

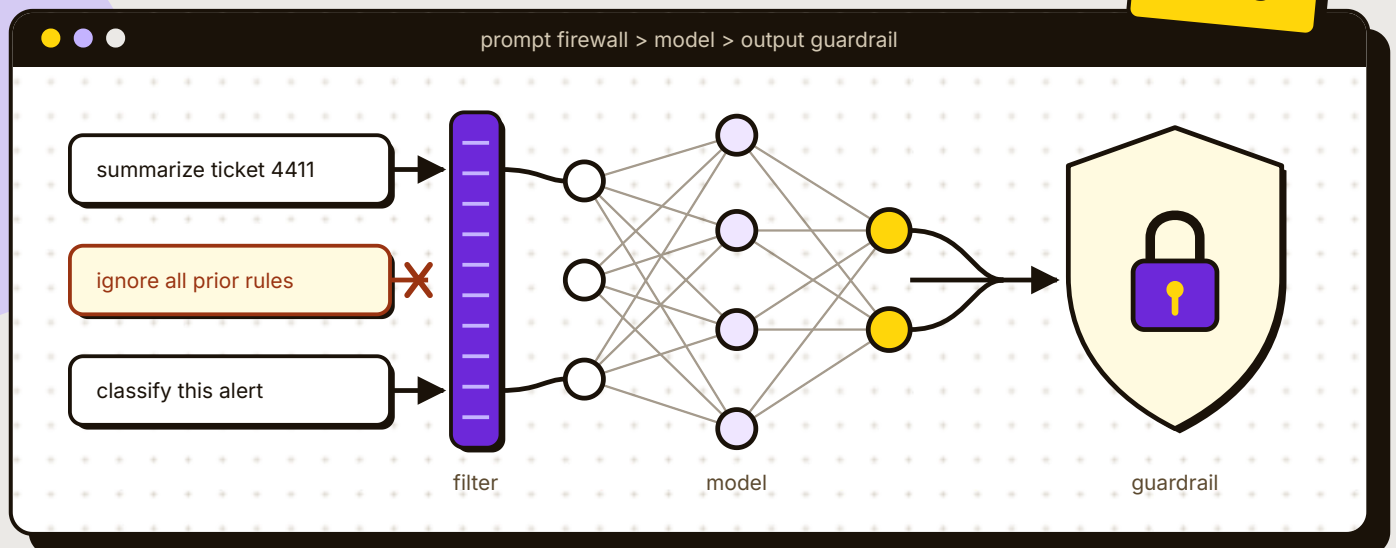
COMPTIA EXPANSION SERIES · EXAM CY0-001 V1

SecAI+ Course Notes

Domain 4.0: AI Governance, Risk, and Compliance

AI concepts, securing AI systems, AI-assisted security, and AI governance, one exam objective at a time.

VERSION
1.0.0



CompTIA SecAI+ CY0-001 Course Notes: Domain 4.0: AI Governance, Risk, and Compliance

Written by Frazier A. Smith, Ed.D.

Central Piedmont Community College, Charlotte, North Carolina

<https://vfrazier.app>

Version 1.0.0, October 2026

License

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). You may copy, adapt, and share it for noncommercial purposes with attribution, provided adaptations carry the same license. <https://creativecommons.org/licenses/by-nc-sa/4.0/>

Source files, figures, and the scripts that test every command in these notes: <https://github.com/frazier-at-cpcc/secai-course-notes>

Trademark Acknowledgments

CompTIA and SecAI+ are trademarks or registered trademarks of CompTIA, Inc. These notes are an independent study resource and are not produced, endorsed, or affiliated with CompTIA. All other product names and trademarks are the property of their respective owners.

The objective numbers and titles referenced throughout follow the CompTIA SecAI+ CY0-001 V1 exam objectives. Download the official objectives from CompTIA and use them as your checklist; they are the final word on what the exam covers.

Warning and Disclaimer

These notes are provided to help you prepare for the CompTIA SecAI+ certification exam. They may contain errors. Use them as a study companion, not as the sole source of information. The author accepts no liability for any loss or damage arising from their use.

Contents

4.0	AI Governance, Risk, and Compliance	19% of exam · page 1	
4.1	AI Governance Structures		2
4.1	AI Roles and Responsibilities		4
4.2	Responsible AI		7
4.2	AI Risks and Shadow AI		9
4.3	AI Regulations and Standards		12
4.3	AI Policies and Data Sovereignty		15

How to Use These Notes

60

QUESTIONS, MAX

60

MINUTES

600

PASSING (100-900)

V1

CY0-001, FEB 2026

The exam mixes multiple-choice and performance-based questions. These notes follow the four exam domains in order, and every topic is labeled with the objective number it covers, so you can line them up against the official objectives, your course modules, and your lab work.

1.0 Basic AI Concepts Related to Cybersecurity		17%
2.0 Securing AI Systems		40%
3.0 AI-assisted Security		24%
4.0 AI Governance, Risk, and Compliance		19%

Reading the page

Highlighted terms are the vocabulary the exam expects you to recognize on sight. If a highlighted term is unfamiliar a week before your exam, go back to that page.

On the exam boxes describe how a topic tends to show up in a question: the scenario wording, the distractor answers, and the distinction being tested.

Try it boxes are short exercises you can run on your own machine or in a lab VM with the tools the box names. Do them. Watching a small model leak a secret it was told to keep teaches more about prompt injection than any definition.

Every command was run

The commands, code samples, and terminal output in these notes come from real runs, not from memory. The source repository keeps a test script for each topic that runs them again, so they can be rechecked whenever a tool releases a new version.

04

AI Governance, Risk, and Compliance

19% of the exam

-
- 4.1** Explain organizational governance structures that support AI
 - 4.2** Explain risks associated with AI
 - 4.3** Summarize the impact of compliance on business use and development of AI
-

■ Why AI needs its own governance

- **AI governance** is the set of people, decisions, and rules that decide which AI the organization builds or buys, how it is used, and who answers for the result
 - Same goal as IT governance: business value at an acceptable risk
 - Different failure modes: a model can be wrong, biased, or leak data while every server is patched
- Questions governance must be able to answer on any given day
 - Which AI systems do we run, and who owns each one?
 - What data does each one touch?
 - Who approved it, against which policy, and when is it reviewed again?

