NextGen GRC

AI GOVERNANCE & INTERNAL AUDIT SERIES

AI Audit Assurance Playbook

A practical, step-by-step framework for auditing AI systems — covering governance review, risk tiering, control testing, bias validation, and continuous monitoring, with worked examples drawn from credit scoring, HR screening, and generative-AI use cases.

NextGen GRC Consultants
Compliance & Internal Audit Advisory

Edition 2026.1
nextgengrc.world

Table of Contents

01 Why AI Systems Need a Distinct Audit Assurance Approach 3

02 The Regulatory & Standards Landscape 4

03 The 7-Step AI Audit Assurance Methodology 5

04 Step Detail — Inventory, Risk Tiering & Governance Review 6

05 Step Detail — Data, Model & Control Testing 8

06 Step Detail — Bias Validation & Continuous Monitoring 10

07 Worked Example A — Credit Risk Scoring Model 11

08 Worked Example B — AI Resume Screening Tool 12

09 Worked Example C — Generative AI Support Assistant 13

10 Sample AI Control Matrix 14

11 Fieldwork Checklist 15

12 Glossary & Disclaimer 16

SECTION 01

Why AI Systems Need a Distinct Audit Assurance Approach

Traditional IT General Controls (ITGC) audit programs were built for deterministic systems: a given input, processed by unchanging logic, always produces the same output. AI systems — particularly machine learning models and generative AI — break that assumption. The same input can produce different outputs depending on model version, prompt construction, retrieval context, temperature settings, or silent data drift in production. A control framework that only asks "is access restricted?" and "is change management documented?" misses the questions that actually drive AI risk: was the model trained on representative data, does its behavior remain stable over time, can a human meaningfully override it, and can the organization explain a specific decision after the fact?

AI Audit Assurance is the discipline of extending internal audit and SOX/ITGC methodology to address these AI-specific risks without discarding the rigor of traditional control testing. It treats an AI system the way a financial auditor treats a complex estimate: the inputs, the methodology, and the output all need independent evaluation — not just the access controls wrapped around them.

Key takeaway

AI audit assurance is not a replacement for ITGC — it is a targeted extension. Access, change management, and operations controls still apply to the infrastructure hosting the model. What changes is the addition of model-specific testing: data lineage, drift detection, bias measurement, explainability, and human-override verification.

Who should use this playbook

- Internal audit and SOX compliance teams adding AI systems to their ITGC scope for the first time
- External auditors evaluating automated controls that incorporate machine learning or generative AI components
- Risk and compliance leaders building an AI governance program ahead of EU AI Act or ISO/IEC 42001 certification
- GRC platform teams designing control libraries and evidence-collection workflows for AI-enabled processes

SECTION 02

The Regulatory & Standards Landscape

No single regulation governs AI audit assurance globally. Instead, auditors are synthesizing requirements from several overlapping frameworks. Understanding how each one contributes to the assurance program helps scope an engagement without duplicating effort.

Framework	Primary Focus	Relevance to Audit Assurance
EU AI Act	Risk-tiered obligations (unacceptable / high / limited / minimal risk) for AI systems placed on the EU market	Provides the risk-tiering vocabulary this playbook adapts in Step 2 — even for organizations outside the EU, the tiering logic is a useful triage tool
ISO/IEC 42001:2023	Management system standard for AI — governance, lifecycle controls, and continual improvement	Defines the AI Management System (AIMS) structure that an auditor can test for design and operating effectiveness, analogous to ISO 27001 for security
NIST AI Risk Management Framework	Voluntary framework organized around Govern, Map, Measure, and Manage functions	Supplies practical control objectives for bias measurement, robustness testing, and documentation that map cleanly onto Steps 4–6 of this methodology
COSO 2023 Supplemental Guidance	Extends the 2013 Internal Control–Integrated Framework to cloud, AI, and algorithmic environments	Confirms that AI-driven decisions affecting financial reporting fall within ICFR scope — meaning SOX 404 assessments must now consider model governance
Sector-specific guidance (e.g., model risk management circulars in financial services)	Model validation, independent review, and ongoing monitoring requirements for quantitative models	Where applicable, sector model-risk rules often set the bar for what "independent validation" must look like — auditors should confirm sector rules first

