rules of engagement: scope, environments, stop conditions, data handling
- 02 Build test cases from 27090 clause 6 and MITRE ATLAS for each applicable threat
- 03 Test before launch, on significant change and at least yearly for external systems
- 04 Track findings in vulnerability management with owners and dates
- 05 Retest fixes and record closure

📄 Evidence an auditor samples

- 01 Rules of engagement
- 02 Red team report
- 03 Finding tracker
- 04 Retest evidence

TEST IT	MEASURE
Confirm each high-impact system was red teamed in the last 12 months and every finding is closed or risk-accepted.	Percentage of findings closed within SLA

27090 SOURCE CLAUSES
5.7

EU AI ACT
15(5), all elements

ISO/IEC 27001:2022
A.8.29; A.8.8

TOOLKIT REFERENCES
Red_Team_Plan; AT-17; Gap G15

ISO/IEC 42001:2023
A.6.2.4

SOA TREATMENT
Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-14

GOVERNANCE

Mitigation conflict review

Trade-offs between mitigations, such as privacy against accuracy or logging against minimisation, are assessed and recorded.

WHY IT MATTERS

27090 clause 7.2 warns that one mitigation can weaken another. Unrecorded trade-offs become findings at audit and gaps in practice.

OWNER	APPLIES TO	THREATS ADDRESSED
Risk owner	Every high-impact AI system	All

≡ Implement

- 01 List the mitigations applied to each high-impact system
- 02 Identify pairs that work against each other
- 03 Quantify the trade-off where possible
- 04 Record the decision with the risk owner's approval
- 05 Revisit when the model, the data or the threat changes

📄 Evidence an auditor samples

- 01 Trade-off decision records
- 02 Risk owner approvals
- 03 Review dates

TEST IT	MEASURE
For a sampled high-impact system, ask for its recorded trade-off decisions and check the approver.	Percentage of high-impact systems with recorded decisions

27090 SOURCE CLAUSES	ISO/IEC 27001:2022	ISO/IEC 42001:2023
7.2	Clause 6.1.3; A.8.27	Clause 6.1.3
EU AI ACT	TOOLKIT REFERENCES	SOA TREATMENT
Art 9 risk management	AT-25; Gap G31	Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3

AIC-15

DETECTIVE

Deterioration re-validation

Models and their mitigations are re-validated on a schedule and when drift or new attack techniques are identified.

WHY IT MATTERS

Models drift and defences age. A control that passed at launch can quietly stop working months later.

OWNER	APPLIES TO	THREATS ADDRESSED
Model owner	Every production model, including third-party models that change version	All

≡ Implement

- 01 Monitor accuracy, drift and input distribution against the release baseline
- 02 Set thresholds that trigger re-validation of the model and its defences
- 03 Re-validate at least yearly even without a trigger
- 04 Re-run red team cases after major provider model updates
- 05 Record the outcome and any action taken

📄 Evidence an auditor samples

- 01 Monitoring configuration and thresholds
- 02 Drift alerts
- 03 Re-validation reports
- 04 Provider change notices

TEST IT	MEASURE
Sample one drift alert and trace it to a re-validation record.	Percentage of models re-validated within schedule

27090 SOURCE CLAUSES	ISO/IEC 27001:2022	ISO/IEC 42001:2023
7.4; 7.3	A.8.16; A.8.32	A.6.2.6
EU AI ACT	TOOLKIT REFERENCES	SOA TREATMENT
15(1); Art 72	AT-19; Gap G33	Organisation-defined control; source ISO/IEC 27090 under ISO/IEC 27001 6.1.3



AIC-16

CORRECTIVE

AI kill switch

Each AI system with actions or external exposure has a documented, tested way to disable its tools, its retrieval or the model itself.

WHY IT MATTERS

When an agent is misbehaving, minutes matter. A switch that nobody has tested will not work when it is needed.

OWNER	APPLIES TO	THREATS ADDRESSED
Service owner with on-call	Agents and external-facing AI systems	T11, T12

≡ Implement

- 01 Document how to disable tools, retrieval or the whole model for each system
- 02 Give on-call staff the authority and access to use it without waiting for approval
- 03 Provide a safe fallback such as a human queue or a static response
- 04 Test it at least twice a year
- 05 Link it from the AI incident playbooks

📄 Evidence an auditor samples

- 01 Kill switch runbook
- 02 Test records with timings
- 03 Incident playbook references

TEST IT	MEASURE
Observe a kill switch test, or review the last one, and record the time taken to disable.	Time to disable the system in test

27090 SOURCE CLAUSES
6.11.3

EU AI ACT
15(5): respond, resolve

ISO/IEC 27001:2022
A.5.26; A.5.24

TOOLKIT REFERENCES
AT-21; Gap G38

ISO/IEC 42001:2023
A.6.2.6; A.8.4

SOA TREATMENT
Organisation-defined control;
source ISO/IEC 27090 under
ISO/IEC 27001 6.1.3