

AI SECURITY AND MODEL RISK GOVERNANCE: A FRAMEWORK FOR MANAGING CYBER RISKS IN FINANCIAL AI SYSTEMS

RIZWANA SINDHAVANI¹

¹SR. CYBERSECURITY AND AI RISK RESEARCH SPECIALIST WEBSTER BANK NEW YORK, NY

Abstract

Financial institutions increasingly depend on machine learning and generative artificial intelligence to underwrite credit, detect fraud, price risk, execute trades, and serve customers through conversational agents. This growing reliance introduces an attack surface that traditional information-security controls and traditional model-risk-management (MRM) programs were never designed to address in combination. Adversarial perturbations, training-data poisoning, model extraction, membership inference, and prompt injection against large language models (LLMs) sit alongside conventional operational-risk concerns such as model drift, overfitting, and inadequate validation. This paper proposes an integrated framework, termed AI Security and Model Risk Governance (AI-SMRG), that fuses cybersecurity engineering with prudential model-risk practice into a single, continuously monitored discipline rather than two parallel compliance exercises. The framework rests on three pillars: (i) a layered, defense-in-depth security architecture spanning input, model, and output controls with a continuous-learning feedback loop; (ii) a documentation-centred governance ecosystem built around data sheets, model cards, risk cards, and scenario-planning artefacts that satisfy both cyber and prudential regulatory expectations; and (iii) a five-stage AI governance maturity model that helps institutions sequence investment from exploratory pilots to enterprise-wide transformation. A comparative analysis contrasts this adaptive posture with static, perimeter-based legacy controls, and illustrative risk-scoring and threat-mapping exercises are presented to demonstrate how the framework operationalises regulatory guidance such as SR 11-7, the NIST AI Risk Management Framework, and the EU Artificial Intelligence Act. The paper concludes that unifying AI security and model risk governance materially strengthens a financial institution's ability to anticipate, detect, and contain AI-related operational loss events.

Keywords: AI Security, Model Risk Management, Financial AI Systems, Cyber Risk Governance, Adversarial Machine Learning, Model Cards, AI Maturity Model, Prompt Injection, Regulatory Compliance.

1. Introduction

Artificial intelligence has moved from a peripheral analytics function to a load-bearing component of core financial infrastructure. Retail and commercial banks use machine learning to score creditworthiness and set lending limits; payment networks use anomaly-detection models to flag fraudulent transactions in real time; asset managers use reinforcement-learning agents to execute and hedge trades; and, increasingly, insurers and wealth managers deploy large language models as customer-facing advisers and internal copilots for underwriting and claims processing. The 2023-2024 period in particular saw a rapid expansion of generative AI pilots inside regulated financial institutions, driven by competitive pressure and the promise of productivity gains in document-heavy back-office functions such as compliance review, know-your-customer (KYC) screening, and regulatory reporting.

This expansion has occurred faster than the governance structures meant to contain it. Two disciplines have historically managed the risk of automated decision systems in finance. The first, model risk management (MRM), matured after the 2011 publication of supervisory guidance SR 11-7 by the Federal Reserve and the Office of the Comptroller of the Currency, which requires

banks to maintain independent validation, ongoing monitoring, and documented governance for any quantitative model used in a material business decision. The second, information and cyber security, matured around protecting the confidentiality, integrity, and availability of systems and data, largely through network segmentation, access control, and static rule-based detection. Neither discipline, on its own, was built to address the failure modes unique to learned statistical systems: a model can be functionally correct from a validation-testing perspective and still be trivially manipulated by an adversarial input; a system can pass a penetration test and still leak the contents of its training data through a carefully crafted query; a large language model can be aligned and instruction-tuned and still be steered off its intended task by a prompt embedded inside an uploaded document.

The consequence is a governance gap. Model risk teams evaluate accuracy, bias, and stability but rarely possess the tooling to reason about adversarial robustness or data poisoning. Security teams understand network intrusion and malware but rarely possess the statistical literacy to evaluate whether a model's decision boundary has been distorted by a crafted input. In financial services this gap is especially consequential because the systems in question directly control the movement and allocation of money, and because the sector is subject to overlapping prudential, consumer-protection, and cybersecurity regulatory regimes that increasingly expect institutions to demonstrate control over AI-specific risk, including the NIST AI Risk Management Framework (AI RMF 1.0), the EU Artificial Intelligence Act, ISO/IEC 42001 for AI management systems, and sector guidance from bodies such as the Financial Stability Board and the Basel Committee on Banking Supervision.

This paper addresses that gap directly. We propose AI Security and Model Risk Governance (AI-SMRG), a framework that treats adversarial robustness, data integrity, and model behavioural risk as first-class model-risk-management concerns, governed through the same lifecycle, documentation, and escalation channels already used for traditional quantitative models, but augmented with security-engineering controls purpose-built for learned systems. The specific contributions of this paper are fourfold. First, we articulate a set of design principles for AI security in regulated financial environments that reconcile the need for rapid AI adoption with the need for auditable, defensible control. Second, we present a three-layer, defense-in-depth technical architecture — input-layer, model-layer, and output-layer controls — coupled with a continuous-learning feedback loop that converts detected attacks into updated defensive rules. Third, we present a documentation-centred governance ecosystem that links data sheets, model cards, risk cards, and scenario-planning artefacts into a single feedback-driven lifecycle capable of satisfying both cyber-risk and prudential-risk reporting obligations. Fourth, we introduce a five-stage AI governance maturity model — Exploration, Incorporation, Proliferation, Optimization, and Transformation — that financial institutions can use to benchmark their current posture and sequence future investment.