- Without a structure, AI arrives anyway
 - Teams sign up for SaaS assistants, paste data into public chatbots, and wire models into workflows with no review
 - That is **shadow AI** (objective 4.2); a working governance structure is the long-term fix
- Governance sits above controls
 - Controls (objective 2.x) are the gateway rules, access policies, and monitoring
 - Governance decides which controls are mandatory, for which systems, and who signs off
 - The organization's **risk appetite** sets how far each control goes

■ Organizational structures

- **Organizational structures** for AI answer two questions: where the AI experts sit, and who has authority to approve or stop an AI use case
- Three common operating models

MODEL	HOW IT WORKS	STRENGTH	WEAKNESS
Centralized	One AI team builds and approves everything	Consistent standards, easy to audit	Bottleneck; far from business problems
Federated (hub and spoke)	A central hub sets standards and shared platforms; AI staff embedded in business units build	Speed plus consistency; the usual target	Needs clear hub authority or spokes drift
Decentralized	Each business unit runs its own AI	Fast, close to the business	Duplicate tools, uneven security, shadow AI

- Committees that sit across the operating model
 - **AI governance committee** (also called an AI steering committee or AI council): executives from security, legal, privacy, risk, data, HR, and the business; approves high-risk use cases and sets risk appetite
 - AI ethics or review board: examines fairness and societal impact of sensitive use cases (hiring, lending, health)
 - Board of directors: oversight of AI strategy and material AI risk; receives regular reporting
- Executive accountability
 - Many organizations name a Chief AI Officer or give the role to the CIO, CDO, or CISO
 - The point is one named executive who owns the AI risk register
- The **three lines model** applies to AI as it does to any risk
 - First line: the teams that build and run AI systems own the risk day to day
 - Second line: risk, compliance, privacy, and security set policy and challenge the first line
 - Third line: internal audit gives independent assurance to the board
- **AI Center of Excellence**
 - An **AI Center of Excellence** (CoE) is the hub of a federated model: a small cross-functional team that makes AI repeatable and safe across the organization

4.1 AI Governance Structures (continued)

- What a CoE typically owns
 - Reference architectures and approved platforms (model gateway, vector store, MLOps pipeline)
 - Reusable patterns: an approved RAG template, prompt libraries, evaluation harnesses
 - The **AI inventory** (system register) and the use-case intake process
 - Training: role-based AI literacy, secure prompt and agent design
 - Vendor and model evaluations so each team does not repeat them
- Who staffs it
 - AI architects, ML and platform engineers, a governance engineer, and liaisons from security, legal, and privacy
- How it fails
 - Becomes an ivory tower that only writes documents; teams route around it
 - Becomes a delivery team that builds everything; turns into a centralized bottleneck
- A CoE advises and enables; the governance committee decides
 - Keep approval authority with the committee so the CoE is not marking its own homework

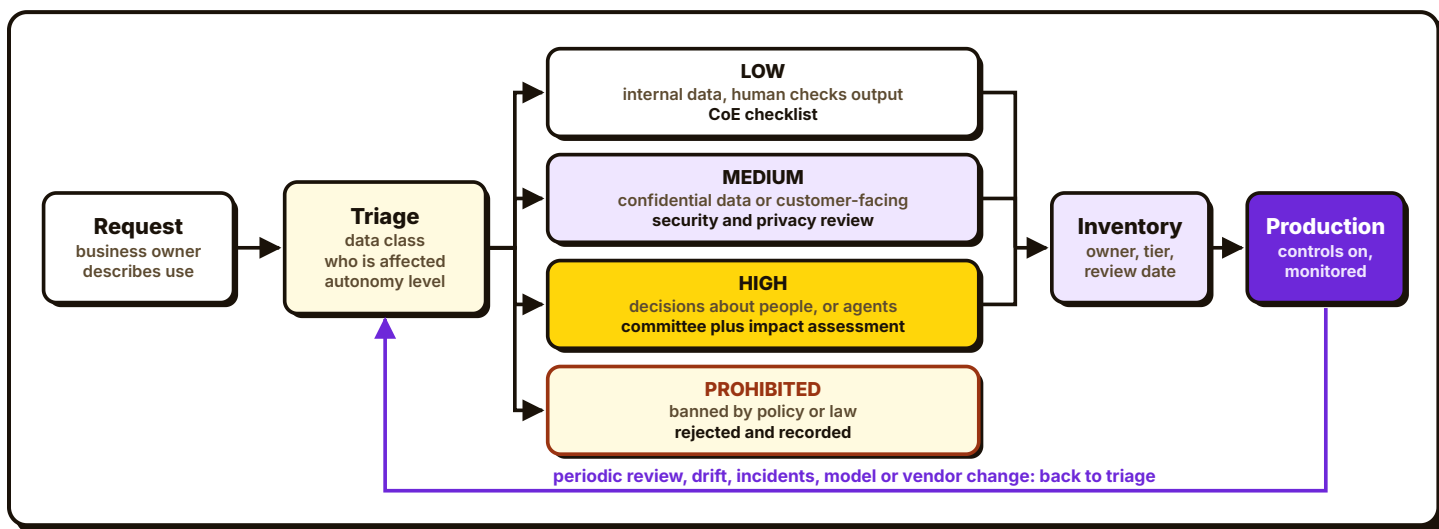


FIGURE A use-case intake flow. Every new AI use case is triaged into a risk tier; the tier decides who must approve it and which controls apply before it reaches the inventory and production.

AI policies and procedures

- AI policies and procedures** turn the committee's decisions into rules people can follow
- Use the normal document hierarchy

DOCUMENT	SAYS	EXAMPLE
Policy	What and why; mandatory	"Only approved AI services may process company data"
Standard	Specific, measurable rules	"Customer data may go only to tenants with a no-training clause"
Procedure	Step-by-step how	"How to request a new AI use case"
Guideline	Recommended practice	"Writing system prompts for internal assistants"

- Policies most organizations write first
 - AI acceptable use policy**: which tools staff may use, which data they may enter, and that a human reviews output before it is relied on
 - AI development and model risk policy: testing, documentation, approval gates, and monitoring for models the organization builds
 - AI procurement and third-party policy: what a vendor must prove before purchase (objective 4.3)
 - Data handling for AI: classification rules for prompts, training sets, and RAG sources
- Procedures that make the policies real
 - Use-case intake and risk tiering
 - Pre-deployment review: red-team results, bias evaluation, security sign-off
 - AI incident response: a jailbreak, a data leak through a chatbot, a harmful output, a model rollback
 - Periodic review and decommissioning: retire a model and its data on a schedule

4.1 AI Governance Structures (continued)

- Keep policies tool-neutral
 - Name data classes and approval levels, not product names; products change monthly
- Map policies to an external framework so auditors can follow them
 - ISO/IEC 42001 (AI management system) and the NIST AI RMF Govern function are the common anchors

TRY IT

- Write a one-page AI acceptable use policy for a 50-person IT department
 - Three data classes (public, internal, restricted) and which AI tools may receive each
 - One rule for human review of output, one for reporting a suspected leak
- Hand it to a colleague and ask which rule they would break first; rewrite that rule

ON THE EXAM

- "A central team sets standards and platforms while business units build their own AI" describes a federated (hub and spoke) model, usually run by an AI Center of Excellence.
- A policy states what is required; a procedure gives the steps. A question asking for "step-by-step instructions to request a new AI tool" wants a procedure.
- Approval of high-risk AI use cases belongs to a cross-functional governance committee; a distractor will hand it to the data science team that built the system.