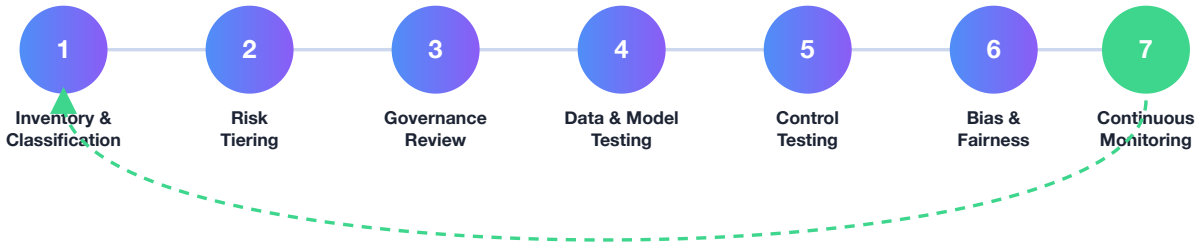
⚠ **Common scoping mistake**

Treating "AI governance" and "AI audit assurance" as the same exercise. Governance defines policy and ownership; audit assurance independently tests whether that policy is actually followed and effective. A mature governance document with no corroborating evidence of testing is, by itself, a control deficiency.

SECTION 03

The 7-Step AI Audit Assurance Methodology

The methodology below sequences naturally from "what AI exists" through "is it still behaving as expected." Each step produces evidence that feeds the next — inventory feeds risk tiering, risk tiering determines testing depth, and testing results feed the ongoing monitoring cadence.



Continuous monitoring findings re-trigger inventory & risk tiering — this is a loop, not a line

Steps 1–3 are largely document- and interview-based and can usually be completed in one to two weeks for a single AI system. Steps 4–6 require technical testing — often with data science or model-validation support embedded in the audit team. Step 7 is not a one-time activity; it defines the recurring cadence (typically quarterly for high-risk systems) that keeps the assurance current between full re-audits.

SECTION 04

Step Detail — Inventory, Risk Tiering & Governance Review

Step 1 — AI System Inventory & Classification

You cannot audit what you have not catalogued. The inventory step builds a register of every AI and machine learning system in scope, including embedded AI inside third-party SaaS tools that the business may not think of as "AI." For each system, record: business owner, purpose, model type (rules-based, classical ML, deep learning, generative/LLM), data sources, deployment environment, and whether the system makes or materially influences a decision affecting customers, employees, or financial reporting.

> ⚠ **Shadow AI risk**
> The most common inventory gap is "shadow AI" — generative AI features quietly enabled inside existing SaaS platforms (CRM next-best-action suggestions, support-ticket auto-summarization, embedded copilots) that were never formally onboarded as AI systems. Interview process owners directly; do not rely solely on a vendor or IT asset register.

Step 2 — Risk Tiering

Once inventoried, each system is assigned a risk tier based on the severity of harm if it fails, and the degree of autonomy it has. This playbook adapts a four-tier model:

- **Tier 1 — Critical:** Materially affects financial reporting, credit/lending decisions, or health and safety, with limited human review before action is taken.
- **Tier 2 — High:** Significantly influences a decision about an individual (hiring, performance, pricing) but with a documented human-in-the-loop checkpoint.
- **Tier 3 — Moderate:** Supports internal operational efficiency (routing, summarization, drafting) where errors are inconvenient but not consequential to a third party.
- **Tier 4 — Minimal:** Low-stakes, fully reversible, internal-only use with negligible downstream impact.

Testing depth in Steps 4–6 should scale directly with tier: Tier 1 systems warrant full data lineage testing, independent bias measurement, and quarterly re-validation; Tier 4 systems may need only an annual governance attestation.

