The Alan Turing Institute

GenAI Model Risk Management and Governance in financial services: from principles to practice

This work is part of the FAIR Programme (Framework for Responsible Adoption of Artificial Intelligence in the Financial Services Industry), an EPSRC-funded Prosperity Partnership EP/V056883/1 at The Alan Turing Institute and the Partnership on AI in Finance (PAIF), a collaboration between the FAIR Programme and our partner members listed below.

The opinions expressed in this publication are those of the contributing authors. They do not purport to reflect the opinions or views of their organisations or its members.

Programme Management:

Isabelle Malcolm, Turing
Tony Zemaitis, Turing

Contributing Authors (Alphabetically Ordered):

Adeline Pelletier, Mastercard
Aleksandra Lukaszewicz, Aviva
Alpay Sabuncuoglu, Accenture
Anita Khadka, University of Warwick
Archie Jones, BNY
Carlo Vicentini, HSBC
Carsten Maple, Turing
John Hearty, Mastercard
Jovan Powar, Turing
Julian Paul Phillips, HSBC
Karina Pretto, HSBC
Kirsten Mycroft, BNY
Lukasz Szpruch, Turing
Søren Mørk, Accenture
Nadeem Chaudhry, Aviva
Oluseyi Akinyede, Aviva
Richard Boorman, Mastercard
Riya Tapwal, University of Warwick
Tom Niven, Accenture
William Flanagan, BNY

Partnership on AI in Finance member organisations:

Contents

1 Introduction

This report presents current best practices for operationalising Model Risk Management (MRM) frameworks, such as **SR 11-7** and **SS1/23**, within artificial intelligence (AI) workflows. It draws on a study conducted with the Partnership on AI in Finance (PAIF) partners and is informed by the latest research, regulatory guidance, and industry developments.

Financial institutions (FIs) form part of every nation's critical national infrastructure and must maintain resilience both individually and across the financial system. Modern FIs make extensive use of models for activities such as credit risk assessment, capital allocation, stress testing, and trading [1, 2]. To ensure these models are reliable, transparent, and well-controlled, FIs rely on MRM frameworks. The key supervisory references for this are the U.S. Federal Reserve's **SR 11-7** [3] and the Bank of England's Prudential Regulation Authority supervisory statement **SS1/23** [4]. These frameworks define principles for identifying, validating, and governing models across their lifecycle, emphasising *reliability, performance validation*, and *resilient governance*.

Introduction

The adoption of GenAI in financial services is largely driven by advances in foundation models capable of automating complex analytical and cognitive tasks [5]. FIs are exploring these technologies for diverse applications, including document summarisation, code generation, customer interaction, compliance monitoring, and information retrieval across risk, audit, and legal functions [6]. These applications promise efficiency gains and richer analytical insight, but they also introduce new dimensions of risk. Unlike traditional machine learning (ML) models, GenAI systems serve a wide range of use-cases thanks to the breadth of their potential behaviours and ability to produce qualitative outputs, through large-scale foundation models provided and managed by third-party vendors [7]. This creates uncertainty around validation, reproducibility, and output integrity [6], while their reliance on dynamic data pipelines and proprietary APIs raises questions about how established MRM principles can be extended to govern GenAI applications in practice.

Supervisory attention to the risks of GenAI systems in finance has intensified. The Financial Stability Board has called for enhanced monitoring of AI adoption and evaluation of whether current regulatory frameworks adequately address AI- and GenAI-specific risks [7]. The Bank for International Settlements highlights hallucination and interpretability risks as distinctive to GenAI, suggesting that existing principles may require augmentation [6]. Similarly, the European Central Bank warns that rapid AI adoption could heighten concentration risk, opacity, and systemic interdependence [8]. While frameworks such as SR 11-7 and SS1/23 provide a robust foundation for managing model risk, the assumptions of stability, interpretability, and bounded data inputs are increasingly challenged by the scale and dynamism of GenAI systems [3, 4].

This report examines the required adaptations to existing MRM practices that FIs should consider in order to effectively operate and govern GenAI-based systems in financial services, by drawing on the experiences of PAIF partners. PAIF partners have been proactive in adapting their risk management and governance practices to meet the challenges of GenAI systems, enabling them to innovate and scale GenAI adoption while ensuring their systems' reliability, performance, and resilience. We drew on this experience through roundtables, workshops, and structured interviews, all supported by literature review. We distil partners' experiences and reflections into discussion, guidance, and best practices to support FIs looking to adapt their MRM practice to GenAI models, or establish a wider AI Governance practice.

The report is structured as follows. Section 2 presents a focused and non-exhaustive discussion of the risks posed by RAG (Retrieval-Augmented Generation) architectures [9]: a GenAI system design pattern which has already seen substantial uptake in financial services, and which serves as a good primer on the challenges posed by GenAI models' broader risk profile compared to traditional financial models. In Section 3 we share two illustrative case studies, detailed examples of how partners are already adopting GenAI in practice and their initial successes in adapting and operationalising the principles of MRM. To support institutions in adapting their existing organisational processes to meet the challenges of AI, Section 4 discusses how FIs can extend their existing governance structures and practices to effectively govern GenAI models, and support a practice of AI Governance that goes beyond traditional Model Risk. Finally, Section 5 concludes with key recommendations for FIs as they continue to adapt to the evolving AI landscape.

2 The RAG risk landscape

The risk landscape of GenAI extends beyond that of traditional statistical models. For example, GenAI models in FIs are usually built on third-party GenAI components such as LLMs, created by vendors which serve a wide range of clients and consumers. This creates dynamism in the system as component artefacts evolve through continual versioning, model deprecations, and back-end routing [10, 11, 12, 13]. This dynamism is amplified by the fact that GenAI models are increasingly architected as pipelines, with patterns such as retrieval-augmented generation (RAG), tool-use, and agents; more complex interdependencies can produce emergent failure modes. Capturing these effectively demands a structured and multidisciplinary risk-elicitation process involving relevant business, technical, and risk functions [14, 15, 16, 17]. Therefore, this section examines the emerging risk dimensions of GenAI, focusing on RAG-architecture models, which are currently being adopted widely in FIs. We present a non-exhaustive examination of risks across four categories: **data, vendor, architecture,** and **human factors**. While this discussion focuses on RAG architectures, many of these risk dimensions also generalise to other GenAI system designs, including agentic workflows.

2.1 Data risks

RAG systems depend on data that are dynamic, heterogeneous, and often semi-structured [14]. The quality, provenance, and legality of this data directly determine model utility [6, 18, 19]. In contrast to traditional models, which operate on curated and relatively more stable datasets, RAG systems rely on corpora that may be updated at much higher frequency, or provided at runtime, such as reports, filings, or market summaries. If there are failures in data quality or data governance (e.g., due to incomplete coverage, outdated sources, or unclear provenance), these issues can propagate throughout the retrieval and generation pipeline [20]. This can influence factual accuracy, completeness and compliance [19]. We will now discuss each of the major dimensions of data risk, which must be closely monitored in RAG GenAI models.

