FREE FIELD GUIDE

AI Under Audit

Field Guide v1.0

Five principles, three checklists, and one regulatory map for putting AI systems under audit in regulated industries.

Paulo Cavallo, PhD

Senior Credit Risk Model Developer
5+ years SR 11-7 regulatory frameworks
Building the bridge between AI and regulated industries

Who this is for: Data scientists, ML engineers, risk and compliance professionals, and product owners who deploy AI where failures have consequences.

SECTION 01

The Five Principles

These mental models underpin everything in this guide. They come from five years of model risk management under SR 11-7 and building AI systems that had to survive regulatory scrutiny. If you internalize nothing else, internalize these.

1

Architecture Beats Prompting

A policy document in SharePoint is not a control. A validation gate that stops bad code from reaching production is a control. The shift from "the AI should do X" to "X happens automatically regardless of what the AI decides" is the same shift from policy to control that model risk management has been teaching for decades. Stop asking the AI more nicely. Build systems that enforce behavior.

2

Grade Outcomes, Not Procedures

A model can have perfect documentation, rigorous development, and a well-credentialed team. None of that matters if the outputs are wrong. The core principle of SR 11-7: evaluate what the model produces, not how it was built. The goal of validation is not to confirm the model works. It is to find where it breaks.

3

Compliance First, Performance Second

Most ML engineers optimize: performance, cost, maybe explainability if there is time. In regulated industries, that hierarchy is inverted. Compliance and auditability are non-negotiable. Explainability is required. Performance is important. Cost is optimized last. This is not conservatism. It is what happens when something goes wrong and you need to explain it to an examiner.

4

Document Failure Modes Before They Happen

In model risk management, we had a saying: if you cannot explain how it breaks, you cannot prove it works. Every AI system has failure modes. The question is whether you discover them during development or during an audit. Systematic risk assessment before deployment is not overhead. It is the thing that lets you ship with confidence.

5

If It Cannot Be Audited, It Cannot Be Trusted

A fine-tuned model that learned patterns from labeled outcomes is, paradoxically, more auditable than the human expert it learned from. You can test it across segments, measure differential performance, version-control the training data, and reproduce it from scratch. Making AI auditable is not a constraint on innovation. It is what makes innovation sustainable in regulated environments.

SECTION 02

The Three Lines of Defense

SR 11-7 structures accountability into three lines. Most AI teams only have the first. If you cannot point to an independent challenge function for your AI systems, you have a gap that regulators will eventually name.

Line	Role	Responsibility	Typical AI Gap
First	Model developers	Build and perform initial testing. Deepest technical knowledge but strongest incentive to believe their work is correct.	Most AI teams stop here.
Second	Independent validation	Review with fresh eyes. Replicate results, stress test assumptions, document weaknesses. Do not trust the developers.	Rarely exists for AI/ML outside banking.
Third	Audit	Verify the first two lines are doing their jobs. Validate the process, not the model directly.	Auditors often lack AI/ML expertise.

What Auditors Actually Ask

After years across the table from model risk reviewers, these are the questions they care about. Notice what they do *not* ask: F1 scores, BLEU scores, perplexity, or benchmark comparisons. They care about whether *this model, in this context*, does what it needs to do reliably.

Scope

- ☐ What is this model supposed to do? What is it explicitly *not* supposed to do?
- ☐ Are those boundaries enforced or just documented?
- ☐ Who approved the use case, and is that approval documented?

Monitoring

- ☐ How do you know the model is still working?
- ☐ What metrics do you track, and what thresholds trigger review?
- ☐ When was the last threshold breach, and what happened?

Lineage

- ☐ Where did the training data come from? How was it processed?
- ☐ Can you reproduce the model from source data if needed?
- ☐ Is there a version history for models, data, and configurations?

Limitations

- ☐ What are the known weaknesses?
- ☐ What populations does it perform poorly on?
- ☐ Are users and downstream systems aware of these limitations?