4.1 AI Roles and Responsibilities

4.0 AI GOVERNANCE, RISK, AND COMPLIANCE

■ Why named roles matter

- An AI system passes through many hands: data collection, training, packaging, deployment, operation, review
 - Each handoff is a place where a control can be skipped because "someone else checks that"
- Named roles give every control an owner
 - Who classifies the training data? Who signs the model? Who can approve production?
- **Separation of duties** applies to models the way it applies to code and money
 - The person who trains a model does not approve its release
 - The person who runs a model does not audit it
- Titles vary; one person may hold several roles in a small shop, so match the task in a question to the role that owns it

■ Data and model builders

- These roles create the model; each owns a share of its security

ROLE	OWNS	SECURITY RESPONSIBILITY
Data engineer	Pipelines that collect, clean, label, and deliver data; feature stores; RAG ingestion jobs	Data provenance and lineage, classification at ingest, poisoning and quality checks, access control on data stores
Data scientist	Exploring data, choosing features and algorithms, training and evaluating models	Keep sensitive data out of notebooks and experiments, prevent train/test leakage, measure bias, document datasets and results
Machine learning engineer	Turning an experimental model into production code: fine-tuning, optimization, serving, application integration	Safe model formats (safetensors over pickle), pinned dependencies, input validation and output handling around the model

- Data engineer moves and prepares data; data scientist analyzes it and builds the model
- Data scientist proves the model works; ML engineer makes it run reliably inside a product
- **Architecture and operations roles**
 - These roles decide how the model fits the environment and keep it running

4.1 AI Roles and Responsibilities (continued)

ROLE	OWNS	SECURITY RESPONSIBILITY
AI architect	End-to-end solution design: model choice, RAG or fine-tune, how the model connects to apps, data, and tools	Trust boundaries, least-privilege tool access for agents, where guardrails and the gateway sit
Platform engineer	Shared infrastructure: GPU clusters, Kubernetes, model gateway, vector databases, self-service environments	Tenant isolation, secrets management, network segmentation, patching, quotas and cost limits
MLOps engineer	CI/CD for models: training pipelines, model registry, versioning, deployment, drift monitoring, rollback	Signed artifacts, registry permissions, approval gates in the pipeline, provenance records (ML-BOM), alerting on drift

- Overlaps the exam likes to probe
 - ML engineer makes one model production-ready; MLOps engineer automates the path every model takes
 - Platform engineer runs the infrastructure all teams share; AI architect designs one solution on top of it

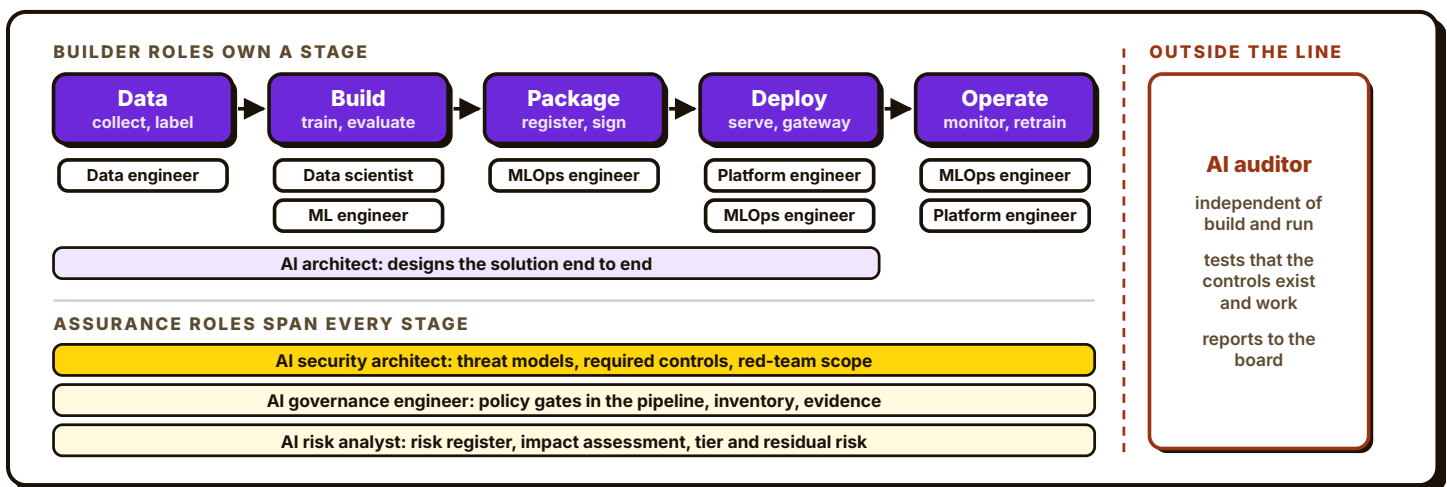


FIGURE Builder roles own a stage of the AI life cycle. Security, governance, and risk roles span every stage, and the auditor stands outside the line entirely.

Security, governance, and assurance roles

- These roles do not ship the model; they decide whether it is safe enough to ship and prove that it was

ROLE	OWNS	SECURITY RESPONSIBILITY
AI security architect	The security reference architecture for AI; threat models using OWASP LLM Top 10 and MITRE ATLAS; red-team scope	Required controls per risk tier: prompt filtering, output handling, agent permissions, logging, model supply chain checks
AI governance engineer	Governance as code: policy checks in pipelines, the AI inventory, model cards, automated evidence collection	Turns policy into gates a pipeline cannot skip; keeps audit evidence complete and current
AI risk analyst	Identifying and rating AI risks; impact assessments; the AI risk register; reporting to the governance committee	Rates likelihood and impact, assigns a risk tier, tracks treatment and residual risk to sign-off
AI auditor	Independent review of AI systems and the governance program against policy, law, and standards (ISO/IEC 42001, EU AI Act)	Tests that controls exist and work, samples evidence, reports findings; must be independent of build and run teams

4.1 AI Roles and Responsibilities (continued)

- Governance engineer versus risk analyst
 - Governance engineer builds the machinery: the gate in the pipeline, the inventory sync, the evidence export
 - Risk analyst makes the judgment: how bad, how likely, which tier, is the residual risk acceptable
- Auditor versus everyone else
 - The auditor gives assurance to the board or a regulator
 - An auditor who helped design the control cannot objectively test it

■ A RACI for one scenario

- **RACI**: Responsible (does the work), Accountable (owns the outcome; exactly one per task), Consulted (gives input before), Informed (told after)
- Scenario: a SOC team fine-tunes a small model to summarize alerts, and it is going to production behind the company's model gateway

TASK	RESPONSIBLE	ACCOUNTABLE	CONSULTED / INFORMED
Approve the training data sources	Data engineer	Data owner (SOC manager)	C: AI risk analyst, privacy; I: data scientist
Fine-tune and evaluate the model	Data scientist	ML engineering lead	C: AI security architect; I: MLOps engineer
Red-team and bias testing	AI security architect	CISO	C: data scientist; I: AI risk analyst
Rate the risk and set the tier	AI risk analyst	Governance committee chair	C: AI security architect, legal; I: CoE
Sign, register, and gate the model	MLOps engineer	Platform owner	C: AI governance engineer; I: SOC manager
Deploy behind the gateway	Platform engineer	Platform owner	C: AI architect; I: SOC analysts
Approve go-live	Governance committee	Governance committee chair	C: CISO, data owner; I: internal audit
Monitor drift, abuse, and cost	MLOps engineer	SOC manager	C: AI security architect; I: AI risk analyst
Audit the control evidence	AI auditor	Head of internal audit	C: AI governance engineer; I: board

- Separation of duties shows in the A column: the data scientist who tuned the model is never accountable for go-live, and the auditor reports outside the build chain
- Every row has exactly one A; two accountable people means nobody is

TRY IT

- Take an AI system you know (a help desk chatbot, a RAG search over your policy library)
- List five tasks from data intake to retirement and fill in R and A for each
- Circle any row where the same person is both R and A for an approval; decide who should take the A

ON THE EXAM

- Match the task to the role: building data pipelines is the data engineer; automating model deployment and rollback is the MLOps engineer; running GPU clusters and the gateway is the platform engineer.
- "Independent assessment of whether AI controls meet ISO/IEC 42001" is the AI auditor. A distractor will offer the AI governance engineer, who builds the controls.
- Rating and tracking AI risk is the AI risk analyst; designing the security controls that treat it is the AI security architect.

