

plzfix.ai Research Labs

Compliant by Design

Building generative AI systems the business can control, after SR 26-2

September 2026

Executive summary

In April 2026 the US banking agencies replaced SR 11-7, the guidance that had governed model risk for fifteen years, with a revision that does two things at once. It modernizes model risk management, trimming the discipline to materiality so that oversight follows inherent risk, exposure and use rather than a blanket inventory. And it draws a line: generative and agentic AI are not within its scope. The key finding of this paper follows from that line. Since 2023, banks had treated a generative AI system as one more model to validate. SR 26-2 separates two disciplines that had been conflated: the validation of quantitative models, which stays with model risk management, and the control of generative systems, which the bank now has to build for itself, as a dedicated set of operational risk controls under the three lines of defense. Generative AI is not deregulated; it is handed to the bank's broader risk management until an announced request for information produces something more specific. In Europe, the Digital Omnibus has moved high-risk obligations to December 2027; in the UK, the PRA keeps AI inside model risk. Three regimes, one practical consequence: the controls around a generative AI system have to be designed, owned and challenged by the bank itself, and the business, not the validation function, is now the owner.

This paper describes how to build generative AI systems that meet that expectation by construction. The principle is compliance by design: every control that a risk function would ask for after the fact is a component of the architecture from the start. The system is modular, so each component can be tested on its own. Its rules are written in plain language, so a process owner can sign them. Its logs and indicators are designed for a second line that is independent of the models doing the work. And its governance layer says who owns what, what the system may decide on its own, and on what evidence it earns more autonomy.

The architecture has seven layers: input management, the core model and its augmentation, output processing and guardrails, explainability and logging, monitoring and alerting, human oversight, and governance. The first three are the pipeline; the next two make it observable; the last two make it a control tower rather than a toolkit. For each layer the paper states what it does and what the second line can test.

What the market is doing confirms the direction. Large banks fold AI oversight into their existing risk structures rather than building parallel regimes; the consulting and vendor offers that exist validate the technical layer for risk and technology functions. The open space, and the subject of this paper, is the business-owned control layer above it.

Compliant by Design: Building Generative AI Systems the Business Can Control

Why this paper, and why now

For fifteen years, a bank deploying a quantitative model knew what to do: document it, validate it independently, monitor it, and keep the evidence for the examiner. SR 11-7 defined the discipline, and when generative AI arrived in 2023, banks applied it by analogy, because nothing else existed. That analogy held until 17 April 2026, when the Federal Reserve, the OCC and the FDIC replaced SR 11-7 with revised guidance, SR 26-2 for state member banks and OCC Bulletin 2026-13 for national banks. The new text is principles-based, risk-tiered and voluntary, and it contains one sentence that matters more than the rest for anyone building with large language models: generative AI and agentic AI models are novel and rapidly evolving and are not within the scope of this guidance.

That sentence has been misread in both directions. It does not mean generative AI is unregulated: the guidance directs institutions to use their broader risk management and governance practices to determine appropriate controls, and every other rule that never depended on SR 11-7 still applies, from third-party risk management to fair lending, customer communication and recordkeeping. Nor does it mean generative AI is now merely monitored, like an ordinary operational process. The letter says nothing about how; it delegates the how to the bank. The agencies have announced a request for information on model risk management and banks' use of AI, including generative and agentic AI, and it has not been issued.

This paper is written for the interval. It describes how to build a generative AI system so that the controls a risk function would ask for after the fact are part of the architecture from the start, and so that they can be owned by the business and challenged by the second line, whichever regime the bank falls under.

What changed: from SR 11-7 to SR 26-2

	SR 11-7 (2011)	SR 26-2 / OCC 2026-13 (April 2026)
Nature	Detailed, prescriptive supervisory guidance	Principles-based, risk-tiered, voluntary; most relevant above \$30 billion in assets
What is a model	A quantitative method turning input data into quantitative estimates	Same definition; scope now set by materiality
Core expectations	Independent validation, effective challenge, ongoing monitoring, governance	Preserved for models in scope, with proportionality
Generative and agentic AI	Not mentioned; applied by analogy	Explicitly outside scope; governed through broader risk management; RFI announced

	SR 11-7 (2011)	SR 26-2 / OCC 2026-13 (April 2026)
Validation independence	Organizational: a separate validation function	Rigor of the work over organizational chart; co-sourced validation defensible
BSA/AML models	Separate 2021 statement	Folded in; prior statement rescinded

The revision does two things for the models that remain in scope. It ties oversight to materiality: the depth of validation and monitoring now follows inherent risk, exposure, purpose and use, so that a low-materiality tool no longer earns the full treatment a capital model does. And it keeps the substance of SR 11-7 for what is material: conceptual soundness, outcomes analysis, effective challenge, governance. Model risk management is not weakened; it is focused.