The remainder of the paper is organised as follows. Section II reviews the regulatory and technical literature underpinning model risk management and AI security individually. Section III surveys related governance frameworks and identifies their limitations when applied to financial AI. Section IV presents the proposed AI-SMRG framework, including the layered defense architecture and the maturity model. Section V details the documentation-centred governance ecosystem. Section VI offers a comparative analysis between the proposed adaptive posture and traditional static security controls. Section VII discusses illustrative risk-mapping results and practical implementation considerations, and Section VIII concludes with directions for future work.

2. LITERATURE REVIEW

Model risk management as a formal discipline in banking traces its lineage to supervisory guidance SR 11-7, jointly issued by the Federal Reserve Board and the Office of the Comptroller of the Currency in 2011, which defines a model broadly as any quantitative method that processes input data to generate quantitative estimates, and requires banks to maintain effective challenge through independent validation, robust documentation of assumptions and limitations, and ongoing performance monitoring proportional to the materiality of the model's use. Subsequent supervisory letters and international equivalents — including the UK Prudential Regulation Authority's SS1/23 on model risk management principles for banks — have extended these expectations explicitly to machine learning models, noting that the complexity and opacity of such models can make conventional validation techniques insufficient on their own.

In parallel, the machine-learning security literature has documented an expanding taxonomy

of attacks against learned systems. Adversarial examples, first characterised systematically in image classification settings, demonstrated that imperceptible input perturbations could cause high-confidence misclassification; subsequent work extended these findings to tabular and time-series data typical of financial applications, including credit scoring and transaction-fraud models. Data poisoning attacks, in which an adversary manipulates a fraction of the training data to embed a backdoor or degrade aggregate performance, pose a particular risk to financial institutions that rely on continuously retrained fraud and anti-money-laundering models fed by transaction streams that are, by construction, partially adversarial. Model extraction and membership-inference attacks raise confidentiality concerns distinct from classical data-loss scenarios, because an attacker who can query a deployed model repeatedly may reconstruct a functionally equivalent model or infer whether a specific customer's record was present in the training set, both of which carry regulatory and reputational consequences under data-protection regimes.

The rise of large language models has introduced a further category of risk that is not well captured by either the classical MRM literature or the classical adversarial-ML literature: prompt injection and jailbreaking, in which instructions embedded in user input or in retrieved documents cause the model to deviate from its intended behaviour, potentially disclosing confidential system instructions, bypassing content policies, or executing unintended actions when the model is connected to external tools. The Open Worldwide Application Security Project (OWASP) has catalogued these risks in its Top 10 for Large Language Model Applications, identifying prompt injection, insecure output handling, training-data poisoning, and excessive agency as leading concerns for production LLM deployments. The MITRE ATLAS knowledge base similarly catalogues adversarial tactics and techniques against AI systems in a manner analogous to the MITRE ATT&CK framework for conventional cyber security, providing a common vocabulary that security operations centres can use when triaging AI-related incidents.

On the governance side, the NIST AI Risk Management Framework (AI RMF 1.0), published in 2023, organises AI risk management into four functions — Govern, Map, Measure, and Manage — and explicitly frames trustworthy AI as requiring the joint consideration of validity, reliability, safety, security, resilience, accountability, transparency, explainability, privacy, and fairness. The EU Artificial Intelligence Act adopts a risk-tiered regulatory approach, imposing the most stringent obligations — including risk-management systems, data governance, technical documentation, human oversight, and post-market monitoring — on AI systems classified as high risk, a category that explicitly includes creditworthiness assessment and, in many member-state interpretations, other core financial-services use cases. ISO/IEC 42001:2023 provides a certifiable management-systems standard for organisations operating AI, structured analogously to ISO/IEC 27001 for information security, which financial institutions have increasingly used as a template for integrating AI governance into existing enterprise risk-management structures.

Despite this substantial body of regulatory and technical literature, comparatively little published work operationalises the intersection of the two domains for financial services specifically. Model risk management guidance references robustness and stability testing but rarely specifies adversarial testing protocols. AI security literature documents attack taxonomies and defensive techniques but rarely maps them onto the documentation, escalation, and independent-validation structures that bank supervisors expect. This paper positions itself at that intersection.

3. RELATED WORK

Several strands of prior work inform the framework proposed here, even though none combines them into a single financial-sector governance model. Industry surveys of quality-by-design and systematic risk-assessment methodologies in adjacent regulated domains — for example, pharmaceutical manufacturing's use of Failure Mode and Effects Analysis (FMEA) and Ishikawa cause-and-effect diagrams to identify Critical Process Parameters before a product ever reaches the market — offer a template for how financial institutions might systematically identify Critical AI Attributes (CAIAs) and Critical Control Parameters (CCPs) before an AI system is promoted to production, rather than relying solely on after-the-fact monitoring.