2.1.1 Document base quality

RAG systems are often employed to process qualitative and unstructured corpora such as PDFs, reports, and message logs [14], in contrast to traditional statistical models whose data inputs tend to be relatively more structured. In RAG use-cases, these sources are often highly heterogeneous in format and semantics, introducing new uncertainties around reliability and interpretability. Consequently, assessing the quality of a RAG system's document base must extend beyond conventional checks for accuracy or data cleanliness. It requires evaluating the semantic fidelity, topical coverage, and contextual completeness of the underlying corpus, all of which directly shape retrieval precision, generation quality, and factual grounding [21, 22, 23]. Within MRM frameworks, data quality is not only a technical consideration but a governance responsibility. Hence, effective MRM requires demonstrable traceability between input quality and output behaviour. For this reason, data quality controls should be explicitly defined and evidenced as part of the model inventory, validation, and ongoing monitoring documentation. Some key quality control dimensions include:

Topical coverage and authority: ensuring that document sets adequately represent the policy or business domains for which the model is used, with reliable authorship and verifiable provenance;

Freshness and versioning: maintaining metadata that records document creation dates and update history to prevent staleness or superseded references;

Extraction and structural fidelity: validating non-trivial preprocessing (e.g. OCR, table or list extraction) to preserve the integrity of information presented to the model;

Chunking and representation: aligning segmentation methods with document structure to avoid semantic distortion or information loss;

Sensitivity detection: identifying and managing personal, confidential, or regulated information embedded in the corpus through SME review or automated screening tools.

Given the potential dynamism of document bases, and the limited ability to surface GenAI model risks entirely during development, some of these quality issues may not arise until the model has been deployed. This would require embedding data governance controls throughout the model lifecycle. Such continuous oversight requires linking input-level data checks with RAG-specific performance metrics (e.g. groundedness, completeness, relevance, and utilisation) to form a closed feedback loop between data quality and model behaviour.

2. The RAG risk landscape

2.1.2 Legal and compliance burdens

RAG systems ingest, index, retrieve, and surface text drawn from heterogeneous sources. As these data sources vary in provenance, ownership, and sensitivity, they expose institutions to a broad spectrum of legal and compliance data risk [24, 25, 26], including: *personal and confidential data handling, contractual permissions for client data, licensing/IP and acceptable use, records management and retention, and data residency/sovereignty.* These risks may appear at any stage of the RAG pipeline, which spans *ingestion, retrieval, and generation*, as well as through *logging/ monitoring* of model operation.

FIs already operate within mature compliance frameworks designed to manage data-related risks [2]. These include establishing lawful bases for processing personal or commercial data, maintaining data retention and deletion schedules across systems (e.g. caches or logs), responding to data breaches, and compliance with individuals' rights around automated decision-making under frameworks such as GDPR. Such controls are typically embedded within wider model governance and risk management processes [3, 27]. However, GenAI systems, and RAG systems in particular, introduce new risk vectors that challenge existing controls and assumptions, such as prompt injections and rephrasing [28, 29]. While many governance principles remain applicable, the distributed, vendor-dependent, and continuously adaptive nature of these systems warrants special attention in several areas, including:

Vendor dataflows: Most GenAI systems rely on third-party vendors for model hosting and inference. Institutions must ensure that all data transfers between internal systems and vendor platforms are securely encrypted, contractually governed, and compliant with cross-border dataflow and sovereignty requirements [30, 31].

Information leakage: Retrieved content in RAG systems may be reproduced or rephrased by LLMs, creating a persistent risk of disclosing sensitive data. Even with utilisation or propagation tests, this risk cannot be fully eliminated [14, 28].

Continuous monitoring: GenAI systems require real-time monitoring to track performance and compliance, which can produce extensive logs that may contain sensitive data, requiring stricter retention and audit controls in line with guidance from NIST and the Bank of England [4, 32].

These issues illustrate how GenAI and RAG systems extend the scope of legal and compliance concerns beyond traditional model governance. FIs' existing MRM frameworks, which already impose responsibilities for data lineage, validation, and oversight in line with SR 11-7 and SS1/23, must now also address the contractual, cross-jurisdictional, and technical dimensions of GenAI supply chains. To support this broader remit, FIs' governance practice will need to treat vendor transparency, utilisation monitoring, and log governance as integral evidence categories within model validation and ongoing monitoring.

2.1.3 Ground truth availability or validity

The availability of ground truth data presents a fundamental challenge for GenAI-based models, whose integration into business processes differs from traditional models. Unlike models which are trained and validated against well-defined quantitative outcomes, GenAI models often produce textual or qualitative outputs such as summaries or narratives, where there may not be one definite or verifiable answer [33, 34]. As a result, the concept of ground-truth becomes fluid, as outputs vary according to subjective dimensions such as tone, framing, or perceived utility, which are challenging to quantify or benchmark objectively [34]. This absence of fixed benchmarks limits the effectiveness of conventional performance metrics and requires alternative validation methods. To address this challenge, FIs are increasingly adopting structured qualitative assessments such as expert review, scenario-based testing, and longitudinal monitoring to evaluate factual grounding, consistency, and usability of GenAI outputs over time.

Moreover, recognising these challenges, regulators such as the Bank of England and the Bank for International Settlements have emphasised the importance of ongoing monitoring to validation, given the limitations of static, point-in-time validation for systems whose behaviour evolves through data updates, vendor changes, or contextual variation [4, 6]. FIs are therefore embedding substantial ongoing post-deployment validation practices into the GenAI model lifecycle to ensure reliable and appropriate oversight where definitive ground truth cannot be established.

2.2 Vendor risks

Vendor risk has always been recognised as a contributor to model risk by MRM frameworks, but for GenAI systems, it becomes an even more prominent source of uncertainty and governance challenges. While traditional models are typically built and operated within a firm's perimeter, GenAI models invariably leverage artefacts sourced from a broad ecosystem of providers for foundation models, embeddings, hosting infrastructures, and orchestration services, and most FIs access GenAI capabilities via external APIs [35, 36, 37]. This dependency introduces risk exposure across a number of touchpoints within both the development cycle and deployment contexts. Accordingly, third-party risk—often managed as a distinct but supportive organisational function in FIs—now arises with greater magnitude across data protection, change management, and operational resilience. This is emerging as an area requiring particular attention and innovation, with supervisory bodies having highlighted vendor governance as an emerging area of scrutiny for AI systems in finance [38, 39].

2. The RAG risk landscape

2.2.1 Artefact re-versioning

A consistent challenge for FIs is the speed and opacity of vendor-driven model updates. For example, foundation models are updated frequently, with minor releases occurring every few months (e.g. GPT-4o versions '*2024-05-13*,' '*2024-08-06*,' and '*2024-11-20*'), while major revisions and upgrades follow a roughly annual cycle (e.g. the transition from GPT-4 to GPT-5) [10, 12]. Although vendors offer "snapshot" version pinning to provide temporary stability, these are updated infrequently and may be deprecated with relatively little notice [12, 13]. With these updates, there may be retraining, alignment, or parameter adjustments which can alter model behaviour without disclosure. Since vendors rarely publish detailed change logs, clients often have limited visibility into what has changed or why [40].