For generative AI, the revision draws a line rather than a tier. The stated reason is only that these systems are novel and rapidly evolving. The reason that matters, and it is this paper's interpretation rather than the regulator's words, is that the validation toolkit was built for deterministic, quantitative models: a fixed specification to replicate, a quantifiable target to test outcomes against, a stable behavior to monitor. An open-ended generative system has none of these. Applying the toolkit by analogy, as banks did from 2023, produced validation reports that described the model and proved little about the system around it. By excluding generative AI, SR 26-2 ends the analogy and separates two disciplines that had been conflated: validating quantitative models, which stays where it was, and controlling generative systems, which becomes the bank's own job.

Three consequences follow for generative AI. First, the discipline moves. Independent validation in the SR 11-7 sense, conceptual soundness review, outcomes analysis and formal effective challenge by a validation function, is no longer required by name. What replaces it is the operational risk framework the bank already runs: risks and controls assessed and owned by the first line, challenged independently by the second, audited by the third. The function of independent challenge survives; the ritual of model validation does not. Second, the controls have to be built, because none exist by default. The guidance says institutions should use their broader risk management to determine appropriate controls; what makes that a must in practice is everything that still applies regardless of SR 11-7, third-party risk management, fair lending and disparate impact, customer communication, recordkeeping, cybersecurity, and the request for information that will ask what the bank did in the meantime. A generative AI system with no dedicated controls is not out of scope of supervision; it is out of scope of the one document that used to tell the bank what to do. Third, the reference framework moves too. Until the RFI produces guidance, the most complete voluntary standard is the Financial Services AI Risk Management Framework published by the Cyber Risk Institute with Treasury in February 2026, built with more than a hundred institutions and organized as roughly 230 control objectives.

Outside the US, the picture is different but not contradictory. In the European Union, the Digital Omnibus on AI, in force since 27 July 2026, defers the high-risk obligations of the AI Act for stand-alone systems, including credit scoring and employment, to 2 December 2027; transparency

obligations for systems that interact with people apply since 2 August 2026. In the United Kingdom, the PRA's SS1/23 keeps AI and machine learning inside model risk management, so the SR 11-7 logic still holds there. A bank operating across the three will need one architecture that satisfies all of them, which is an argument for building the controls into the system rather than into a jurisdiction-specific process around it.

Where generative AI now sits: operational risk and the three lines

A generative AI system in a bank does not estimate a probability of default or produce a valuation. It drafts, summarizes, retrieves, classifies, answers and, increasingly, acts. Its failure modes are not the ones model risk was built for: hallucination, prompt-dependent behavior, sensitivity to retrieval, drift when a third-party provider updates the model without notice, and, for agents, the combination of individually acceptable steps into an unacceptable workflow. The traditional validation toolkit, replication of a fixed specification and outcomes analysis against a quantifiable target, has nothing to replicate and no fixed target.

Operational risk has the right shape for this. It asks, for any process, what can go wrong, what controls prevent it, whether the controls are operating, and who is accountable. It separates the people who run the process from the people who challenge it. And it already governs the things a generative AI system is made of: data, third parties, technology, conduct, customer communication. What operational risk lacks, and what this paper supplies, is a description of the controls specific to generative AI and of the evidence the second line should demand.

Two functions still matter, then, but not the two the old discipline named. The first line owns and documents the controls: which inputs are allowed, what the model may see and say, what checks its output passes, when a person decides. The second line challenges those controls independently, with the rigor of validation but from outside the system: are the rules right, are they operating, did anything slip past them. Monitoring is what the second line does continuously; challenge is what it does to the design. The third line audits both.

Compliant by design: three drivers

Building for that split, rather than retrofitting it, rests on three principles.

Layered architecture. The system is modular, with a clear separation of concerns, so that the language model is one component within a larger, risk-managed framework rather than the framework itself. Each layer can be tested on its own, documented on its own, and replaced without disturbing the rest. This is what makes independent challenge possible: a second line cannot test a monolith, and a process owner cannot sign one.