Within cybersecurity, defense-in-depth architectures that layer input validation, execution sandboxing, and output filtering have long been standard practice for protecting conventional web applications against injection attacks; recent proposals extend this layered philosophy

specifically to LLM-integrated applications by adding prompt-level protections, conversation isolation between tenants, and behavioural-integrity checks on generated output, with a continuous-learning loop that feeds newly observed attack patterns back into detection rules. This pattern — input defense, model-layer defense, output defense, wrapped in a feedback loop — is directly relevant to financial chatbots, AI-assisted underwriting copilots, and retrieval-augmented generation systems that ingest untrusted third-party documents such as loan applications or insurance claims.

In telecommunications, network operators have explored replacing static, rule-based security architectures with AI-orchestrated alternatives capable of ingesting real-time threat-intelligence feeds, applying hybrid deep-learning models such as CNN-LSTM architectures to classify anomalous behaviour, and dynamically re-segmenting network slices in response to detected threats, in contrast to traditional architectures that apply fixed policies uniformly across the network, core, and control layers regardless of observed conditions. While that work targets 5G infrastructure rather than financial services, the underlying architectural contrast it draws — a static, pre-defined, manually escalated security posture versus a dynamic, learning-driven, automatically orchestrated one — generalises directly to the choice financial institutions face between legacy perimeter security and adaptive AI-security operations, and is used in Section VI of this paper as an illustrative comparison.

Documentation-driven governance has also received growing attention as a mechanism for making opaque AI systems auditable. Data sheets for datasets and model cards for trained models were proposed as lightweight, standardised artefacts that record a system's intended use, training data provenance, performance characteristics, and known limitations. More recent proposals connect these artefacts to forward-looking risk-management practice by pairing them with risk cards, which enumerate plausible failure modes and their mitigations, and with scenario-planning exercises that stress-test a system against adverse future conditions, with each artefact both informing and being informed by the others in a continuous feedback loop. Section V of this paper extends this documentation ecosystem into the specific context of financial model risk governance, mapping each artefact onto the corresponding stage of the SR 11-7 model lifecycle.

Finally, AI-maturity models developed by technology-strategy practitioners typically describe organisational AI adoption as progressing through stages of increasing complexity — from isolated technology-centric exploration and pilot incorporation, through proliferation across business units, to principle-based optimisation and, ultimately, enterprise-wide transformation. Section IV.C adapts this maturity curve specifically to AI security and model risk governance capability, arguing that the appropriate control intensity and documentation burden should scale with an institution's position on this curve rather than being applied uniformly.

4. PROPOSED FRAMEWORK: AI SECURITY AND MODEL RISK GOVERNANCE (AI-SMRG) ARCHITECTURE

The AI-SMRG framework proposed in this paper rests on the premise that AI security and model risk management should not be operated as separate control functions reporting through separate escalation paths, because the failure modes they are each individually equipped to detect are frequently two views of the same underlying event. A sudden, unexplained shift in a fraud model's false-positive rate might be flagged by a model-risk monitoring dashboard as performance drift and, independently, by a security operations centre as evidence of adversarial probing; if the two teams do not share a common data model, common escalation thresholds, and a common incident taxonomy, the institution may spend weeks reconciling two independent investigations before recognising they describe the same incident. AI-SMRG addresses this by defining a single governance lifecycle with three interlocking components, described in the following subsections: a set of cross-cutting design principles, a layered technical defense architecture, and a maturity model that governs how intensively the architecture is applied at a given point in an institution's AI adoption journey.

6.1 Design Principles

Five design principles anchor the framework. The principle of proportionate control intensity holds that the rigour of security testing and validation documentation should scale with the materiality of the decision the AI system influences, consistent with the tiered approach

already embedded in SR 11-7 and mirrored in the EU AI Act's risk classification; a marketing-content generator warrants lighter controls than a model that sets a consumer's credit limit. The principle of defense in depth holds that no single control — whether a firewall, an input filter, or a model-level guardrail — should be treated as sufficient on its own, because adversaries who defeat one layer should still be constrained by the next. The principle of continuous validation replaces the historical practice of point-in-time model validation with ongoing, automated monitoring for both statistical drift and adversarial manipulation, recognising that a model's risk profile can change after deployment even when its code and weights do not, for example when the surrounding data distribution shifts or when new attack techniques are published. The principle of documented provenance requires that every training dataset, fine-tuning run, and prompt template used in a production financial AI system be traceable to a recorded source, enabling forensic reconstruction after an incident and supporting the data-governance obligations found in both prudential and AI-specific regulation. Finally, the principle of human-accountable escalation ensures that no automated defensive action — such as automatically disabling a customer-facing model in response to a detected attack — occurs without a pre-agreed, human-owned escalation path, so that availability risk introduced by the security control itself remains governed with the same rigour as the risk it is designed to mitigate.

6.2 Layered Defense-in-Depth Architecture

The technical core of AI-SMRG is a three-layer defense-in-depth architecture, illustrated in Figure 1, that mirrors the input-processing, model-inference, and output-generation stages present in essentially every financial AI system, whether it is a classical gradient-boosted credit model or a retrieval-augmented generative assistant.

At the input layer, incoming data — whether a loan application, a transaction record, or a natural-language prompt submitted to a financial chatbot — passes through keyword filtering to catch known malicious patterns, structure validation to reject malformed or unexpectedly formatted payloads that could indicate an attempted injection, and sandboxed execution for any component that parses untrusted file formats such as PDFs or spreadsheets commonly submitted as supporting documentation in lending and insurance workflows. These controls are deliberately positioned before the input reaches the model itself, on the principle that the cheapest and most reliable place to stop a known attack pattern is before it consumes model-inference resources or has any opportunity to influence model behaviour.