Step 3 — Governance & Documentation Review

This step tests whether the paperwork that should exist, actually exists and reflects reality. Request and evaluate:

- A model card or system card describing intended use, known limitations, and out-of-scope uses
- Data lineage documentation showing training data sources, collection method, and consent basis where personal data is involved
- Evidence of a pre-deployment approval — who reviewed the model, what criteria they applied, and whether the criteria were actually met
- A named accountable owner (not just a development team) responsible for the system's ongoing performance
- An incident log capturing prior model failures, complaints, or override events and how each was resolved

Red flag

If the model card or system documentation was generated after the audit was announced rather than maintained as a living artifact throughout the model's lifecycle, treat the underlying control as not operating — documentation produced reactively for an audit is evidence of a deficiency, not of a functioning control.

Step Detail — Data, Model & Control Testing

Step 4 — Data & Model Testing

This is where AI audit assurance most clearly diverges from traditional ITGC. The auditor independently tests the substance of the model, not just the access controls around it.

- **Training data provenance:** Confirm the data used to train or fine-tune the model was lawfully obtained, appropriately licensed, and representative of the population the model will be applied to in production.
- **Data drift testing:** Compare the statistical distribution of current production inputs against the original training distribution. Significant drift is an early warning that model performance may be degrading even if no errors have been reported yet.
- **Holdout / challenger testing:** Where feasible, run a held-out test set (or an independently sourced sample) through the model and compare results against documented performance benchmarks.
- **Adversarial and edge-case testing:** For generative AI, test prompt-injection resistance and confirm the system does not act outside its documented scope (e.g., a support chatbot should not be able to authorize refunds it was never granted authority to approve).

Step 5 — Control Testing: The ITGC Extension for AI

Traditional ITGC domains still apply but need AI-specific test attributes added:

ITGC Domain	Traditional Test	AI-Specific Extension
Access to Programs & Data	Review of user access to application and database	Add review of access to model weights, training pipelines, prompt/system-instruction configuration, and fine-tuning datasets
Program Changes	Change tickets, approvals, testing evidence for code deployments	Add model version control: every retrain or prompt change should have its own versioned record, approval, and rollback plan
Computer Operations	Job scheduling, backup, incident management	Add monitoring of inference-time failures, latency-driven fallback behavior, and logging retention sufficient to reconstruct a specific decision
Human Override	Not applicable in traditional ITGC	New domain: test that a human reviewer can and does override or escalate AI output for Tier 1–2 systems, with evidence the override path is actually used, not just theoretically available

Key takeaway

The single highest-value test in AI control testing is usually the human-override walkthrough: pick five real decisions for a Tier 1 or Tier 2 system and confirm a human could meaningfully intervene before harm occurred — not merely that an "override button" exists in the interface.

Step Detail — Bias Validation & Continuous Monitoring

Step 6 — Bias, Fairness & Performance Validation

For any system that touches individuals (Tier 1–2), independently measure outcome distributions across protected or sensitive groups where lawfully permitted to do so, using metrics appropriate to the use case:

- **Disparate impact ratio:** the selection rate for one group divided by the selection rate for the group with the highest rate — a commonly used screening threshold is 0.8, though context and legal counsel should set the actual bar for a given jurisdiction and use case.
- **Demographic parity / equalized odds:** whether error rates (false positive / false negative) are comparable across groups, not just overall accuracy.
- **Performance stability over time:** tracking the chosen fairness and accuracy metrics on a recurring cadence, not as a one-time validation at launch.

> ⚠ **Methodology note**
> Bias testing requires the right sensitive-attribute data to even run the analysis, which itself raises privacy and legal considerations. Engage legal/privacy counsel before collecting or using protected-class data for testing purposes, and document the lawful basis relied upon.