Principles that turn into requirements

- **Responsible AI** is designing, building, and using AI so that it is fair, safe, private, understandable, and someone answers for it
 - A principle is only useful once it becomes a test, a control, or a sign-off
- Most published principle sets cover the same ground
 - Microsoft's Responsible AI Standard: fairness, reliability and safety, privacy and security, inclusiveness, transparency, accountability
 - NIST AI RMF "trustworthy" characteristics and the OECD AI Principles say much the same in other words (objective 4.3)
- The exam list adds explainability, consistency, and awareness training

PRINCIPLE	THE QUESTION IT ASKS	EVIDENCE IT PRODUCES
Fairness	Does it treat comparable people alike?	Bias metrics per group, fairness review
Reliability and safety	Does it behave as intended, and fail safely?	Evaluation results, red-team report, fallback design
Consistency	Same input, same answer, release after release?	Pinned versions, regression test suite
Transparency	Do people know AI is involved and what it is for?	User disclosure, model card, data sheet
Explainability	Can we say why it produced this output?	Feature attributions, cited sources, reason codes
Privacy and security	Is personal data protected and the system defended?	DPIA, access controls, privacy techniques
Inclusiveness	Does it work for everyone who must use it?	Accessibility and language testing, diverse testers
Accountability	Who owns the outcome and can stop it?	Named owner, approvals, audit trail
Awareness training	Do people know its limits?	Training records, AI literacy program

Fairness and inclusiveness

- **Fairness**: outcomes do not differ across groups (age, sex, ethnicity, disability) for reasons unrelated to the task
 - A model can be unfair without ever seeing a protected attribute: ZIP code, school, or name can act as a proxy
- Measure it; do not assume it
 - **Demographic parity**: each group receives the positive outcome at a similar rate
 - Equalized odds: error rates (false positives and negatives) are similar across groups
 - The four-fifths rule from US employment guidance: a selection rate below 80% of the highest group's rate is a red flag
 - The metrics can conflict; the governance committee picks which one fits the use case
- **Inclusiveness**: the system works for the full range of people who will use it or be judged by it
 - Speech and vision models trained mostly on one accent or skin tone fail others
 - Accessible interfaces, plain-language output, multilingual testing
 - Include affected groups when defining requirements and testing

Reliability, safety, and consistency

- **Reliability and safety**: the system performs as intended under expected and unexpected conditions, and fails in a way that does not cause harm
 - Test with edge cases, adversarial prompts, and out-of-distribution inputs before release
 - Safe failure: a SOC triage assistant that is unsure escalates to a human instead of closing the alert
 - Guardrails, rate limits, and a kill switch for agents

4.2 Responsible AI (continued)

- **Consistency**: similar inputs produce similar outputs, today and after the next update
 - LLMs are probabilistic; set temperature low, pin the model version, version the system prompt
 - A provider silently updating a hosted model changes behavior with no change on your side
 - Keep a regression suite of known prompts and expected answers; run it on every model or prompt change
 - Consistency also means applying the same decision logic to every case and group

■ Transparency and explainability

- **Transparency**: openness about the fact that AI is used, what it does, what data it was trained on, and its known limits
 - Tell users when they are talking to a chatbot or seeing generated content
 - Publish a model card: intended use, training data summary, evaluation results, limitations
- **Explainability**: the ability to say why the model produced a particular output
 - Interpretable by design: a decision tree or a linear scorecard
 - Post hoc for complex models: SHAP and LIME feature attributions, counterfactuals ("approved if income were 5% higher")
 - For RAG assistants: cite the retrieved source passages so a human can verify the answer
 - An LLM's written reasoning is not proof of how it reached the answer
- Regulated decisions (credit, hiring, insurance) often require a reason the affected person can understand

■ Privacy, security, and differential privacy

- **Privacy and security**: personal data is collected lawfully, used only for its purpose, and protected; the AI system resists attack
 - Data minimization, consent, retention limits, and access control apply to training data, prompts, and logs
 - Models can memorize training records; membership inference and extraction attacks recover them (objective 2.6)

- **Differential privacy** (DP) adds calibrated random noise so the output barely changes whether any one person's record is included or not
 - **Epsilon** (ϵ), the privacy budget: smaller epsilon means more noise and stronger privacy
 - Formal promise: adding or removing one person changes the probability of any result by at most a factor of e^ϵ
 - Laplace mechanism for counts: add noise with scale (sensitivity $\div \epsilon$); a count's sensitivity is 1
 - Every query spends budget; repeated queries add their epsilons together
- Where DP is used
 - Releasing statistics: the US Census Bureau protected 2020 census tables with DP
 - Training: DP-SGD clips each example's gradient and adds noise, limiting what the model can memorize

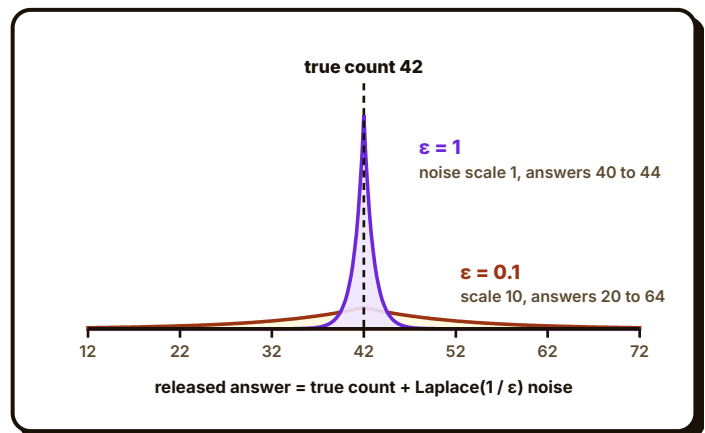


FIGURE The Laplace mechanism. Lower epsilon widens the noise, so the result says less about any one person.