Change Management

- ☐ How are changes approved? Who signs off?
- ☐ Can you roll back to the previous version if something breaks?
- ☐ Is there documented history of what changed, when, and why?

SECTION 03

The Regulatory Map

Three frameworks form the foundation of AI governance in U.S. financial services. If your governance program maps exclusively to SR 11-7, you are building on an incomplete blueprint.

Dimension	SR 11-7 (2011)	NIST AI 100-1 (2023)	FHFA AB 2022-02
Scope	Quantitative models at banks	All AI systems, cross-industry	All AI/ML at GSEs
Core structure	Develop, validate, govern	Govern, Map, Measure, Manage	Three lines of defense + four risk areas
Model definition	Quantitative methods only	Broad: includes GenAI via 600-1	All AI/ML including vendor-embedded
Explainability	Assumed (interpretable models)	Distinguishes explainability from interpretability	Names "black box risk" as primary concern
Bias / fairness	"Representative" data required	Fairness as core trustworthiness characteristic	Fair lending testing across all lifecycle stages
Third-party AI	Same rigor as internal	Map function includes third-party risk	Flags AI embedded in vendor software
Monitoring	"Commensurate with use"	Continuous risk tracking, drift detection	Heightened expectations for AI/ML
Status	Mandatory for banks	Voluntary (but cited in enforcement)	Mandatory for GSEs

Three Gaps That Matter

The Map Function Gap. NIST introduces a formal Map function with no direct SR 11-7 equivalent. Mapping means establishing context *before* you build: stakeholders, operational environment, intended and unintended uses, third-party dependencies. SR 11-7 assumes you do this during development. NIST makes it a separate, accountable step.

The GenAI Gap. NIST AI 600-1 (2024) extends coverage to hallucinations, factual reliability, adversarial inputs, confabulation monitoring, and third-party model risk for LLM vendors. None of this exists in SR 11-7 or FHFA AB 2022-02. If your institution deploys GenAI, you need 600-1 in your governance map.

The Regulatory Temperature. NIST AI 100-1 is technically voluntary. But the Colorado AI Act cites it for safe harbor protection with penalties up to $20,000 per violation. Federal agencies introduced 59 AI-related regulations in 2024 alone. "Voluntary" today means "expected" tomorrow.

SECTION 04

AI Governance Quick Assessment

Score your AI system or program. For each question, mark Yes (2 points), Partial (1 point), or No (0 points). Total out of 24.

#	Question	Y	P	N
1	Can you identify every AI/ML system operating in your organization, including those embedded in vendor software?	☐	☐	☐
2	Does each AI system have a documented owner, defined scope, and explicit boundaries for intended use?	☐	☐	☐
3	Is there an independent challenge function (second line) that validates AI systems separately from the development team?	☐	☐	☐
4	Can you trace model outputs back to training data and explain the decision pathway to a non-technical stakeholder?	☐	☐	☐
5	Are failure modes documented before deployment, including edge cases and distribution tails?	☐	☐	☐
6	Do you have automated, continuous monitoring for performance degradation, data drift, and output anomalies?	☐	☐	☐
7	Are bias and fairness evaluations performed across protected segments at development and ongoing in production?	☐	☐	☐
8	Does your framework address GenAI-specific risks: hallucination, confabulation, prompt injection, third-party model risk?	☐	☐	☐
9	Is there a formal change management process with approval gates, version history, and rollback capability?	☐	☐	☐
10	Can you produce on demand a complete audit package: purpose, data, methodology, limitations, monitoring, and incidents?	☐	☐	☐
11	Are controls architectural (enforced by the system) rather than behavioral (dependent on human compliance)?	☐	☐	☐
12	Does your AI inventory definition cover LLMs, agents, and AI-assisted tools, not just traditional quantitative models?	☐	☐	☐