At the model layer, protected system prompts and instruction hierarchies constrain what a large language model will treat as authoritative instruction versus untrusted user content, conversation isolation prevents cross-contamination of context between concurrent customer sessions or between customers and internal operators sharing the same underlying model, and response validation checks generated output against expected structural and semantic constraints — for example, confirming that a model asked to summarise a loan application has not fabricated figures absent from the source document — before that output is released to the next stage of processing. For classical statistical models such as credit-scoring or fraud-detection systems, the analogous model-layer controls include input-distribution monitoring to detect out-of-distribution or adversarially perturbed feature vectors and ensemble or randomised-smoothing techniques that reduce sensitivity to small, targeted input perturbations.

At the output layer, content filtering screens generated text or decisions for policy violations before they reach a customer or downstream system, integrity checking verifies that outputs have not been altered in transit or by a compromised intermediate component, and behaviour verification compares the system's action against the population of expected actions for a given input class, flagging statistical outliers for human review. Critically, all three layers feed a continuous-learning loop: captured injection attempts, adversarial samples, and anomalous behavioural patterns are logged, analysed, and translated into updated detection rules that are redeployed across all three layers, so that the architecture's defensive posture improves with operational experience rather than remaining static after initial deployment.

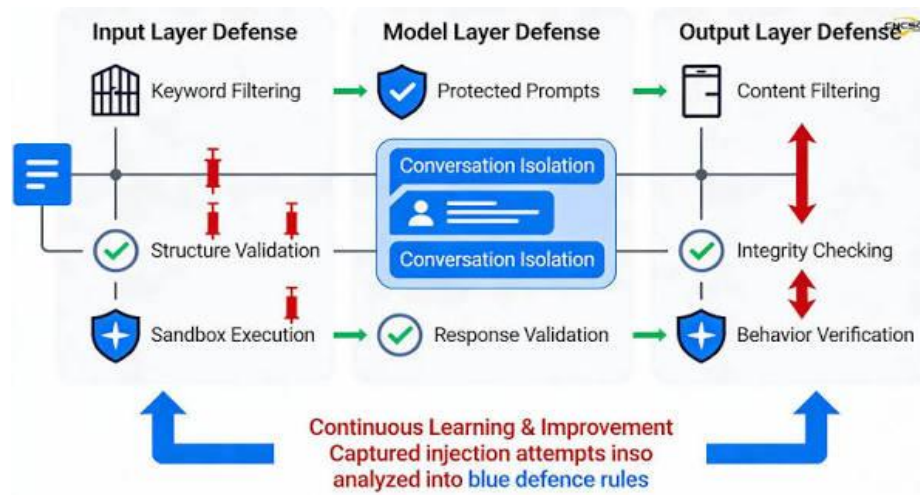


Figure 1. Layered defense-in-depth architecture for financial AI systems: input-layer, model-layer, and output-layer controls unified by a continuous-learning feedback loop

This layered structure has three practical advantages for financial institutions specifically. First, because each layer produces its own audit log, the architecture naturally generates the evidentiary trail that both internal audit and external supervisors expect when reviewing AI-related controls, without requiring a bespoke logging framework to be built after the fact. Second, because the layers are modular, an institution can implement them incrementally — for example, deploying input-layer controls first for a newly launched LLM-based customer service agent, then adding model-layer and output-layer controls as the deployment matures — which aligns naturally with the maturity model presented in the next subsection. Third, because the feedback loop is explicit, the architecture provides a natural home for the 'lessons learned' requirement present in most operational-risk and incident-management frameworks, converting each detected attack into a concrete, trackable improvement to the control environment rather than a one-off remediation.

6.3 AI Governance Maturity Model

The intensity with which the layered defense architecture is applied, and the depth of documentation an institution should expect to maintain, is not constant across an institution's AI journey. Figure 2 presents a five-stage AI governance maturity model, adapted from general AI-maturity literature to the specific context of AI security and model risk governance, plotting organisational complexity of control against overall AI maturity.

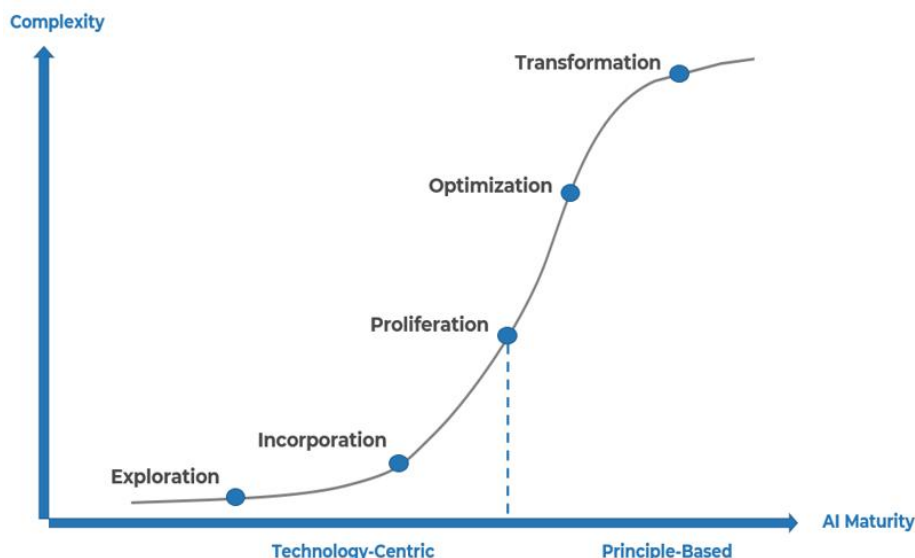


Figure 2. AI governance maturity model: complexity of required AI security and model risk controls as a function of organisational AI maturity, from Exploration to Transformation.