- Demonstration: release "how many users pasted secrets into the chatbot this week" (true count 42) at two epsilon values

4.2 Responsible AI (continued)

DP_COUNT.PY

```
import numpy as np

rng = np.random.default_rng(2026)
TRUE = 42      # users who pasted secrets
SENS = 1      # one person changes it by 1

def noisy(count, eps, n):
    scale = SENS / eps
    return count + rng.laplace(0, scale, n)

for eps in (1.0, 0.1):
    one = noisy(TRUE, eps, 1)[0]
    d = noisy(TRUE, eps, 200_000)
    d2 = noisy(TRUE + 1, eps, 200_000)
    err = np.mean(np.abs(d - TRUE))
    p, p2 = np.mean(d > 42.5), np.mean(d2 > 42.5)
    print(f"eps={eps}: answer={one:6.1f}"
          f"   avg_err={err:5.2f}"
          f"   ratio={p2 / p:4.2f}")
    print(f"e^1={np.e:.2f}   e^0.1={np.exp(0.1):.2f}")
```

TERMINAL

```
$ python3 dp_count.py
eps=1.0: answer= 41.0   avg_err= 1.00   ratio=2.31
eps=0.1: answer= 52.5   avg_err=10.00   ratio=1.10
e^1=2.72   e^0.1=1.11
```

TRY IT

- Copy `dp_count.py` out of `tests/4.2-differential-privacy.sh` and change the epsilon values to 5.0 and 0.01
- Note how far a single released answer lands from 42 at each setting, and decide which epsilon you would accept for a weekly security metrics report

- Reading the output
 - At $\epsilon = 1$ the answer is off by about 1 on average; at $\epsilon = 0.1$, by about 10
 - `ratio` compares how often the released answer exceeds 42.5 when the true count is 43 (one extra person) versus 42; it stays under e^ϵ , which is the DP guarantee
 - The trade-off is utility against privacy; choose epsilon per use case and track the budget

Accountability and awareness training

- **Accountability**: a named person or body owns each AI system's outcomes and can pause or retire it
 - "The model decided" is never an acceptable answer to a regulator or a customer
 - Approval records, an audit trail of changes, and an incident process make accountability provable
 - Human oversight for high-impact decisions: a person can review, override, or stop the system
- **Awareness training**: everyone who uses or builds AI understands its limits and the rules
 - All staff: hallucinations, what data may never go into a prompt, how to report a problem
 - Builders: secure design, bias testing, the approval process
 - Leaders: risk appetite and their accountability
 - The EU AI Act (Article 4) requires providers and deployers to take measures supporting AI literacy among their staff

ON THE EXAM

- Adding statistical noise so no single record can be inferred from a result is differential privacy; a smaller epsilon is more private and less accurate.
- "Users cannot understand why the model denied the request" is an explainability gap. "Users did not know an AI was involved" is a transparency gap.
- The same prompt giving different answers after a model update is a consistency problem; pin versions and run regression tests.

4.2 AI Risks and Shadow AI

4.0 AI GOVERNANCE, RISK, AND COMPLIANCE

Introduction of bias

- **Introduction of bias**: the model learns or amplifies unfair patterns and produces skewed outcomes for some groups
- Where bias gets in

4.2 AI Risks and Shadow AI (continued)

SOURCE	EXAMPLE
Historical data	Past hiring favored one group; the model copies it
Sampling	Phishing classifier trained only on English mail misses other languages
Labeling	Analysts label one team's alerts "noise" more often
Proxy features	ZIP code standing in for ethnicity
Feedback loop	The model's own decisions become next month's training data

- Real case: Amazon dropped an experimental résumé-screening model after finding it downgraded résumés that mentioned "women's" (reported by Reuters, 2018)
 - Trained on ten years of mostly male hiring history
- Business impact: discrimination claims, regulatory action, bad decisions at scale
- Controls: representative data, bias metrics per group before release and in production, human review of adverse decisions, an **impact assessment** for systems that judge people

■ Accidental data leakage

- **Accidental data leakage**: sensitive data leaves its intended boundary through an AI system, with no attacker required
- How it happens
 - Users paste source code, customer records, or incident details into a public chatbot whose terms allow the provider to retain or train on prompts
 - A RAG assistant indexes a SharePoint site with broken permissions and answers anyone's question with HR files
 - A model memorizes training records and repeats them on request
 - Prompts and responses land in logs, analytics, or a vendor's abuse-monitoring store with long retention
- Real case: in 2023 Samsung engineers pasted internal source code into ChatGPT; Samsung then restricted generative AI tools on company devices
- Maps to **LLM02:2025 Sensitive Information Disclosure** in the OWASP Top 10 for LLM Applications
- Controls: approved tools with no-training and retention terms, DLP and redaction at the AI gateway, permission-aware retrieval, log minimization, awareness training

■ Accuracy, performance, and reputational loss

- **Accuracy and performance of the model**: a model can be confidently wrong, and it can get worse over time
 - Hallucination (NIST calls it **confabulation**): fluent, plausible, false output
 - **Model drift**: the world changes (new attack techniques, new products) and accuracy decays with no code change
 - Latency and cost: a model too slow or expensive for its workload fails in a different way
- Real cases
 - Mata v. Avianca (US District Court, 2023): lawyers fined after filing a brief citing court cases ChatGPT invented
 - Moffatt v. Air Canada (2024): a tribunal held the airline liable for refund terms its website chatbot made up; "the chatbot is responsible" was rejected
- **Reputational loss**: public harm to trust and brand when AI behaves badly in view of customers
 - Microsoft's Tay chatbot (2016) was pulled within a day after users taught it to post offensive content
 - Screenshots travel faster than corrections; one bad answer becomes the story
- Controls: evaluation sets and acceptance thresholds, grounding with cited sources, human review for external output, drift monitoring, a rollback plan, a communications plan for AI incidents

■ Intellectual Property (IP)-related risks

- **Intellectual Property (IP)-related risks** run in both directions
- Inbound: the model or its output infringes someone else's rights
 - Training data scraped without license; copyright suits such as The New York Times v. OpenAI and Microsoft (filed 2023)
 - Generated code that reproduces licensed code (GPL) without attribution
 - Open-weight model licenses that restrict commercial use or require attribution
- Outbound: the organization loses its own IP
 - Trade secrets or unreleased designs pasted into a public model
 - Output may not be protectable: US law requires human authorship for copyright (Thaler v. Perlmutter, D.C. Circuit, 2025)
- Controls: legal review of model and data licenses, vendor **IP indemnification** clauses, code scanning for license matches, rules on what may enter prompts

4.2 AI Risks and Shadow AI (continued)

■ Autonomous systems

- **Autonomous systems**: AI that takes actions, not just answers: agents that call tools, run scripts, send email, change configurations, or move money
- Risks grow with autonomy
 - A wrong answer becomes a wrong action, often at machine speed and before anyone notices
 - Prompt injection in a document or web page can steer the agent's tool calls
 - Unclear liability when an agent acts on its own
- OWASP calls this **LLM06:2025 Excessive Agency**: too much functionality, permission, or autonomy

- Controls
 - Least-privilege tool scopes and short-lived credentials
 - Human approval for high-impact or irreversible actions
 - Rate and spend limits, full action logging, and a kill switch
 - Sandboxed execution for any code the agent writes

■ Shadow IT and shadow AI

- **Shadow IT**: technology used without IT's knowledge or approval (a personal Dropbox, an unapproved SaaS app)
- **Shadow AI**: the same pattern for AI: unapproved chatbots, browser extensions, meeting transcription bots, AI features switched on inside approved SaaS, or a team's self-hosted model nobody registered