This pace of change conflicts with the slower validation cycles typical in FIs, where model reviews may occur only a few times a year. As a result, systems dependent on external models risk behavioural drift between validation checkpoints. Hence, existing MRM processes, which are typically applied to static, internally owned models, are being tested by the unprecedented velocity of change introduced by GenAI systems [41]. Hence, to mitigate this, institutions are embedding vendor-version management within their MRM lifecycle with key controls. They include maintaining explicit version inventories, obtaining vendor attestations, and conducting structured change impact assessments when models are updated. Additionally, recent regulatory frameworks such as SR 11-7, SS1/23, and guidance such as the NIST AI RMF [42] highlight vendor transparency, version traceability, and update governance as essential to the AI lifecycle.

2.2.2 Availability and cost

The operational performance and cost structure of GenAI systems are largely determined by their external vendors. Unlike traditional in-house models, most GenAI capability is accessed through third-party APIs, placing both availability and unit economics beyond the direct control of FIs [43]. This creates two types of risk:

Availability and latency: Factors like vendor outages, regional disruptions, or usage throttling can delay or interrupt model operations. Even minor backend changes or routing adjustments can affect response times and breach service-level objectives (SLOs) [43, 30].

Cost fluctuations: Operating costs can vary due to changes in token pricing, larger data retrievals in RAG systems, longer user sessions, or the introduction of additional safety layers [32].

The significance of these risks depends on the use case. High-stakes or real-time applications demand consistent uptime and predictable costs, whereas support functions can tolerate greater variation [30, 43]. To manage these dependencies, institutions should integrate vendor performance and cost oversight directly into their model risk frameworks.

2.2.3 Open-weight or on-premises GenAI artefacts

Open-weight models are foundation models whose trained parameters (weights) are made publicly available for download and use, though not necessarily alongside the training data or code. The use of open-weight models and on-premises hosting of GenAI artefacts can reduce certain direct vendor risks, such as data disclosure, provider lock-in, and unpredictable availability or version deprecation. They also increase the FI's ability to control factors such as latency, data residency, and configurability. However, they do not eliminate vendor risk, but redistribute it across a different supply chain and move certain risk management burdens in-house [44].

Hosting of foundation models poses a substantial operational burden, with significant complexities posed by capacity planning, security, observability, and cost management: creating institutional capacity for this requires substantial investment in technical expertise and uplifting governance to handle greater model load, and may not outperform high-scale vendors on cost [45]. These operational and cost risks cannot be mitigated through service-level agreements (SLAs) in the same way as they can for vendor-hosted artefacts.

Open-weight artefacts are not a perfect solution to supply chain risks: these artefacts still inherit risks from their providers' technical stacks; and although they will often inherit provider-developed hardening fixes or patches, the scheduling of these will remain a third-party dependency [46], and these changes will still require change management and re-validation. FIs' legal and third-party risk functions will still require oversight over these upstream vendors, even if they do not host the artefacts. Furthermore, 'open' does not imply permissive or licence-free [47]; these artefacts may still entail legal, contractual, or other compliance limitations on use, including restrictions on specific behaviours or use-cases [48]. Furthermore, indemnity or liability protections offered under open licensing are likely to be less comprehensive than vendor terms [49].

Finally, open-weight artefacts may suffer from functional limitations: the most powerful foundation models have to date remained proprietary to vendors [50]; out-of-the-box utility will be limited without the functional tooling, tuning, and high-availability patching provided by vendors; and open projects often suffer from fragmentation issues, complicating regression testing and the artefact updates which will now be required of in-house teams.

Open-weight and on-premises GenAI artefacts are a powerful tool in an FI's GenAI toolbox if the economics of scale allow, but it is doubtful that they will or should entirely replace vendor-supplied artefacts. Instead, those institutions should leverage these assets as risk mitigation following a diversification strategy, in line with sectoral guidance such as the UK PRA's SS2/21 [51].

2. The RAG risk landscape

2.3 Architecture risks

GenAI models are typically constructed as composed architectures, such as coupling LLMs with external retrieval stores in RAG systems, and in more advanced implementations incorporating tool interfaces or evaluation components such as 'LLM-as-a-Judge' modules for automated monitoring [14, 15]. In such modular pipelines, small specification gaps between components, or behavioural shifts in a single module, can propagate and create wider system-level effects that are difficult to predict, diagnose, or reproduce [52, 53]. This can escalate into several interrelated risk categories, such as:

Interface coupling: Weakly defined or unversioned interfaces between modules (e.g., retrieval or prompt schemas) can cause silent performance degradation and hinder validation [52].

Behavioural drift: Updates to embeddings, retrievers, or orchestration logic may shift outputs unpredictably, producing non-linear system behaviour [54].

Evaluation fragility: LLM-based evaluators exhibit similar biases and instability to the LLM-based models they monitor, especially in cases of common foundation models, creating circular dependencies and unreliable metrics [55].

Configuration complexity: Misaligned versions or undocumented parameter changes across subsystems weaken reproducibility and auditability [52].

These risks highlight the need to approach GenAI model risk as an end-to-end system governance problem, in which the integration of components is itself a key risk surface. It is not sufficient to validate the foundation model, or any other component, in isolation. This perspective must be reflected across FIs' MRM practice: developers should document explicit interface contracts and perform change-impact testing between model components; version control and change management must be baked into orchestration; dependencies should be inventoried; and reproducibility of model behaviour should be a target of validation and ongoing monitoring. In some cases, the scope of a MRM framework's model identification criteria may require re-evaluation to address model components that perform critical or sufficiently complex behaviour, or that are developed sufficiently independently of the other components (we discuss this challenge further in Section 4.2.1).

2.4 Human factors

Human factors are a substantial driver of risk in GenAI systems. Even when technical architectures and controls are sound, the way humans interact with, interpret, and rely on model outputs can introduce systemic vulnerabilities. Four recurrent patterns, *automation bias, algorithmic aversion, cognitive offloading,* and *mode confusion,* capture the spectrum of trust and role-alignment failures observed in GenAI-assisted environments as presented below:

Automation bias: Users may over-trust GenAI outputs and skip verification, creating oversight gaps; calibrated interface cues can help sustain critical scrutiny [56, 57, 58, 59, 60].

Algorithmic aversion: Conversely, users may resist model recommendations despite evidence of superior performance; clearer role definition and interaction design that increases users' confidence in their control over the system can reduce this effect [61, 62].

Cognitive offloading: Reliance on GenAI tools may erode domain expertise over time, heightening dependence and reducing human challenge [63, 64, 65].

Mode confusion: Misunderstanding an AI system's role or authority can lead to incorrect actions, a risk well-documented in safety-critical domains [60, 66].

These patterns demonstrate that risks due to GenAI models extend beyond conceptual soundness and output validity to include human trust calibration and role alignment. For FIs, effective risk management therefore requires system-centric oversight, linking technical validation with human-factor controls such as training, interface design, and operational guardrails. This integrated approach can provide data governance, vendor transparency, architectural traceability, and human reliability that can reinforce within a unified risk framework.

In summary, the risk landscape of RAG-based GenAI systems spans multiple interconnected dimensions: *data, vendor, architectural,* and *human factors*; which extend far beyond questions of model accuracy alone. It is essential to highlight that effective oversight requires continuous traceability between inputs, outputs, and governance actions, close coordination between technical developers and the full spectrum of FI risk functions, and explicit recognition of evolving vendor dependencies. These dimensions can be summarised in a structured taxonomy that aligns each risk category with its primary risk management priorities under SR 11-7 and SS1/23, as shown in Table 1. Building on this taxonomy, the next section examines how these principles are applied in practice through two financial institution case studies. Together, they demonstrate how RAG and multi-agent architectures can be developed, validated, and monitored within SR 11-7- and SS1/23-aligned MRM frameworks.