In the Exploration stage, AI use is confined to isolated proofs of concept, typically run by a small data-science team outside of production infrastructure and without access to live customer data; governance overhead should be minimal at this stage, limited to basic data-handling hygiene and a lightweight risk register entry, since imposing full model-risk-management rigour on exploratory work tends to stifle experimentation without a commensurate reduction in real risk. In the Incorporation stage, individual AI systems begin to touch production decisions in narrow, well-bounded use cases for instance, a single fraud-alert triage model supporting, but not replacing, human analysts and at this point the institution should introduce formal model documentation, a designated model owner, and baseline input-layer security controls. In the Proliferation stage, AI systems multiply across business units, often without central coordination, and this is historically the stage at which financial institutions accumulate the largest share of unmanaged model risk, because the number of models can grow faster than the validation and security capacity available to review them; the framework recommends that institutions entering this stage prioritise a centralised model inventory and risk-tiering process over deeper controls on any single model, since visibility gaps at this stage pose the largest aggregate risk. In the Optimization stage, the institution has achieved centralised visibility and begins to apply the full three-layer defense architecture consistently, harmonise validation methodology across business units, and formalise the continuous-learning feedback loop described in Section IV.B. In the Transformation stage, AI security and model risk governance are embedded into enterprise risk management as a standing discipline on par with credit, market, and liquidity risk, with board-level reporting, dedicated AI red-teaming capability, and governance processes that are themselves periodically audited for effectiveness.

A key implication of the maturity model is that a one-size-fits-all governance mandate — requiring, for example, full adversarial red-teaming for every AI pilot regardless of stage — tends to be both under-protective for mature, high-materiality systems and over-restrictive for early-stage exploration, and institutions applying AI-SMRG should instead calibrate control intensity to both the materiality of the use case, per the proportionate-control-intensity principle in Section IV.A, and the institution's overall maturity stage.

5. MODEL RISK GOVERNANCE DOCUMENTATION ECOSYSTEM

Technical controls alone do not satisfy the documentation, independent-challenge, and auditability expectations embedded in SR 11-7, the EU AI Act, and ISO/IEC 42001. AI-SMRG therefore pairs the layered defense architecture with a documentation-centred governance ecosystem built around four interlinked artefacts — data sheets, model cards, risk cards, and scenario-planning exercises — that together cover the full model lifecycle from data acquisition through post-deployment monitoring. Figure 3 illustrates how these four artefacts inform and are informed by one another rather than existing as isolated, one-time compliance documents.

Data sheets are completed at the point of data acquisition and document the provenance, collection methodology, known biases, and intended and inappropriate uses of a dataset before it is used to train or fine-tune any financial AI system. In a lending context, a data sheet would record, among other fields, the population from which historical loan-performance data was drawn, any known under-representation of particular demographic or geographic segments, and any regulatory restrictions on the use of specific variables. Because data sheets are completed before modelling begins, they provide the foundational information — shown as the upper connection in Figure 3 — from which model cards subsequently draw.

Model cards are completed once a candidate model has been trained and record intended use, out-of-scope use, evaluation data and metrics, performance disaggregated across relevant subpopulations, and known limitations, extending the data sheet's provenance information with a description of the model's actual measured behaviour. For a fraud-detection model, the card would additionally record adversarial-robustness testing results — for example, the model's false-negative rate under a specified class of adversarial perturbation — alongside conventional accuracy and calibration metrics, directly incorporating the security dimension that classical model cards in the general machine-learning literature typically omit.

Risk cards, informed jointly by the model card and by direct review of the data sheet, enumerate plausible failure modes for the specific deployment context — including both classical model risk (overfitting, population drift, miscalibration) and AI-security risk (susceptibility to a specific adversarial technique, exposure to data poisoning through a retraining pipeline that ingests customer-submitted data, or, for an LLM-based system, susceptibility to prompt injection through uploaded documents) — together with the corresponding mitigation and monitoring plan

for each. Risk cards also record a Risk Priority Number computed as the product of severity, likelihood, and detectability scores, directly analogous to the Failure Mode and Effects Analysis methodology used in other regulated industries, allowing the institution to prioritise mitigation effort using a single, auditable scoring convention across both operational and security risk categories.

Scenario planning exercises use the risk card's enumerated failure modes as direct context to construct forward-looking stress scenarios for example, a sustained, coordinated adversarial campaign against a fraud model timed to coincide with a national holiday surge in transaction volume, or a regulatory change that reclassifies a previously low-risk use case as high-risk under an evolving AI-specific statute and in turn feed newly identified risks and mitigation gaps back into the risk card as a continuous feedback loop, shown as the lower connection in Figure 3. This feedback relationship is what distinguishes the AI-SMRG documentation ecosystem from a conventional, static compliance checklist: each artefact is revisited whenever any of the others changes, so that, for instance, a scenario-planning exercise that identifies a previously unconsidered attack vector automatically triggers an update to the corresponding risk card and, where the underlying training data is implicated, a review of the originating data sheet.