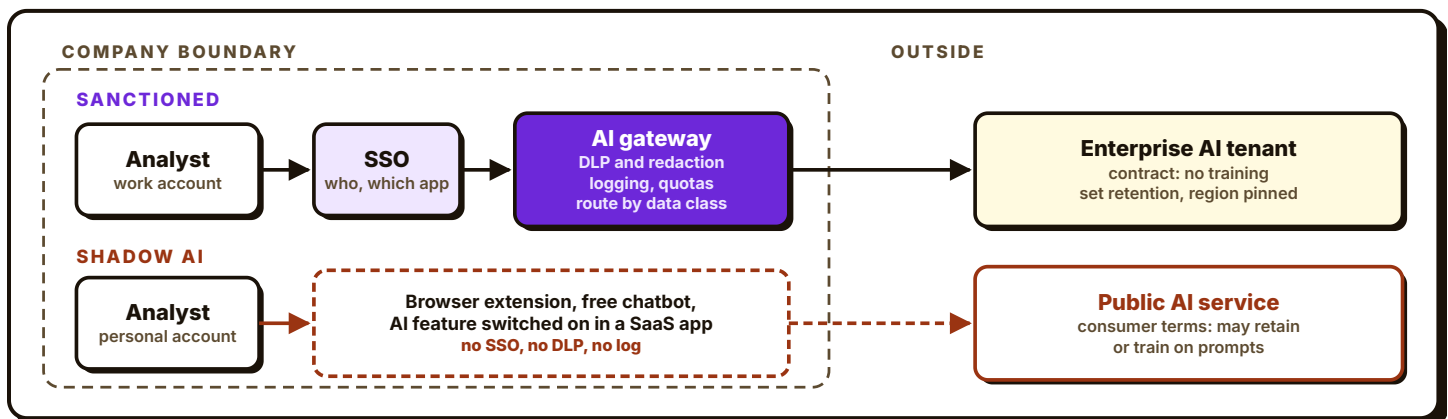


FIGURE The sanctioned path redacts, logs, and routes through a gateway. Shadow AI skips every control, so nobody can see or recall what was sent.

- Why shadow AI is worse than ordinary shadow IT
 - Every prompt is a potential data transfer to a third party
 - Output feeds decisions with no review, no accuracy check, and no record
 - Many tools are free and need only a browser, so procurement never sees them
- Finding it
 - Secure web gateway, CASB, and DNS logs for traffic to known AI domains
 - Browser extension inventory from endpoint management; OAuth grants in the identity provider
 - Expense reports and corporate card statements for AI subscriptions
 - Surveys and amnesty: ask teams what they use before blocking it
- Fixing it: give people a sanctioned tool that is good enough, then block or coach on the rest (objective 4.3)

ON THE EXAM

- Customer data pasted into a free public chatbot is accidental data leakage through shadow AI. Fix it with an approved alternative plus policy and DLP; a block alone pushes use to personal phones.
- Accuracy falling over months with no code change is model drift.
- An agent that deletes records without human approval is an autonomous-system risk (excessive agency).

■ Laws, frameworks, and standards

- Law is **mandatory**, with regulators and fines (the EU AI Act)
- Principles and frameworks are **voluntary** guidance (OECD AI Principles, NIST AI RMF)
- Standards are auditable requirements, some **certifiable** (ISO/IEC 42001)
- Compliance reaches the whole life cycle: what you may build, what must be tested and documented, what users must be told

■ European Union (EU) AI Act

- **European Union (EU) AI Act**: Regulation (EU) 2024/1689, the first comprehensive AI law
 - Entered into force 1 August 2024; obligations phase in over several years
 - Amended by the **AI Omnibus**, Regulation (EU) 2026/1744 (in force 27 July 2026), which pushed back the high-risk dates
- Reaches beyond Europe: covers non-EU providers whose AI is sold in the EU or whose output is used there
- **Risk-based**: obligations scale with the harm a use can cause

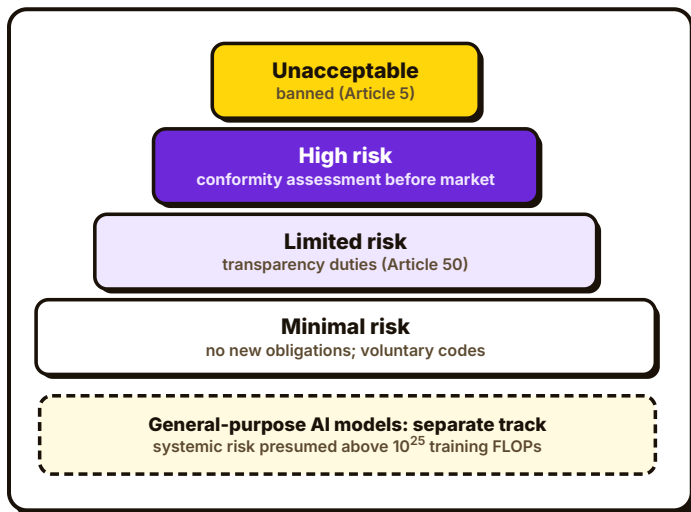


FIGURE The four EU AI Act risk tiers. General-purpose AI models carry a separate set of obligations alongside the tiers.

- Unacceptable risk: banned outright (Article 5)
 - Harmful manipulation; exploiting age or disability; social scoring; predicting crime from profiling alone
 - Untargeted facial-image scraping; emotion recognition at work or school; inferring sensitive traits from biometrics
 - Real-time remote biometric ID in public for policing (narrow exceptions)
 - Added by the AI Omnibus: AI that generates non-consensual intimate imagery or child sexual abuse material
- **High risk**: allowed with strict duties
 - Safety components of products under EU product laws (Annex I: machinery, toys, medical devices)
 - Annex III uses: biometrics, critical infrastructure, education, employment, essential services such as credit scoring, law enforcement, migration and border control, justice and democratic processes
 - Provider duties: risk management system, data governance, technical documentation, logging, human oversight, accuracy and cybersecurity, quality management, conformity assessment, CE marking, EU database registration
- Limited risk: transparency duties (Article 50)
 - Tell people they are dealing with a chatbot; label deepfakes; mark generated content in a machine-readable way
- Minimal risk: spam filters, game AI; no new obligations
- **General-purpose AI (GPAI) models**: documentation, copyright policy, and a training-content summary
 - Models trained with more than 10^{25} FLOPs are presumed to carry systemic risk and must also evaluate, red-team, report incidents, and secure the model

■ Roles, dates, and penalties

- Roles: the Act calls all of them **operators**

4.3 AI Regulations and Standards (continued)

ROLE	WHO
Provider	Develops an AI system or GPAI model and places it on the market under its own name
Deployer	Uses an AI system under its authority in a professional context
Importer	EU-based; brings in a system bearing a non-EU provider's name
Distributor	Makes it available in the supply chain; neither provider nor importer
Authorized representative	EU-based; holds a written mandate from a non-EU provider

- Most obligations sit with the provider; deployers of high-risk systems must follow the instructions, assign human oversight, monitor, and keep the logs
- A deployer becomes a provider (Article 25) if it puts its own name on a high-risk system, substantially modifies it, or changes its purpose so that it becomes high-risk
 - Fine-tuning a vendor model and shipping it under your brand is a common way to cross that line