Step 7 — Continuous Monitoring, Re-certification & Reporting

A point-in-time audit answers "was this model acceptable on the day we tested it." Continuous monitoring answers the more important question: "is it still acceptable today." A sustainable program includes:

- Automated drift and performance dashboards reviewed by the accountable owner on a defined cadence (monthly for Tier 1, quarterly for Tier 2)
- A re-certification trigger whenever the model is retrained, the prompt/system instructions change materially, or the underlying foundation model provider issues a new version
- An audit-ready evidence repository — inventory, risk tier, test results, override logs, and incident history — producible on demand
- A standing reporting line to the audit committee or risk committee for Tier 1 systems, summarizing testing coverage and open issues each cycle

> **Highest-risk finding pattern**
> The most frequent finding in second-year AI audits is a Tier 1 system that passed its original validation but was silently retrained or had its prompt logic modified afterward with no corresponding re-test. Treat "model or prompt changed without triggering re-certification" as a control failure equivalent to an unauthorized production code change in traditional ITGC.

SECTION 05 · WORKED EXAMPLE A

Credit Risk Scoring Model

TIER 1 — CRITICAL

System: A gradient-boosted model used by a regional lender to generate a risk score informing consumer loan approval and pricing decisions.

Why Tier 1: Directly determines access to credit and pricing for individual consumers, with regulatory exposure under fair-lending requirements and direct ICFR relevance through loan-loss provisioning.

Audit steps performed

1. Obtained the model card, training data lineage documentation, and the most recent independent model validation report
2. Reconstructed the disparate impact ratio across three protected classes using a held-out sample of 12 months of live decisions
3. Walked through five declined applications to confirm a human underwriter reviewed and could override the model recommendation before the adverse decision was communicated
4. Tested change management evidence for the two model retrains that occurred during the period, confirming each had a documented approval and a corresponding re-validation of the fairness metrics

Representative finding

One retrain event was approved for deployment but the fairness re-validation was completed nine days *after* go-live rather than before — a sequencing deficiency that exposed the lender to nine days of unvalidated production decisions. Remediation: a hard gate was added to the deployment pipeline preventing promotion to production until the fairness re-validation sign-off is recorded.

SECTION 05 · WORKED EXAMPLE B

AI Resume Screening Tool

TIER 2 — HIGH

System: A third-party applicant tracking system feature that ranks incoming resumes against a job description and surfaces a "top match" shortlist to recruiters.

Why Tier 2: Significantly influences who advances in a hiring process, but a recruiter reviews the full applicant pool, not only the shortlist, providing a documented human checkpoint.

Audit steps performed

1. Confirmed the vendor's published model documentation and requested the vendor's own bias testing results, since the model is not independently trainable by the organization
2. Tested whether recruiters in practice reviewed only the AI-ranked shortlist versus the full applicant pool, via interview and sampling of actual hiring decisions
3. Verified the contractual right to audit and the vendor's incident notification obligations if a bias issue were identified post-deployment

Representative finding

Sampling showed recruiters reviewed only the top-ranked shortlist in 80% of sampled roles despite policy requiring full-pool review — meaning the documented human checkpoint was not consistently operating. This was escalated as a significant deficiency in the control design's practical execution, not the AI model itself. Remediation: the ATS workflow was reconfigured so the full applicant list, not just the AI shortlist, is the default recruiter view.

Lesson

When the AI system is third-party and not independently testable, audit focus shifts to vendor due diligence, contractual audit rights, and — critically — whether the human-in-the-loop control is actually exercised in practice, which is fully within the organization's ability to test regardless of model access.

SECTION 05 · WORKED EXAMPLE C

Generative AI Customer Support Assistant

TIER 2 — HIGH

System: An LLM-based chatbot that answers customer billing questions and is authorized to issue account credits up to a defined dollar threshold without human approval.

Why Tier 2: Has standing financial authority (bounded), and incorrect or manipulated behavior carries direct financial and reputational impact.

Audit steps performed

1. Reviewed the system prompt and tool-permission configuration to confirm the credit-issuance ceiling was enforced at the application layer, not merely instructed to the model in natural language
2. Ran a structured set of adversarial prompts attempting to convince the assistant to exceed its credit authority or disclose other customers' account information
3. Sampled 30 days of transcripts and credit-issuance logs to confirm every issued credit stayed within the configured ceiling and was logged with a reconstructable rationale
4. Confirmed escalation logic correctly routed out-of-scope or distressed-customer conversations to a human agent

Representative finding

The credit ceiling was enforced correctly at the application layer (a hard-coded check, not a prompt instruction) and held under adversarial testing — a control operating effectively. However, log retention for transcripts was only 30 days, insufficient to support a 12-month SOX evidence retention requirement. Remediation: retention extended to align with the organization's financial-records retention policy.