Scoring Guide

Score	Assessment	Priority
20-24	Strong foundation. Focus on continuous improvement and emerging risks (GenAI, agentic AI).	Maintain and extend
14-19	Partial coverage. Core governance exists but gaps in monitoring, independence, or AI-specific risks.	Close gaps systematically
8-13	Significant gaps. Governance may exist on paper but lacks enforcement, independence, or AI-native coverage.	Prioritize fundamentals
0-7	Pre-governance. AI systems operating without structured oversight. Regulatory risk is high.	Start with inventory and ownership

How to use this assessment: Run it once for your overall AI program, then once per high-risk AI system. The per-system scores reveal where governance is uneven. A program that scores 18 overall but has individual systems scoring 6 has a false sense of security.

"Governance, done right, is not a tax on productivity.
It is what makes productivity sustainable."

SECTION 05

Controls That Execute

The most common failure in AI governance: the controls are suggestions. A policy the AI can ignore is not a control. A compliance checklist nobody verifies is not governance.

Governance Theater	Real Controls
"The AI should follow these guidelines"	Validation gate that blocks non-compliant outputs
Policy document in SharePoint	Pre-commit hooks that enforce standards automatically
Quarterly manual review of outputs	Continuous automated monitoring with threshold alerts
"We have guardrails in place"	Input validation, hard-coded limits, approval workflows
Annual bias assessment	Ongoing fairness metrics tracked per segment in production
"Our team reviews all AI outputs"	Deterministic audit trail logging every decision with source attribution

Pre-Deployment Readiness Checklist

Before any AI system goes to production. If you cannot check every box, you have documented the gaps. That documentation itself is governance.

Architecture and Design

- ☐ System purpose, scope, and limitations documented
- ☐ Architecture decisions recorded with rationale for alternatives rejected
- ☐ Third-party dependencies inventoried with risk assessment

Data and Model

- ☐ Training data provenance documented and reproducible
- ☐ Model outputs traceable to input data and decision logic
- ☐ Bias and fairness evaluations completed across relevant segments

Testing and Validation

- ☐ Independent validation performed (not by the development team)
- ☐ Edge cases, adversarial inputs, and failure modes tested and documented
- ☐ Performance metrics defined with acceptance thresholds

Operations

- ☐ Monitoring dashboard: drift detection, performance tracking, alerting
- ☐ Incident response plan: notification, rollback triggers, documentation
- ☐ Change management: approval gates, version control, rollback capability

Governance

- ☐ System registered in model/AI inventory with assigned owner
- ☐ Regulatory mapping completed (which frameworks apply, requirements met)
- ☐ Audit package can be produced on demand

GO DEEPER

Continue the Conversation

This field guide is the starting point. **Exit Code 2 | AI Under Audit** goes deeper every issue with case studies from real AI systems, regulatory analysis, and practical frameworks you can implement immediately.

Subscribe to Exit Code 2 | AI Under Audit

Bi-weekly newsletter for practitioners who need AI governance that works. Every issue gives you at least one concrete framework to make your AI systems more auditable, explainable, and regulator-ready.

linkedin.com/newsletters/exit-code-2-ai-under-audit

PRIMARY SOURCES

SR 11-7 Guidance on Model Risk Management, Federal Reserve/OCC, 2011
NIST AI 100-1 AI Risk Management Framework, NIST, 2023
NIST AI 600-1 Generative AI Profile, NIST, 2024
FHFA AB 2022-02 AI/ML Risk Management, FHFA, 2022

ABOUT THE AUTHOR

Paulo Cavallo holds a PhD in Public Policy and Political Economy and has spent five years developing and validating credit risk models under SR 11-7 regulatory frameworks. He builds AI systems for regulated industries and writes about what it takes to make them auditable, explainable, and production-ready.

LINKEDIN
linkedin.com/in/paulocavallo

PORTFOLIO
pmcavallo.github.io

NEWSLETTER
Exit Code 2 | AI Under Audit

"The frameworks are the map. The work is the territory."