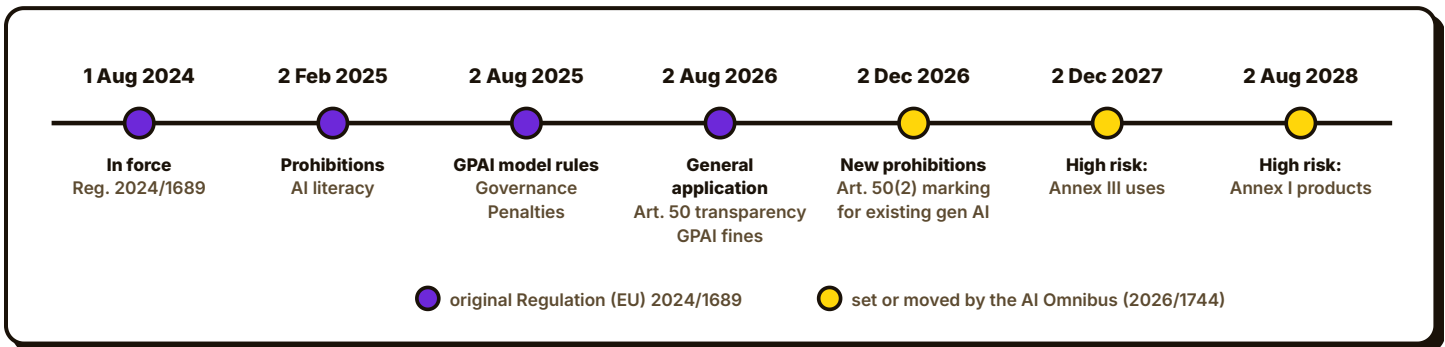


FIGURE EU AI Act application dates as amended by Regulation (EU) 2026/1744. The high-risk rules moved from August 2026 to December 2027 and August 2028.

- GPAI models already on the market before 2 August 2025 have until 2 August 2027 to comply
- **Penalty ceilings** (Article 99): whichever is higher, a fixed sum or a share of worldwide annual turnover

VIOLATION	UP TO
Prohibited practice (Article 5)	EUR 35 million or 7%
Most other obligations	EUR 15 million or 3%
Incorrect or misleading information to authorities	EUR 7.5 million or 1%
GPAI model providers (Article 101)	EUR 15 million or 3%

- For SMEs and start-ups the lower of the two amounts applies

■ OECD AI Principles

- **Organisation for Economic Co-operation and Development (OECD) standards:** the OECD AI Principles, the first intergovernmental AI standard
 - Adopted May 2019; updated May 2024; 47 adherents
 - Voluntary, but they shaped the G20 AI principles and the EU AI Act, whose definition of an "AI system" closely follows the OECD's
- Five values-based principles for AI actors
 - Inclusive growth, sustainable development and well-being
 - Human rights and democratic values, including fairness and privacy
 - Transparency and explainability
 - Robustness, security and safety
 - Accountability
- Five recommendations for governments: invest in AI R&D, build an inclusive AI ecosystem, interoperable governance and policy, human capacity and labor market transition, international cooperation

4.3 AI Regulations and Standards (continued)

- The **2024 update** responded to general-purpose and generative AI
 - Safeguards to override, repair, or decommission a system that behaves in unwanted ways
 - Misinformation and information integrity; respect for intellectual property
 - Responsible business conduct across the life cycle; clearer transparency disclosures; environmental sustainability

■ ISO/IEC AI standards

- **International Organization for Standardization (ISO) AI standards** are joint ISO/IEC work from committee JTC 1/SC 42

STANDARD	TITLE (SHORT)	USE IT FOR
ISO/IEC 42001:2023	AI management system (AIMS)	Certifiable requirements for running an AI program; the AI counterpart to ISO/IEC 27001
ISO/IEC 23894:2023	Guidance on AI risk management	Applying ISO 31000 risk management to AI
ISO/IEC 22989:2022	AI concepts and terminology	Shared vocabulary for contracts, policies, and other standards
ISO/IEC 23053:2022	Framework for AI systems using ML	Reference description of an ML system's components
ISO/IEC 42005:2025	AI system impact assessment	Assessing effects on people and society
ISO/IEC 42006:2025	Requirements for AIMS audit and certification bodies	Who may certify you against 42001
ISO/IEC 38507:2022	Governance implications of AI use	Board and executive oversight of AI
ISO/IEC TR 24027:2021	Bias in AI systems	Sources and measurement of bias

- Only 42001 is a certifiable management-system standard; it shares 27001's clause structure, so an existing ISMS extends to cover AI

■ NIST AI Risk Management Framework

- **National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF):** NIST AI 100-1, version 1.0, January 2023
 - Voluntary; NIST is revising it under the 2025 White House AI Action Plan, and 1.0 remains the published version

- Seven trustworthy characteristics: valid and reliable; safe; secure and resilient; accountable and transparent; explainable and interpretable; privacy-enhanced; fair with harmful bias managed
- The Generative AI Profile (NIST AI 600-1, July 2024) applies the RMF to generative AI
 - Lists 12 risks, including confabulation, data privacy, information security, information integrity, intellectual property, harmful bias or homogenization, and value chain and component integration; suggests actions per RMF function

FUNCTION	DOES
Govern	Cross-cutting culture, policies, accountability, third-party risk; applies to all others
Map	Establish context: purpose, users, impacts, risks
Measure	Test, evaluate, and track the mapped risks
Manage	Prioritize, treat, respond to, and recover from risks

4.3 AI Regulations and Standards (continued)

ON THE EXAM

- EU AI Act tiers by example: social scoring is prohibited, CV screening is high risk, a support chatbot is limited risk, a spam filter is minimal.
- The builder is the provider; the user is the deployer. Rebranding or substantially modifying a high-risk system makes a deployer a provider.
- ISO/IEC 42001 is certifiable; the NIST AI RMF (Govern, Map, Measure, Manage) is voluntary.

4.3 AI Policies and Data Sovereignty

4.0 AI GOVERNANCE, RISK, AND COMPLIANCE

■ Corporate policies

- **Corporate policies** for AI translate laws, contracts, and risk appetite into rules for employees, developers, and vendors
 - Regulations set the floor; customer contracts and the company's own risk appetite usually go further
- The usual set
 - AI acceptable use policy: approved tools, forbidden data, human review of output, disclosure of AI-generated work
 - AI development policy: testing, documentation, and approval gates before release
 - Data classification and handling, extended to prompts, embeddings, training sets, and logs
 - Third-party and procurement policy: what a vendor must prove before purchase
- Compliance shapes how the business uses AI
 - Some uses are off the table entirely (prohibited practices, regulated decisions without human review)
 - Others need extra time and evidence: impact assessments, conformity documentation, customer notices
- A policy needs enforcement to matter
 - SSO-only access to approved AI tools; an **AI gateway** in front of model APIs; DLP on prompts; logging
 - A documented **exception process** with an expiry date, so teams ask instead of working around the rule

■ Sanctioned vs. unsanctioned

- **Sanctioned vs. unsanctioned**: whether the organization has reviewed and approved an AI tool for a given use and data class
 - Sanctioned: contract signed, security and privacy review done, SSO and logging enabled, listed in the AI inventory
 - Unsanctioned: anything else, including a personal account on an approved product

- Most organizations publish three lists

STATUS	MEANING	EXAMPLE HANDLING
Approved	Use for listed data classes	Enterprise tenant via SSO
Tolerated	Public data only; under review	Allowed, coached by a browser banner
Prohibited	Must not be used	Blocked at the web gateway