Table 1: Summary of key RAG-related risk dimensions and corresponding MRM control priorities.

Risk category	Representative risks	MRM control priorities
Data	Unstructured input quality, data provenance, compliance gaps, lack of ground truth	Corpus validation, lineage tracking, utilisation testing, and continuous data–output traceability
Vendor	Frequent re-versioning, API reliability, cost volatility, opaque change logs	Version inventories, vendor attestations, fallback planning, and resilience testing
Architecture	Interface drift, orchestration errors, configuration misalignment, evaluator fragility	End-to-end validation, interface contracts, version control, and reproducibility checkpoints
Human factors	Automation bias, overreliance, misunderstanding of model boundaries	User training, role definition, explainability controls, and operational guardrails

3 Case studies

This section presents two partner case studies from major FIs: the *Digital Credit Platform (DCP)* and the *Lead Recommendation Engine (LRE)*. Both illustrate practical applications of LLM-based systems within established governance and MRM frameworks, focusing on model identification, validation and governance for RAG and multi-agent architectures. These case studies were selected as they are embedded in mature banking workflows and reveal concrete adaptations to MRM practices. They are not presented as templates, but as concrete, bounded examples whose lessons may need to be adapted for different institutional sizes, regulatory environments, and levels of GenAI maturity.

3.1 Digital Credit Platform

This case study explores how a major financial institution deployed a RAG model within its credit-lending workflow to enhance efficiency while maintaining regulatory control. The DCP illustrates a pragmatic approach to adopting GenAI within existing processes and a strong internal MRM governance practice, establishing clear model boundaries and continuous monitoring.

The following subsections outline (i) the use case and model design, (ii) governance arrangements aligned with SR 11-7 and SS1/23, and (iii) validation and monitoring mechanisms adapted for RAG systems.

3.1.1 Use case description

The application employs a RAG-LLM architecture to automate sections of wholesale credit applications. Specifically, the model generates draft write-ups for *Company Background* and *Business/ Industry Risk* fields based on user-uploaded company files and industry reports. The objective is to streamline the drafting process while maintaining factual accuracy and professional tone.

The model operates within the bank's federated GenAI platform, which provides access to vendor-hosted LLMs through secure APIs and applies common data-classification and access controls. This RAG pipeline transforms uploaded documents into vector embeddings, retrieves relevant content, and generates two candidate drafts per section for user selection. The design choice to surface multiple candidate drafts was intended to reinforce the user's awareness of the potential variability in the generator LLM's outputs; the developers attempted to minimise hallucination and related issues by setting a very low 'temperature' value, but there remained residual uncertainty.

The system automates low-impact components of the credit application; all remaining sections are still completed manually by the team responsible for lending. This design keeps human review central to the process and provides GenAI outputs to augment rather than replace expert judgment. The model's development was jointly led by the business unit and the an internal AI Centre of Excellence, reflecting shared ownership across technical and operational lines.

Operating boundaries specify that all outputs must be factually accurate, professional in tone, and fully grounded in the uploaded materials. The DCP's value, therefore, lies not in decision automation but in efficiency gains, workflow consistency, and demonstrable accountability. This illustrates how GenAI can improve productivity while remaining aligned with supervisory expectations for human control.

3. Case studies

3.1.2 Governance

The DCP was developed under existing SR 11-7–aligned MRM structures, with additional GenAI controls. Ownership and development were shared between business and engineering teams. In addition to oversight and independent validation performed by the Model Risk function, use-case owners sought approvals from risk teams, including Data, Cyber, IT, and Legal. Development of all GenAI models in the institution is also subject to central AI governance forums.

Independent validation followed a two-line defence structure. The first line (the development team) conducted internal testing and performance benchmarking, while the second line (independent validators from the Model Risk team) performed a formal validation using additional tests, based both on the principles applied to traditional statistical models and extensions to those principles to more directly address GenAI-specific risks. A pilot review exercise conducted collaboratively between the first and second lines established performance baselines and informed the validation test plan prior
to deployment.

Controls were implemented through the federated platform and supplementary safeguard processes, including standardised data classification, prompt version control, vendor model tracking, and defences against hallucination and prompt injection. The model's materiality was initially rated as high due to uncertainty given the novelty of GenAI models to the teams involved, but was later revised downward to directly reflect limited business exposure and effective mitigations. This adjustment demonstrates how proportionality and existing governance structures remain relevant.

3.1.3 Model validation and monitoring

Independent validation followed standard MRM principles, augmented for GenAI-specific risks such as hallucination, toxicity, prompt injection, and stability. Testing employed metrics including *ROUGE-L*, RAG-specific measures, and SME scoring. The utility of lexical methods such as *ROUGE-L* applied to LLM outputs was found to be limited, so these were augmented by semantic methods for judging RAG metrics such as completeness. Each LLM prompt within the pipeline was inspected, and validation outcomes focused on end-to-end output quality.

Performance monitoring was implemented through an LLM-as-a-Judge framework that evaluates groundedness, completeness, relevance, and other metrics on a daily basis. Daily dashboards track these metrics, providing early detection of performance drift and version-related degradation. This real-time monitoring capability represented a material advancement over traditional periodic monitoring cycles.

The monitoring framework also serves as an early warning system for vendor model changes, and enables evaluators to compare performance across LLM versions and optimise cost-efficiency while maintaining control. Overall, the DCP demonstrates how dynamic monitoring, coupled with traceable validation, translates the intent of SR 11-7 and SS1/23 into practice for continuously evolving GenAI environments.

3.2 Lead Recommendation Engine

This case study examines how a large financial institution implemented a multi-agent GenAI system to support client relationship management and cross-selling activities. The LRE demonstrates how complex, modular architectures can be governed while remaining aligned to established MRM frameworks. The following subsections outline (i) the use case and system architecture, (ii) governance structures integrating business, compliance, and technical teams, and (iii) validation and monitoring practices tailored to multi-agent configurations.

3.2.1 Use case description

The LRE was developed to assist Client Relationship Managers (CRMs) in identifying cross-selling opportunities across the bank's existing client base. It analyses CRM system data, regulatory filings, and market intelligence to recommend financial products relevant to specific clients.
The system is built on a multi-agent architecture operating within the bank's internal GenAI platform. This platform provides access to vendor LLMs (GPT-4.1, 4o, and Gemini 1.5), orchestration tools, and standardised data-governance and monitoring frameworks.

The LRE comprises distinct RAG agents for data retrieval and analysis, Persona agents that model client, product, and peer attributes, and Critic and Consultant agents that synthesise insights to generate the final recommendation. The orchestration layer manages agent sequencing and ensures output consistency.

The development of the system followed a collaborative model involving AI engineers and business subject-matter experts (SMEs), who provided iterative feedback on generated leads and narrative quality. This partnership enabled alignment between model performance and business expectations, while ensuring that human oversight remained central to decision-making. The system was piloted on a controlled client subset that had consented to participate in AI-assisted service trials.