SECTION 06

Sample AI Control Matrix

Use this as a starting template — adapt control objectives to the specific AI system's risk tier and use case.

Control Objective	Control Activity	Test of Design	Sample Evidence
AI systems are inventoried and risk-tiered	Quarterly inventory refresh owned by AI governance committee	Inspect inventory register and refresh log	Inventory export, committee meeting minutes
Models are approved before production use	Pre-deployment sign-off by accountable owner and risk function	Trace a sample of deployments to a recorded approval	Approval ticket, model card version history
Model changes are controlled	Versioned retrain/prompt-change process with mandatory re-validation gate	Confirm deployment pipeline blocks promotion without re-validation sign-off	Pipeline configuration, CI/CD gate logs
Human override is meaningful for Tier 1–2 systems	Defined escalation/override workflow with logged usage	Sample real decisions and confirm override path was available and used appropriately	Override logs, case-level walkthrough
Fairness and performance are monitored on an ongoing basis	Recurring drift/bias dashboard reviewed by owner	Inspect dashboard history and review sign-off cadence	Dashboard exports, review attestations
Evidence is retained for audit and regulatory inquiry	Retention policy aligned to financial-records requirements	Confirm retention period and test sample retrievability	Retention policy, sample log retrieval

SECTION 07

Fieldwork Checklist

Before fieldwork begins

- [ ] AI system inventory obtained and cross-checked against at least one independent source (procurement records, SaaS spend, or IT asset register)
- [ ] Risk tier assigned to every in-scope system, with rationale documented
- [ ] Data science or model-validation resource identified to support technical testing in Steps 4 and 6

During fieldwork

- [ ] Model card / system card obtained for each Tier 1–2 system
- [ ] Training data lineage and consent basis reviewed
- [ ] Drift and/or holdout testing performed for Tier 1 systems
- [ ] Access to model weights, prompts, and fine-tuning pipelines reviewed under standard ITGC access-testing procedures
- [ ] Change management evidence traced for at least one retrain or prompt-change event per system
- [ ] Human-override walkthrough completed for Tier 1–2 systems using real decision samples
- [ ] Bias/fairness metrics independently recalculated or vendor-provided results corroborated
- [ ] Log and evidence retention period confirmed against records-retention policy

Before reporting

- [ ] Findings rated by severity using the organization's standard deficiency classification (control deficiency / significant deficiency / material weakness)
- [ ] Remediation owner and target date agreed for each finding
- [ ] Continuous monitoring cadence confirmed or established for the next review cycle

SECTION 08

Glossary & Disclaimer

Glossary

- **Model card / system card:** A structured document describing an AI system's intended use, training data, known limitations, and performance characteristics.
- **Data drift:** A change in the statistical properties of production input data relative to the data the model was trained on.
- **Disparate impact ratio:** A fairness metric comparing the favorable-outcome rate of one group to that of the highest-rated group.
- **Human-in-the-loop:** A control design in which a human reviews, approves, or can override an AI-generated output before it takes effect.
- **Shadow AI:** AI functionality embedded in existing tools or processes that has not been formally inventoried or risk-assessed.
- **Re-certification:** The process of re-validating a model's risk tier, fairness metrics, and control effectiveness after a material change.

Disclaimer

This playbook is provided for general informational and educational purposes only and does not constitute legal, regulatory, audit, or professional compliance advice. It reflects the authors' synthesis of publicly available AI governance and audit concepts as of the publication date and does not represent the official position of any standards body or regulator referenced herein. Regulatory requirements and audit expectations for AI systems are evolving rapidly; organizations should consult qualified legal, audit, and AI governance professionals before designing or relying on an AI audit program. NextGen GRC Consultants makes no representations or warranties regarding the completeness, accuracy, or applicability of this content to any specific organization, system, or regulatory situation.