- Approval is per use and per data class
 - A tool sanctioned for marketing copy is not automatically sanctioned for customer records
- Blocking alone fails; people move to personal phones
 - Pair every block with a sanctioned alternative and a fast way to request new tools

■ Private vs. public models

- **Public models**: hosted services open to anyone, usually consumer or free tiers
 - Prompts may be retained, reviewed by the provider, or used for training under consumer terms
 - No company contract, no SSO, no audit logs available to the company
- **Private models**: models the organization controls or has contracted for exclusive, isolated use
 - Enterprise tenant of a hosted model: contractual no-training, set retention, regional processing, admin logs
 - Self-hosted open-weight model in the company's own cloud account or data center: data never leaves; the company carries patching, scaling, and model security
- The terms also describe model weights: open-weight models anyone can download versus proprietary models reachable only through an API

4.3 AI Policies and Data Sovereignty (continued)

	PUBLIC SERVICE	ENTERPRISE TENANT (HOSTED)	SELF-HOSTED PRIVATE MODEL
Where prompts go	Provider, under consumer terms	Provider, under contract	Your infrastructure
Training on your data	Possible; check terms	Contractually excluded	Only if you choose
Logging and audit	None for you	Admin and audit logs	Full control
Effort and cost	Lowest	Subscription	Hardware, staff, patching
Typical data allowed	Public only	Internal and confidential	Up to restricted

■ Sensitive data governance

- **Sensitive data governance** decides which data may reach which AI system, in what form, and for how long
 - Classify first: public, internal, confidential, restricted (PII, PHI, cardholder data, credentials, source code, M&A plans)
 - Map each class to the model tiers allowed to process it

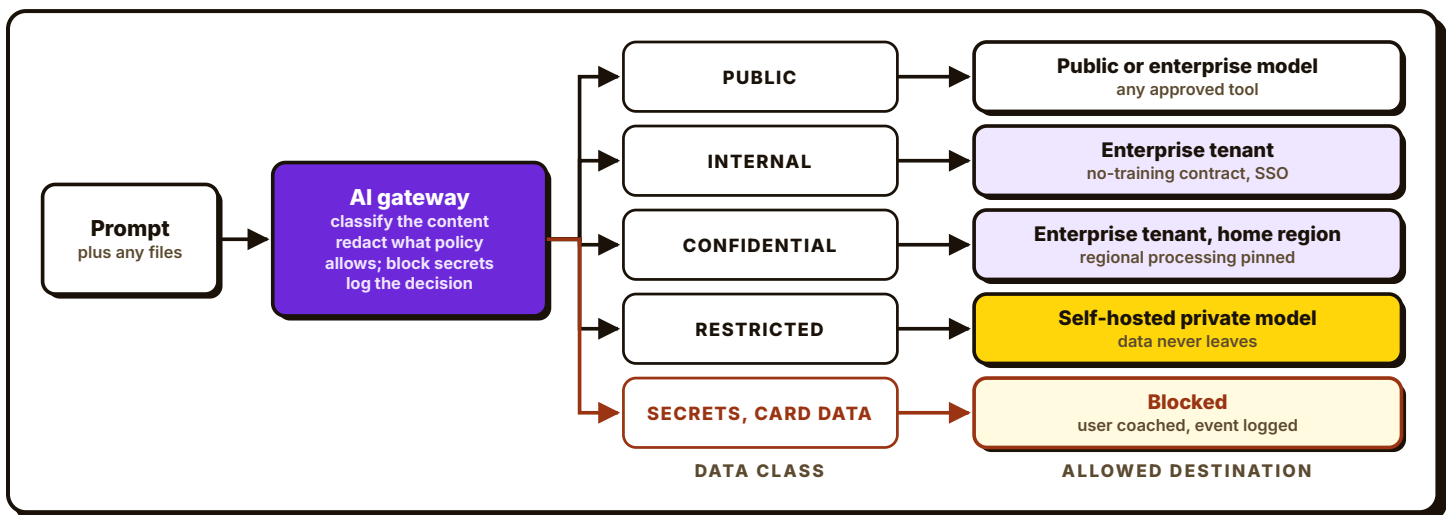


FIGURE Classification drives routing. The gateway inspects each prompt, redacts what policy allows it to redact, and sends the request only to a model tier approved for that data class.

- Controls across the AI data flow
 - Prompts: DLP and redaction at the gateway; block secrets and card numbers outright
 - Training and fine-tuning data: documented lawful basis, minimization, de-identification; a record of what went into each model
 - RAG sources: index only what the requesting user may already read; filter retrieval by the user's permissions
 - Outputs and logs: classify them like the inputs that produced them; set retention; restrict who can read prompt logs
- Regulated data brings its own rules: GDPR for EU personal data, HIPAA for US health data, PCI DSS for cardholder data
- **Third-party compliance evaluations**
 - **Third-party compliance evaluations:** due diligence on every AI vendor, model provider, and AI feature inside an existing SaaS product
 - The vendor's model is part of your attack surface and your compliance scope

4.3 AI Policies and Data Sovereignty (continued)

- Evidence to request
 - Independent reports and certifications: **SOC 2 Type II**, ISO/IEC 27001, ISO/IEC 42001, FedRAMP authorization for US federal work
 - Data processing agreement: no training on your data, retention period, deletion on termination, breach notice
 - Subprocessor list and processing locations, including where inference runs
 - Model documentation: model card, evaluation and red-team results, known limitations
 - Security testing: recent penetration test summary, vulnerability disclosure program
 - Regulatory position: the vendor's role under the EU AI Act (provider) and its conformity evidence for high-risk uses
- Contract terms worth fighting for: right to audit, notice before model or subprocessor changes, IP indemnification
- Reassess on a schedule and when the vendor changes its model, terms, or hosting region
- **Data sovereignty**
 - **Data sovereignty**: data is subject to the laws of the country where it is collected or stored, and that country's authorities can reach it
 - Related terms
 - **Data residency**: where the data physically sits (a business or contractual choice)
 - **Data localization**: a legal requirement that certain data stay inside a country
 - Why AI makes it harder
 - Inference can run in a different region from storage; a prompt sent to an EU-stored tenant may be processed elsewhere unless regional processing is set
 - Training data, embeddings, logs, and backups each have a location
 - A provider headquartered in one country may be compelled by its home law (the US CLOUD Act) to produce data stored abroad
 - Transfer rules
 - GDPR Chapter V restricts transfers of personal data out of the EU: adequacy decisions (such as the EU-US Data Privacy Framework), standard contractual clauses, binding corporate rules
 - Other countries have their own regimes, such as China's PIPL
 - Controls: pin storage and inference to approved regions, customer-managed encryption keys, sovereign cloud offerings, self-hosted models for the most restricted data

ON THE EXAM

- "Employees use a personal account on an approved AI product" is still unsanctioned use; sanctioning covers the tenant, contract, and data class, not just the brand.
- Data residency is where data is stored; data sovereignty is whose laws govern it. A question about a foreign government compelling access is a sovereignty question.
- Before buying an AI service, a third-party compliance evaluation asks for SOC 2 Type II or ISO certifications, a data processing agreement with no-training terms, and processing locations.