3. Case studies

3.2.2 Governance

Risk management followed an existing architecture-agnostic taxonomy: Model, Data, Architecture, and Vendor risks. The only failure mode with business impact was judged to be a bad recommendation provided to the CRM, and so was the only *business* risk considered material. Architecture and Vendor risks were also judged material, including: handoff/sequence failures, compounded hallucinations, limited explainability of GenAI systems as well as multi-agent systems, vendor API uptime, jurisdictional issues around third-party dependencies, and unannounced version deprecations. Risks of sensitive data disclosure were also judged to be material given the use of vendor LLMs hosted on third-party infrastructure.

Development was jointly owned by the business team and an engineering team, and the Model Risk function (MRM) engaged continuously; the business team set thresholds, and an interdisciplinary AI Release Board (business, compliance, engineering, information security) ratified thresholds and acted as deployment gatekeeper. The engineering team was not a decision-maker in these tasks, and the AI Release Board assumed ultimate ownership over development gating.

This approach to governance demonstrates alignment to existing MRM frameworks while also horizontally broadening the input of risk expertise from across the institution, and shifting the balance of responsibilities over development further in the direction of interdisciplinary fora.

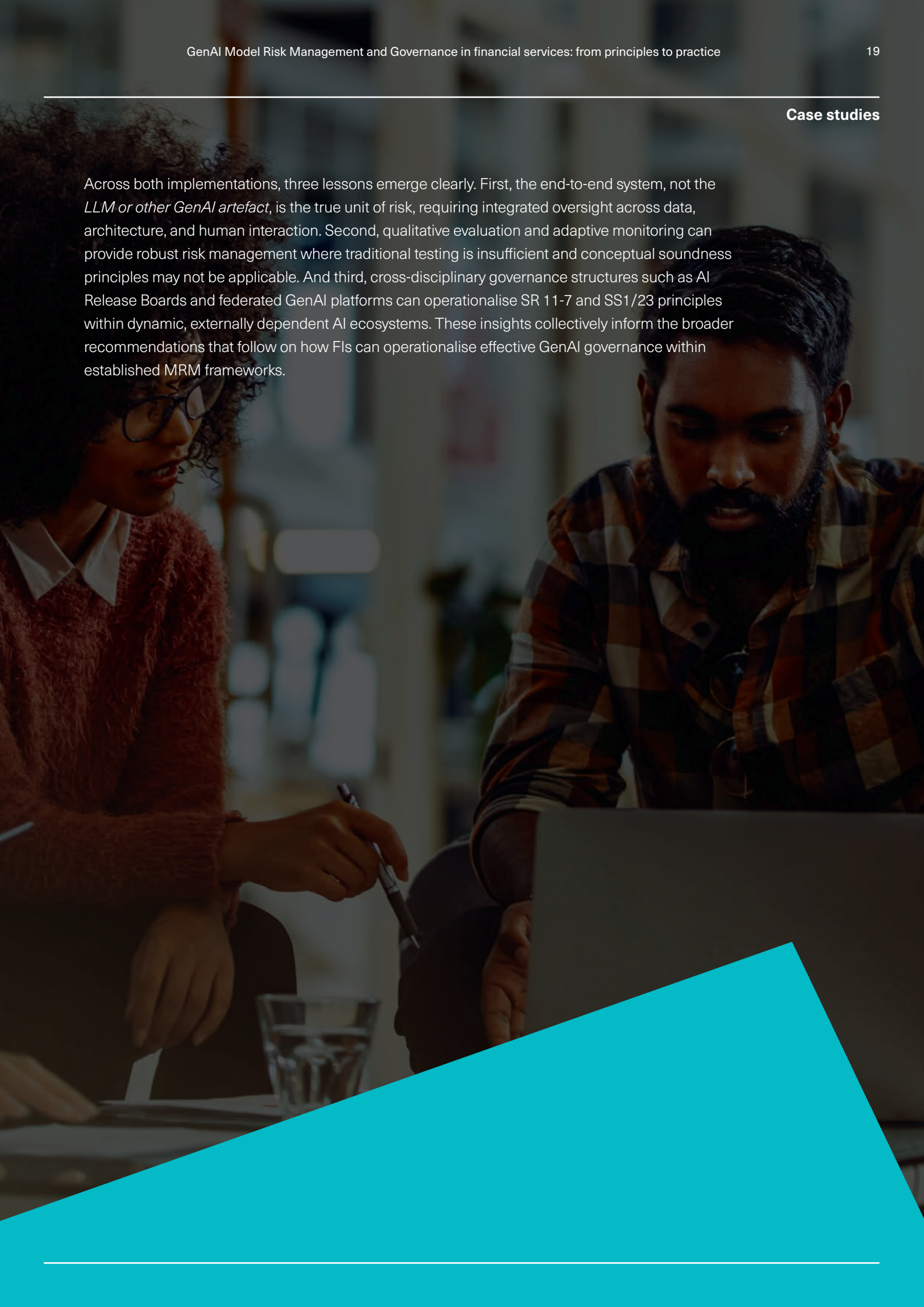
3.2.3 Model validation and monitoring

Model performance was evaluated primarily through SME review using a 1–5 scoring and qualitative feedback. Engineering teams applied quantitative metrics such as accuracy, hallucination rate, and tone to meet business-defined thresholds. Each agent was validated independently as a discrete task, and overall system validation focused on output utility and consistency. These assessments established the operational baseline for deployment: the model was approved only when SMEs deemed its recommendations meaningfully helpful to their workflow.

Traditional challenger-model testing was infeasible; instead, pre-development justification was required to demonstrate that LLMs were the appropriate solution over statistical or ML alternatives. SME evaluation served as the operational baseline; deployment was approved only if SMEs deemed the tool practically valuable. Post-deployment monitoring tracks performance at the overall model level via qualitative CRM feedback. Identified architectural risks include data handoff errors, sequencing issues, and compounded hallucinations. Vendor risks such as API uptime, jurisdictional limitations, and version deprecation are continuously monitored.

In summary, the DCP and LRE case studies illustrate how financial institutions are adapting long-standing MRM principles to the realities of GenAI deployment. Both use cases demonstrate that traditional validation, documentation, and governance processes remain applicable but require recalibration for systems that evolve through external vendor updates, multi-agent orchestration, and continuous data interaction. The DCP highlights the feasibility of embedding GenAI within mature business processes under tight human oversight, showing that near-real-time monitoring can enhance transparency and responsiveness. By contrast, the LRE exposes the expanded risk surface created by modular architectures and vendor reliance, highlighting the need for cross-functional governance and SME-driven validation where quantitative benchmarks are unavailable.

Across both implementations, three lessons emerge clearly. First, the end-to-end system, not the *LLM or other GenAI artefact*, is the true unit of risk, requiring integrated oversight across data, architecture, and human interaction. Second, qualitative evaluation and adaptive monitoring can provide robust risk management where traditional testing is insufficient and conceptual soundness principles may not be applicable. And third, cross-disciplinary governance structures such as AI Release Boards and federated GenAI platforms can operationalise SR 11-7 and SS1/23 principles within dynamic, externally dependent AI ecosystems. These insights collectively inform the broader recommendations that follow on how FIs can operationalise effective GenAI governance within established MRM frameworks.