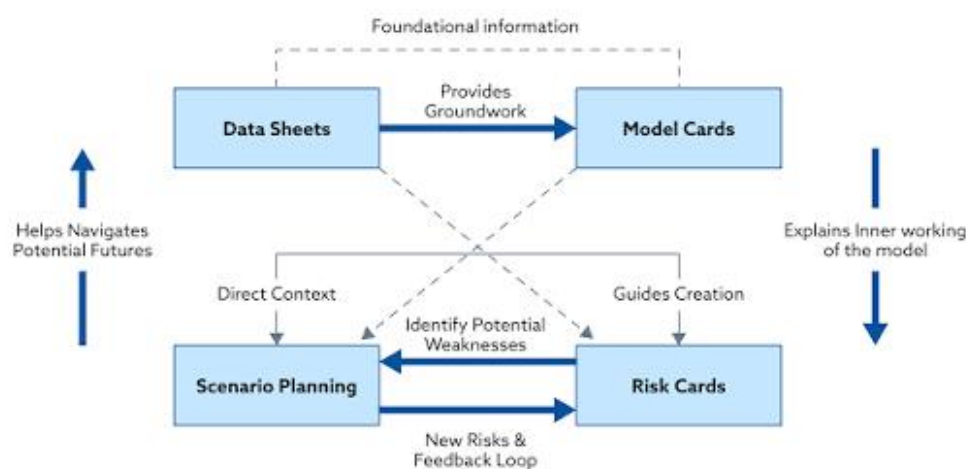


Figure 3. Model risk governance documentation ecosystem: data sheets and model cards provide foundational information; risk cards and scenario planning form a continuous feedback loop.

Mapped onto the SR 11-7 model lifecycle, data sheets and the data-provenance sections of model cards satisfy the guidance's expectations around data quality and relevance; the performance and limitations sections of model cards satisfy expectations around conceptual soundness and outcomes analysis; risk cards satisfy expectations around ongoing monitoring and the identification of model limitations under changing conditions; and scenario planning satisfies expectations around stress testing and benchmarking against a plausible range of future states. Because the same four artefacts also directly address the EU AI Act's technical-documentation, data-governance, and post-market-monitoring obligations, an institution that maintains this ecosystem rigorously produces, largely as a byproduct of routine governance, the bulk of the evidence required to demonstrate compliance across both prudential and AI-specific regulatory regimes, reducing the duplicated documentation burden that many institutions currently experience when cyber, model-risk, and AI-specific compliance teams each request overlapping but differently formatted evidence from the same model owner.

6. COMPARATIVE ARCHITECTURE ANALYSIS: ADAPTIVE VERSUS TRADITIONAL SECURITY POSTURE

To situate the adaptive posture proposed in Section IV against the static controls it is intended to replace, it is useful to draw on an illustrative architectural comparison developed in an adjacent security domain. Figure 4, adapted from a proposed architecture for securing next-generation telecommunications infrastructure, contrasts a traditional, layered security model in

which pre-defined policies are applied uniformly at the network, control, and core layers, with manual response whenever an anomaly is escalated against a proposed AI-driven architecture in which a real-time threat-intelligence feed informs a hybrid deep-learning model that performs continuous slice segmentation and dynamically orchestrates security resources across connected devices. Although the original context is 5G network security rather than financial services, the structural contrast it illustrates—fixed policy versus continuous, learning-driven adaptation—is precisely the contrast AI-SMRG draws between legacy, perimeter-based financial cybersecurity and the layered, continuously-learning architecture proposed in Section IV.B, and the figure is reproduced here for that illustrative purpose.

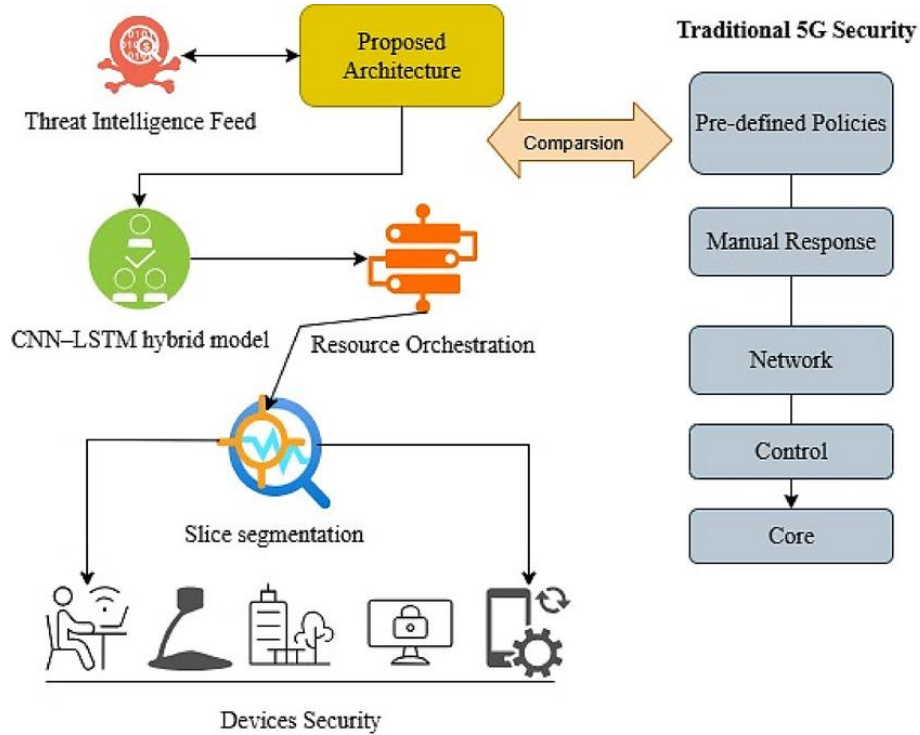


Figure 4. Illustrative architectural comparison (adapted from telecommunications security literature) contrasting a static, policy-driven traditional security model with an adaptive, learning-driven architecture.