Comprehensive documentation. Every component, decision and limitation is written down: the selection rationale for the model, the data used, the rules enforced, the thresholds set, the exceptions

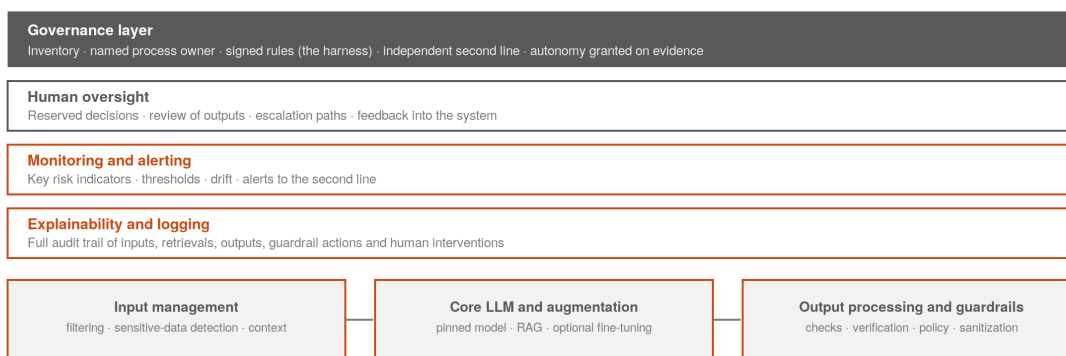
reserved for people. Documentation is not the price of compliance; it is the control that makes the other controls examinable.

Transparency and control. The system is observable and controllable at every layer. Observable means logged, measured and explained; controllable means bounded, with limits that hold whether or not the model behaves. A control the business cannot read is not a control the business owns.

The layered architecture

Compliant by design: the layered architecture

Each layer is a control the business signs and the second line tests



Red frames: the first line, inside the system. Grey: what the business and the second line do around it.
Below all of this: the technical control plane (identity, permissions, tool access). Plumbing, not governance.

Figure 1. The seven layers, and the split between the first line and the functions around it.

The seven layers fall into three groups. The pipeline, input management, the core model with its augmentation, and output processing, does the work. Explainability and logging, and monitoring and alerting, make the work observable. Human oversight and governance make it a control tower: they decide what the system may not do, who answers for it, and on what evidence it earns more room. Each layer is described below in the same form: what it does, and what the second line can test.

Input management and pre-processing

The first layer decides what reaches the model. Content filtering blocks harmful, biased or non-compliant input before inference. Prompt engineering structures requests for consistency and better performance, so that the same question produces the same shape of answer. Sensitive-data detection identifies and redacts personal and confidential information, which is where the data boundary of the system is enforced: what the model may see is a rule here, not a property of the model. Contextual enrichment adds the metadata and context the model needs to answer relevantly.

What the second line tests: that every filtering and redaction decision is documented and reproducible; that the data boundary matches the one the process owner signed; that sensitive-data detection catches a known set of cases; and that the prompt templates in production are the versions that were approved.

Core model and augmentation

The second layer is the model itself and what surrounds it. The base model carries a documented selection rationale, known limitations and a statement of suitability for the financial task; its version is pinned, so that a provider update does not silently change behavior, and it is isolated behind interfaces. Retrieval-augmented generation, where used, adds a curated knowledge base, an embedding model, retrieval logic and source attribution; the knowledge sources are versioned and quality-controlled, because an agent is only as reliable as what it retrieves. Fine-tuning, where used, is an isolated process on curated, documented and bias-audited datasets.

What the second line tests: the selection criteria and the version-control protocol; the fine-tuning data and method; and, for retrieval, the curation process and the retrieval accuracy, which can be measured independently against a reference set. The confidentiality tier of the model, from public providers to zero-retention agreements to on-premises deployment, is also a control here, and one the business should have chosen.

Output processing and guardrails

The third layer decides what leaves the system. Automated checks screen for hallucination, toxicity and bias. Factual verification checks outputs against trusted sources. Policy enforcement checks them against business rules and internal compliance policies. Output sanitization scrubs personal and sensitive information before anything is shown or sent.

What the second line tests: each guardrail directly, against a library of known problematic inputs, with pass/fail criteria and documented examples of what was intercepted. This is the most testable layer in the system, and the one where the second line can establish, on evidence, whether the controls operate.

Explainability and logging

The fourth layer makes the first three examinable after the fact. Audit trails log every input, every intermediate step including retrievals, every prompt, every raw and processed output, every guardrail action and every human intervention, with timestamps and identities. Explainability tools provide insight into why an output was produced, with the techniques used and their limits documented. Behavioral testing describes how the model responds across scenarios and provides the data for auditability.

What the second line tests: that the log is complete and immutable, that a given output can be traced back to its inputs and retrievals, and that the behavioral tests cover the scenarios the process owner identified as risky. Without this layer, the second line has opinions; with it, it has evidence.