4 Operationalising GenAI governance in FIs

The use of GenAI models creates a wide range of risks, as discussed in Section 2. It also puts pressure on existing governance processes and organisational structures in FIs. Partners report that safe, production-level use of GenAI models can stretch current MRM processes. In some cases, this has led institutions to adapt their existing MRM practice. In others, institutions have added specific governance activities for GenAI models to complement traditional MRM. This section draws on partner reflections to describe the main challenges and to propose recommendations for institutions that want to build an AI model governance practice based on the principles and practices of MRM.

We structure the governance challenge in two parts. First, we discuss high-level challenges that affect the whole organisation. These are cross-cutting issues that shape how GenAI models are adopted and governed across functions and business lines. Second, we discuss more specific pressures on institutions that already use MRM-style model governance and need to extend it to GenAI systems. We then suggest concrete enhancements that institutions can adopt in response. These recommendations assume a basic governance approach in line with the principles of SR 11-7 and SS1/23-style MRM. However, we also show that GenAI governance requires deeper involvement from a wider set of risk functions, including Third Party Risk Management (TPRM) and Information Security. In addition, institutions will need new cross-disciplinary centres of expertise that focus on AI-related risks.

4.1 AI Governance for FIs: organisation-level challenges

We now describe the generic organisation-level challenges that FIs face when adapting existing model governance frameworks, such as those based on SR 11-7 and SS1/23, to cover GenAI models. Partners noted that, in their institutions, governance of GenAI models often needs a wider and more interdisciplinary effort than their current practice for traditional models. In several institutions, discussions are underway on how to adapt existing structures that bring together risk and business functions, so that risks can be considered in a more holistic way than within a single Model Risk team. Our goal with this discussion is to highlight the top-level challenges and insights from our partners' experience adapting their governance practices for GenAI models.

4.1.1 Adoption at scale: volume and velocity

Partners raised concerns about whether existing MRM governance practices can scale to the volume and speed of GenAI model adoption. Traditional model governance frameworks evolved to manage a limited number of stable, high-materiality models that were developed and validated over many months. GenAI adoption breaks these assumptions. Developers can now build many lightweight GenAI tools for diverse use cases in a short time, partly by reusing common components and external vendor APIs. As a result, partners question whether the current pace and staffing model for traditional MRM validation and governance can be sustained.

Current responses focus on making validation more scalable without weakening standards. Several partners use standard onboarding for new foundation models, applying a common evaluation suite before they are made available to developers. This provides a baseline view of risks and a shared set of generic controls. Some are also building component-based frameworks in which common building blocks are validated once, so that subsequent validation focuses on how these components are combined. In parallel, partners are starting to group low-materiality GenAI tools into shared reviews and to reuse evidence from earlier validations where this is appropriate.

Partners also described tension between first-line developers, who are keen to deploy GenAI models quickly, and second-line validators, who take a more cautious view. This tension can increase when senior management encourages rapid adoption. One partner is addressing this at a cultural level, using interdisciplinary forums to promote a shared expectation that any technical solution, including GenAI, must be justified as the most appropriate way to solve a given problem.

4. Operationalising GenAI governance in FIs

4.1.2 Internal AI-specific expert groups and their responsibilities

Partners described creating new organisational structures to coordinate expertise across technical, risk, and business functions. The most common example is the establishment of an internal Centre of Excellence (CoE) for AI. This acts as a central forum for oversight and guidance. These bodies help institutions handle the many open questions around GenAI model development. For example, which use cases are acceptable, what data can be used, how to measure and monitor risks like hallucination or bias, while staying aligned with formal MRM-style governance.

The remit and authority of AI CoEs vary. In some institutions, the CoE has an advisory role, supporting first-line developers and second-line validators with input from other risk teams such as Data Protection, Compliance, and Information Security. In others, the CoE has formal decision-making powers, for example, approving development standards, reviewing documentation, or gating model approvals. Across these models, partners find that engaging a broad set of teams early, especially those responsible for data protection, vendor risk, and enterprise risk, helps to surface issues sooner and speeds up coordination when GenAI systems change.

Partners consistently described these arrangements as evolutionary rather than disruptive. The core MRM principles, such as clear ownership, independence, documentation, and effective challenge, remain in place, but they now apply to a wider set of teams. For some partners, a central point of AI expertise offers a practical way to embed accountability across the GenAI model lifecycle and to promote consistent practices across the institution.

4.2 Implications for SR 11-7 and SS 1/23- style governance practice

The organisation-level challenges described above translate into a number of concrete adjustments to SR 11-7 and SS1/23-style MRM. Partners highlighted that, in practice, they are not replacing existing frameworks. Instead, they are applying the same principles to GenAI systems in a more explicit and system-wide way. The main areas of change relate to how GenAI systems are identified and scoped, how risk tiering is applied, how deployments are linked to business processes and controls, how functions work together, and how monitoring and change management are organised. We discuss these fundamental questions in the following sections.

4.2.1 Model identification and scope

The first practical question is what counts as 'the model' in a GenAI setting. Under established MRM principles, the model is not defined as a single algorithm or foundation model. For GenAI use cases, the relevant unit of risk is the entire GenAI workflow, including data sources, retrieval or search components, prompt templates, orchestration logic, and one or more LLMs. Behaviour often depends on how these parts interact, not just on any one component in isolation. In response, partners now describe the end-to-end GenAI pipeline in the model inventory entry. In practice this means recording system boundaries, key data sources, important components, and any reliance on third-party services. Some partners are also starting to treat complex or widely reused components, such as shared retrievers or evaluator modules, as separate entries in the inventory. This allows institutions to oversee and update those components without having to re-open a full validation for every downstream system each time a small change is made.

Within this broader system view, prompts and templates are treated as part of the formal model specification rather than as an informal configuration. Hence, partners emphasised that any LLM prompts or prompt templates should be subject to risk management in the same way as other modelling decisions, including version control, review, and revalidation. In some cases, where the prompt set is static, such as the DCP use-case (Section 3.1), this may be straightforward. However, other prompt sets may require changes in governance practice if they are liable to be updated after deployment, determined by run-time input, or reused across models. In practice, only a minority of partners currently treat prompt templates and orchestration logic as first-class, regularly maintained artefacts in the inventory; this weakens effective challenge and traceability and is an area for improvement.

4.2.2 Risk tiering for GenAI systems

The second area of adjustment concerns risk tiering. Existing model tiering frameworks, such as three-tier materiality schemes that classify models as high, medium, or low risk based on materiality and complexity in line with SS1/23, remain the starting point. These traditional dimensions, including business impact, regulatory use, potential customer harm, and data sensitivity, still apply to GenAI models. However, partners have found that GenAI systems introduce additional elements, such as stronger dependence on external vendors, more frequent changes to models and corpora, greater behavioural opacity, and a wider range of possible uses. These factors affect how stable a system is over time and how easy it is to explain and control, and so they also need to be reflected in tiering decisions.

To address this, extending tiering rubrics with a small number of GenAI-specific factors, such as vendor dependency, change velocity, and the degree of autonomy given to the system, could be a way forward. Additionally, introducing explicit triggers for review or re-tiering can be beneficial for auditability, for example changes to foundation model versions, substantial changes to the retrieval corpus, major updates to guardrail logic, or observed drift in key risk indicators. Recording tiering decisions and subsequent changes with a short explanation of why a given tier was chosen or amended improves transparency and supports proportional oversight. This is consistent with SR 11-7 and SS1/23 and helps keep oversight proportionate but responsive to the more dynamic behaviour of GenAI systems.