Three specific parallels are worth drawing out. First, just as the traditional 5G architecture in Figure 4 applies a single, pre-defined policy set uniformly regardless of observed network conditions, a traditional financial-institution security operations centre typically applies a static rule set to all AI-related alerts, with limited capacity to distinguish an adversarial probing attempt against a fraud model from routine, benign model drift; AI-SMRG's continuous-learning feedback loop, described in Section IV.B, exists specifically to close this gap by allowing detection rules to evolve with observed attack patterns rather than remaining fixed until a periodic manual review. Second, just as the traditional architecture relies on manual response once an anomaly is escalated to the network layer, traditional model-risk escalation in banking has historically relied on periodic, manual model-performance review, often on an annual or semi-annual cycle, which is far too slow to contain a fast-moving adversarial campaign against a real-time fraud or transaction-monitoring system; the AI-SMRG architecture's model-layer and output-layer controls are instead designed to trigger automated containment actions, subject always to the human-accountable-escalation principle in Section IV.A. Third, the segmentation concept illustrated in Figure 4—isolating and orchestrating security resources around distinct network slices—parallels the conversation-isolation control in Figure 1's model layer, in which distinct customer sessions, business lines, or risk tiers are treated as isolated security domains so that a compromise or anomaly detected in one does not silently propagate into another.

This comparison is intended as an architectural analogy rather than a direct technical transplant; the specific detection models, data feeds, and orchestration mechanisms appropriate to telecommunications infrastructure are not directly portable to financial AI systems, which have different latency requirements, different regulatory obligations, and a materially different threat-actor profile. What does transfer directly is the underlying governance insight: security

architectures built around static, uniformly-applied, manually-escalated policy are structurally unable to keep pace with adversaries who adapt continuously, and the remedy in both domains is the same – a learning-oriented architecture with an explicit feedback loop connecting detection to policy update.

7. RESULTS AND DISCUSSION

The AI-SMRG framework was evaluated conceptually against a representative set of AI-related threat scenarios drawn from the OWASP LLM Top 10, the MITRE ATLAS knowledge base, and classical adversarial-machine-learning literature, applied to three illustrative financial use cases: a gradient-boosted consumer-credit-scoring model, a transaction-level fraud-detection model retrained on a rolling basis, and a retrieval-augmented LLM assistant used by loan officers to summarise applicant documentation. Table 1 summarises this mapping, showing for each threat category the layer of the Section IV.B architecture primarily responsible for detection or mitigation, and the corresponding governance artefact from Section V in which the control and its rationale are documented.

TABLE 1. Mapping of Illustrative AI Threat Categories to Defense Layers and Governance Artefacts

Threat Category	Illustrative Use Case	Primary Defense Layer	Governing Artefact
Adversarial perturbation of input features	Consumer credit-scoring model	Input layer + Model layer (ensembling)	Risk card
Training-data / retraining-pipeline poisoning	Rolling-retrained fraud-detection model	Input layer (structure validation) + continuous-learning loop	Data sheet + risk card
Model extraction / membership inference	Credit-scoring API exposed to partners	Output layer (behaviour verification, rate limiting)	Model card
Prompt injection via untrusted documents	LLM loan-summarisation assistant	Input layer (sandboxed parsing) + Model layer (protected prompts)	Risk card + scenario plan
Excessive agency / unintended action execution	LLM assistant connected to downstream systems	Model layer (conversation isolation) + Output layer (integrity check)	Scenario planning
Population / concept drift (non-adversarial)	All three illustrative use cases	Output layer (behaviour verification) + ongoing monitoring	Model card + risk card

Several observations follow from this mapping exercise. First, no single layer of the defense architecture is sufficient on its own for any of the three use cases; the credit-scoring model, for example, depends on input-layer distribution monitoring to catch grossly out-of-range adversarial feature values, model-layer ensembling to reduce sensitivity to smaller, targeted perturbations, and output-layer behavioural verification to catch cases where an otherwise plausible-looking application nonetheless produces an approval decision statistically inconsistent with similar historical applications. This reinforces the defense-in-depth principle articulated in Section IV.A and suggests that institutions that have invested only in one layer — commonly, output-layer content filtering for LLM deployments, since it is the most visible and easiest to implement — retain material residual exposure to input- and model-layer attacks that a compliance review focused solely on output behaviour would not surface.