Monitoring and alerting

The fifth layer tracks the system in production. A monitoring plan documents the metrics, data sources, frequency and protocols. Continuous monitoring follows key performance and key risk indicators specific to generative AI: drift in user queries and response patterns, guardrail interception

rates, escalation rates, the gap between the system's own assessment of success and the independent check, and usage that suggests misuse. Benchmarks and thresholds are documented; alerts fire on anomalies and breaches.

What the second line tests: that the indicators measure what the risk assessment said mattered; that thresholds are set and reviewed; that alerts reach someone with authority to act; and that the dashboard reflects the system's behavior rather than its self-reporting. This layer is the second line's screen, and it exists before go-live.

Human oversight

The sixth layer is where people sit. Human review procedures state which outputs qualified reviewers examine and how. Escalation paths state where problems go and how fast. Feedback mechanisms return human corrections to the system. Training programs prepare the people who work with the system. Above all, this layer holds the reserved decisions: what the system may never decide alone, the cases that always escalate, the amounts above which a person signs.

What the second line tests: reviewer qualifications, intervention records, the effectiveness of escalation, and whether the reserved decisions are actually reserved in the system rather than only in the policy. The question is not whether humans are involved but whether their involvement is proportionate to the risk, and enforced.

Governance: the bridge to the control tower

The seventh layer is where the technical modules meet the business, and it is the one that changes most under the new regime. A model inventory records every generative AI system in use, with its attributes and its named owner. A validation framework holds the quantitative and qualitative tools used to assess performance and compliance. A documentation repository centralizes design documents, test results and data sheets. Security testing covers adversarial testing and red-teaming.

To that, compliance by design adds three things. A process owner who signs the plain-language rules that the six layers below enforce: the objective, the standard of proof, the limits, the forbidden behaviors, the data boundary, the escalation thresholds. A second line that is independent of the models doing the work, in the model it uses, in the evidence it examines and in who owns what it measures. And a rule for autonomy: every system starts by suggesting, then acts with approval, then acts alone, and moves up only on evidence from the monitoring layer. The six layers below are the first line of defense. This layer is what makes them a control tower rather than a toolkit.

What the others are doing

Three observations from the market, as of September 2026.

Banks are not building parallel regimes. JPMorgan, BBVA and Standard Chartered have folded AI oversight into their existing risk structures rather than creating a separate governance track. Citigroup gave roughly 140,000 employees access to internal document tools with consent-order controls built

into the deployment from the start, and reports review times cut from sixty minutes to fifteen on document-heavy tasks. Morgan Stanley describes its approach as evaluation-based. Practitioner guidance has converged on the shape described in this paper: modular systems validated component by component, with an evaluator that does not share the executor's assumptions and monitoring for silent drift.

Consultants and vendors sell the technical layer. KPMG launched AI Assurance services in September 2025, covering model risk assessment, control testing, model validation and formal attestation. H2O.ai markets a model risk management framework for generative AI. Accenture, QuantumBlack and BCG X sell generative AI programs aligned with model risk. All of it is validation of the technical layer, sold to risk and technology functions.

The open space is the business-owned control layer. Nobody sells the plain-language rules a process owner signs, the second line independent of the models, or autonomy granted on evidence rather than at go-live. With generative AI now outside model risk guidance in the US, that layer is not optional; it is where the accountability has moved.

What to do in the interval

Four actions fit the period between SR 26-2 and whatever the RFI produces.

Rewrite the policy that cites SR 11-7. The inventory, the materiality tiers and the control design should be reassessed against the revised guidance, and generative AI systems should be moved into the operational risk framework with an owner, a risk assessment and a control set of their own.

Adopt a reference framework now. The FS AI RMF's control objectives are the most complete voluntary standard available, and mapping the seven layers to them produces the evidence the RFI, and the examiner after it, will ask for.

Build the next system compliant by design. Retrofitting the seven layers onto a system built without them is possible but slower and less convincing than building the next one with them, on one process, with a signed harness and a dashboard the second line reads before go-live.

Decide who signs. The single largest change is that the business now owns the controls. That requires a named process owner per system, a second line with the technical means to challenge, and a rule for autonomy that the executive committee has approved. None of this is a technology decision.

Conclusion

The regulator has, for the moment, stepped back from generative AI in the one discipline banks knew how to apply to it. That is not a reprieve. It moves the burden from a validation function that knew how to review a model to a business that has to say, in its own words, what its systems may do, and to a second line that has to test whether they do it. Systems built compliant by design make that possible: modular enough to test, documented enough to sign, observable enough to challenge, and governed by

people who answer for the outcome. The architecture in this paper is one way to get there. The alternative is to wait for the RFI, and to explain, when it comes, why the systems already in production were built without it.