4. Operationalising GenAI governance in FIs

4.2.3 Organisation, processes, and skills

The third area includes broader aspects of organisations, their processes and skills. GenAI systems are now used in many different parts of FIs. Some are embedded in core decision workflows, while others act as assistive tools for tasks such as drafting, summarising, or searching information. The same GenAI capability may also be reused across several processes and business units. This can make it harder to see where a model is used, who is accountable for it, and which controls apply. To address this, strengthening the link between model inventories and process documentation can be a way forward. For each material GenAI system, institutions aim to record where it is deployed, what decisions it supports, who owns the process, and how it fits within the wider control environment, so that there is a clear audit trail to support accountability. Standard operating procedures and user guidance can then be aligned with this documented role, so staff can understand when the tool is advisory, when human review is required, and which existing controls must be applied.

This wider use, as well as the wider risk profile of GenAI models, also broadens the set of functions involved in governance. Some partners described a shift from a model risk–centric picture towards a more distributed one that includes Data Governance, Data Protection, Third-Party Risk, Information Security, Legal, Compliance, and business teams. Institutions are updating role descriptions and escalation routes so that ownership, challenge, and support are clear, and are investing in basic training and documentation standards so that teams with different backgrounds can apply existing MRM principles consistently to GenAI systems. MRM therefore remains one part of a broader AI governance landscape, focused on technical model risk, while other functions address questions of ethics, systemic risk, and wider regulatory expectations.

4.2.4 Monitoring and change management over the lifecycle

The final areas of concern include monitoring and change management. Partners find that GenAI systems change more often than many traditional models. For instance, vendors update foundation models, retrieval corpora are refreshed, and prompts, templates, and guardrails are revised. Because behaviour depends on how these elements interact, the role of monitoring in validation is especially important, and institutions are moving to a lifecycle view of assurance. In practice, firms define what counts as a material change for each system and specify the level of review required. Material changes are assessed before implementation, rolled out in a controlled way where needed, and checked afterwards against agreed quality and risk measures. Key elements such as versions, configurations, and key test results are logged so that behaviour at a given time can be reconstructed. Vendor and data changes are routed through change management processes so that they trigger appropriate re-testing and, where necessary, re-validation. Some partners use automated evaluations, including language-model-based checks, alongside periodic human review, while others rely more heavily on business metrics and incident reports. The aim is to apply SR 11-7 and SS1/23 expectations on ongoing monitoring to a more fluid GenAI environment, while keeping effort proportionate to the model's tier and use.

5 Conclusions

This report has examined how financial institutions can adapt established Model Risk Management (MRM) frameworks, such SR 11-7 and SS1/23, to govern GenAI systems in practice. Drawing on partner discussions, a structured taxonomy of RAG-related risks, and two case studies, it concludes that GenAI does not require a separate governance regime, but targeted refinements to existing MRM practice to reflect the dynamism, modularity, and vendor dependence of GenAI workflows.

Across partners, three themes are consistent. First, the relevant unit of risk is the end-to-end GenAI workflow, not any single artefact such as an LLM. Second, static, point-in-time validation must be complemented by continuous monitoring and integrated vendor oversight to keep pace with change. Third, effective control depends on cross-functional governance and human-factor measures, not technical controls alone.

Conclusions

We identify the following practical priorities for FIs seeking to operationalise GenAI governance within established MRM frameworks:

- Extend model inventories to capture end-to-end GenAI workflows. Entries should describe system boundaries, key components (including retrievers, prompts, evaluators, and orchestration logic), vendor dependencies, and the business processes and decisions affected.
- Refine tiering frameworks with GenAI-specific dimensions. In addition to traditional materiality and complexity, tiering should reflect degree of automation, exposure to untrusted inputs, change velocity, and concentration of vendor and data dependencies.
- Define and implement lifecycle monitoring and change triggers. Institutions should identify relevant performance and risk indicators and link them to clear thresholds for investigation, rollback, or re-validation when vendor, data, or configuration changes occur.
- Integrate vendor oversight into model documentation and validation. Model files should record which external artefacts and services are in scope, how they are monitored, and how vendor changes are mapped to internal change-management processes, including responsibilities for assessing impact.
- Formalise cross-functional governance and human-factor controls. Clear mandates for AI governance functions, explicit role descriptions, training, and user guidance can help ensure that GenAI systems are deployed consistently with institutional risk appetite and that human oversight remains meaningful.

It should also be noted that traceability is important for regulatory confidence. Supervisors need to see that institutions can explain how GenAI systems behave in practice, and how that behaviour relates to design choices, data sources, vendor updates, and human interventions. The priorities above are valuable not only because they reduce operational and conduct risk, but because they make this traceability concrete and demonstrable.

Glossary

API
Application programming interface, a clearly documented point of connection at which one computer system interacts with another. The term can also be used to describe the specification of that point of connection via a specific set of requests one system may make of the other.

CoE
Centre of Excellence, an internal forum or team tasked with developing expertise around a specific topic or technology.

FI
Financial institution.

GenAI
Generative AI, a subset of machine learning methods which learn the distribution of (often very large) training data and can produce new data of that category given an input, often a natural language 'prompt.'

LLM
A large language model, a GenAI artefact with a deep learning architecture trained on large corpora of data, which models the operation of language and can be used as a general-purpose 'AI' component.

MRM
Model Risk Management, a practice required of banking organisations wherein risks arising from the use of quantitative methods or systems (models) in business decisions are managed through a framework addressing model development, implementation, and use, independent validation, and sound governance, practices, and controls.

PRA
The Prudential Regulation Authority, the UK's financial services regulator and a division of the Bank of England.

RAG
Retrieval-augmented generation, a design pattern in GenAI-based systems where an LLM prompt is augmented by an information retrieval pipeline which provides useful contextual information from a document base, in order to ground or constrain the generator LLM's output such that it relates directly to a specific supporting corpus.

SLA
Service-level agreements, agreements between a service provider and a customer, usually backed by contract, which define minimum levels of service quality through service-level indicators, such as mean time between failures and uptime.

SME
Subject matter experts, persons with direct and deep knowledge of a particular topic or business process and whose expertise is relied upon to ensure that organisations make informed decisions about their area.

SR 11-7
The supervision and regulation letter providing guidance on MRM issued by the Board of Governors of the US Federal Reserve and Office of the Comptroller of the Currency, first published in 2011 and which remains one of the foundational regulatory references for MRM.

SS1/23
The Bank of England Prudential Regulation Authority's Supervisory Statement on MRM principles for banks, published in 2023 and one of the key regulatory guidance documents in contemporary MRM.

Bibliography

[1] UK Government, "National cyber strategy 2022." Available: *https://www.gov.uk/government/publications/national-cyber-strategy-2022*, 2022.