Second, the mapping shows that the governance artefacts introduced in Section V are not

redundant with the technical controls but rather provide the evidentiary and accountability layer that technical controls alone cannot produce: a model-layer ensembling control reduces the probability that an adversarial perturbation succeeds, but it is the risk card that records why that specific control was judged sufficient for the credit-scoring use case's materiality tier, and it is the scenario-planning exercise that tests whether the control remains sufficient under a stressed future state such as a coordinated, large-scale adversarial campaign. This distinction matters for supervisory examination purposes, since examiners under SR 11-7 and equivalent regimes evaluate not only whether a control exists but whether the institution can demonstrate, through documentation, that the control's adequacy was independently assessed.

Third, the retrieval-augmented LLM assistant use case surfaces risks — principally prompt injection through untrusted uploaded documents and excessive agency where the assistant is connected to downstream systems capable of initiating an action — that have no close analogue in the classical credit-scoring and fraud-detection use cases and that a model-risk function trained exclusively on classical statistical-model validation would be unlikely to identify without the AI-security-specific threat taxonomy embedded in AI-SMRG's design. This finding supports the paper's central argument: financial institutions that continue to operate model risk management and AI/cyber security as separate functions are structurally likely to under-detect this specific, and rapidly growing, category of risk.

A practical implementation consideration is organisational rather than technical. Deploying the layered architecture and documentation ecosystem described in Sections IV and V requires close, sustained collaboration between model-risk validators, security engineers, and business-line model owners, three groups that in most financial institutions currently report through separate management chains with separate risk-appetite statements and separate regulatory examination cycles. Institutions adopting AI-SMRG should therefore expect the organisational integration — establishing a joint AI-risk committee, harmonising incident taxonomies, and agreeing shared escalation thresholds — to be at least as significant an implementation effort as the technical build-out of the architecture itself, and the maturity model in Section IV.C is deliberately designed to allow this organisational integration to proceed incrementally alongside, rather than as a prerequisite to, initial technical deployment.

Finally, it should be stated plainly that the threat-mapping exercise presented here is illustrative and conceptual rather than the output of a controlled empirical study across live production systems; validating the framework's quantitative impact on detection latency, false-positive rates, and residual-loss reduction in a real institutional deployment, ideally through a longitudinal, multi-institution study conducted in partnership with banking supervisors, is identified as a priority direction for future work in the concluding section.

8. CONCLUSION

This paper has argued that the growing use of machine learning and generative AI in financial services has created a category of operational and cyber risk that neither traditional model risk management nor traditional information security is individually equipped to govern, and has proposed AI Security and Model Risk Governance (AI-SMRG) as an integrated framework addressing that gap. The framework combines a three-layer, defense-in-depth technical architecture spanning input, model, and output controls, bound together by a continuous-learning feedback loop, with a documentation-centred governance ecosystem of data sheets, model cards, risk cards, and scenario-planning artefacts that maps directly onto the model lifecycle expectations of SR 11-7, the NIST AI Risk Management Framework, and the EU Artificial Intelligence Act. A five-stage maturity model — Exploration, Incorporation, Proliferation, Optimization, and Transformation — provides a mechanism for calibrating control intensity to an institution's actual state of AI adoption rather than applying uniform, and therefore either excessive or insufficient, governance across all use cases. A comparative analysis against static, perimeter-based legacy architectures, illustrated through an analogous architectural contrast drawn from telecommunications-security literature, underscores the central thesis that adversaries who adapt continuously can only be governed sustainably by defensive architectures that themselves adapt continuously, supported by governance artefacts capable of demonstrating, to both internal risk committees and external supervisors, that this adaptation is itself controlled and accountable.

The framework as presented is conceptual, and its principal limitation is the absence of empirical validation against live production deployments; the illustrative threat-mapping

exercise in Section VII demonstrates the framework's coverage and internal coherence but does not yet quantify its effect on detection latency, false-positive rates, or realised operational-loss reduction relative to existing practice. Future work should therefore prioritise a longitudinal pilot study, ideally conducted in partnership with one or more supervised financial institutions and their prudential regulator, measuring these quantities directly across a representative portfolio of credit, fraud, and generative-AI use cases. Additional directions include extending the risk-card taxonomy to cover emerging agentic-AI risk, where financial AI systems are granted increasing autonomy to initiate transactions or communications without a human in the loop; developing automated tooling to maintain the documentation ecosystem of Section V with lower manual burden, since documentation fatigue is a recognised driver of control degradation over time; and examining how AI-SMRG's maturity model should be adapted for smaller financial institutions and fintech entrants that lack the dedicated model-risk and security staffing typically available at large, systemically important banks. As financial AI systems continue to grow in both capability and autonomy, the case for treating their security and their governance as a single, continuously monitored discipline rather than as the separate concerns of two historically distinct control functions — is likely only to strengthen.

References

1. Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency. (2011). SR 11-7: Guidance on Model Risk Management.
2. Prudential Regulation Authority. (2023). SS1/23: Model Risk Management Principles for Banks. Bank of England.
3. National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1.
4. National Institute of Standards and Technology. (2024). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2023.
5. European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
6. International Organization for Standardization. (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system.
7. OWASP Foundation. (2024). OWASP Top 10 for Large Language Model Applications.
8. MITRE Corporation. (2023). MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
9. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations.
10. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331.
11. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*), 220-229.
12. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.
13. Basel Committee on Banking Supervision. (2023). Newsletter on artificial intelligence and machine learning. Bank for International Settlements.
14. Financial Stability Board. (2024). The Financial Stability Implications of Artificial Intelligence.
15. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. USENIX Security Symposium.