[2] Basel Committee on Banking Supervision, Ed., *Principles for effective risk data aggregation and risk reporting,* Jan. 2013. Basel: Bank for International Settlements, 2013. Available: *https://www.bis.org/publ/bcbs239.pdf*

[3] Board of Governors of the Federal Reserve System, "Supervisory guidance on model risk management (SR 11-7)," Board of Governors of the Federal Reserve System, Washington, D.C., SR 11-7, 2011. Accessed: Oct. 09, 2025. [Online]. Available: *https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm*

[4] Bank of England Prudential Regulation Authority, "Supervisory statement SS 1/23: Model risk management principles for banks," Bank of England Prudential Regulation Authority, London, United Kingdom, SS 1/23, May 2023. Accessed: Oct. 09, 2025. [Online]. Available: *https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/supervisory-statement/2023/ss123.pdf*

[5] McKinsey & Company, "The economic potential of generative AI: The next productivity frontier," McKinsey Global Institute, 2023. Available: *https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier*

[6] Crisanto, J, and Leuterio, C,and Prenio, J, and Yong, J, "Regulating AI in the financial sector: Managing opportunities and risks," Bank for International Settlements (BIS) Financial Stability Institute, FSI Insights on Policy Implementation No. 63, 2024. Available: *https://www.bis.org/fsi/publ/insights63.htm*

[7] Financial Stability Board, "The financial stability implications of artificial intelligence," Financial Stability Board, Nov. 2024. Available: *https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/*

[8] European Central Bank, "The rise of artificial intelligence: Benefits and risks for financial stability," Financial Stability Review – Special Feature, May 2024, Available: *https://www.ecb.europa.eu/press/financial-stability-publications/fsr/special/html/ecb.fsrart202405_02~58c3ce5246.en.html*

[9] Cervicorn Consulting, "Retrieval augmented generation market size, share, growth, report 2025 to 2034." Available: *https://arxiv.org/abs/2312.10997*

[10] OpenAI, "Deprecations — OpenAI API." Available: *https://platform.openai.com/docs/deprecations*, 2023.

[11] Anthropic, "Model deprecations — anthropic (claude) docs." Available: *https://docs.claude.com/en/docs/about-claude/model-deprecations*, 2024.

[12] Microsoft, "Model retirements — azure AI foundry (azure OpenAI)." Available: *https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/model-retirements*, 2025.

[13] Amazon Web Services, "Model lifecycle — amazon bedrock." Available: *https://docs.aws.amazon.com/bedrock/latest/userguide/model-lifecycle.html*, 2025.

[14] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in neural information processing systems (NeurIPS 2020), 2020. Available: *https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html*

[15] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," arXiv preprint arXiv:2210.03629, 2022, Available: *https://arxiv.org/abs/2210.03629*

[16] T. Schick et al., "Toolformer: Language models can teach themselves to use tools," arXiv preprint arXiv:2302.04761, 2023, Available: *https://arxiv.org/abs/2302.04761*

[17] M. Wicker, L. Szpruch, and M. Søren, "Move fast without breaking the bank: Model risk management of GenAI workflows," London, United Kingdon, 2025. Available: *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5682603*

[18] McKinsey & Company, "How financial institutions can improve their governance of gen AI," McKinsey & Company Insights, 2024, Available: *https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/how-financial-institutions-can-improve-their-governance-of-gen-ai*

[19] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven Failure Points When Engineering a Retrieval Augmented Generation System," in Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI, in CAIN '24. New York, NY, USA: Association for Computing Machinery, Jun. 2024, pp. 194–199. doi: *10.1145/3644815.3644945*.

[20] S. Wu et al., "Retrieval-Augmented Generation for Natural Language Processing: A Survey." Accessed: Nov. 17, 2025. [Online]. Available: *http://arxiv.org/abs/2407.13193*

[21] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. Aroyo, "'Everyone wants to do the model work, not the data work': Data cascades in high-stakes AI," in Proceedings of the 2021 CHI conference on human factors in computing systems (CHI), 2021. Available: *https://dl.acm.org/doi/10.1145/3411764.3445518*

[22] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of retrieval-augmented generation: A survey," in Big data, Springer Nature Singapore, 2025, pp. 102–120. doi: *10.1007/978-981-96-1024-2_8*.

[23] T. Ding, A. Banerjee, L. Mombaerts, Y. Li, T. Borogovac, and J. P. D. la Cruz Weinstein, "VERA: Validation and evaluation of retrieval-augmented systems," 2024, Available: *https://arxiv.org/abs/2409.03759*

[24] European Parliament and Council, Regulation (EU) 2016/679 (general data protection regulation). 2016. Available: *https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng*

[25] National Institute of Standards and Technology, "Artificial intelligence risk management framework: Generative AI profile," U.S. Department of Commerce, NIST AI 600-1, 2024. doi: *10.6028/NIST.AI.600-1*.

[26] European Parliament and Council, *Regulation (EU) 2024/1689 (artificial intelligence act).* 2024. Available: *https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng*

[27] Information Commissioner's Office (ICO) and The Alan Turing Institute, "Explaining decisions made with AI." Accessed: Oct. 09, 2025. [Online]. Available: *https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/*

[28] OWASP Foundation, "OWASP top 10 for large language model applications (2025)." Available: *https://genai.owasp.org/llm-top-10/*, 2025.

[29] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." Accessed: Oct. 17, 2025. [Online]. Available: *http://arxiv.org/abs/2302.12173*

[30] Federal Deposit Insurance Corporation, Board of Governors of the Federal Reserve System, and Office of the Comptroller of the Currency, "Interagency guidance on third-party relationships: Risk management," Federal Deposit Insurance Corporation; Board of Governors of the Federal Reserve System; Office of the Comptroller of the Currency, Washington, DC, Jun. 2023. Available: *https://www.fdic.gov/news/financial-institution-letters/2023/fil23029.html*

[31] European Banking Authority, "EBA guidelines on outsourcing arrangements," EBA, 2019. Available: *https://www.eba.europa.eu/sites/default/files/documents/10180/2551996/38c80601-f5d7-4855-8ba3-702423665479/EBA%20revised%20Guidelines%20on%20outsourcing%20arrangements.pdf*

[32] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," U.S. Department of Commerce, NIST AI 100-1, 2023. doi: *10.6028/NIST.AI.100-1*.

[33] U.S. Department of the Treasury, "Uses, opportunities, and risks of artificial intelligence in financial services," U.S. Department of the Treasury; Available: *https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf*, 2024.

[34] S. Lebovitz, N. Levina, and H. Lifshitz-Assaf, "Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what," *MIS Quarterly*, vol. 45, no. 3, pp. 1501–1525, 2021, doi: *10.25300/MISQ/2021/16564*.

[35] A. Chandrasekaran, "The 2025 hype cycle for GenAI highlights critical innovations." [Online]. Available: *https://www.gartner.com/en/articles/hype-cycle-for-genai*

[36] Deloitte, "Generative AI for enterprises." [Online]. Available: *https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/articles/generative-ai-for-enterprises.html*

37] AI Index Steering Committee, "AI index report 2024." Stanford Institute for Human-Centered AI, 2024. Available: *https://hai.stanford.edu/ai-index/2024-ai-index-report*

[38] IBM Institute for Business Value, "2025 global outlook for banking and financial markets," Jan. 2025. Available: *https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/2025-banking-financial-markets-outlook*

[39] ABA Banking Journal, "Survey: Banks boosting cybersecurity due to AI while also investing in technology," Jun. 2025, Available: *https://bankingjournal.aba.com/2025/06/survey-banks-boosting-cybersecurity-due-to-ai-while-also-investing-in-technology/*

[40] R. Bommasani et al., "The 2024 Foundation Model Transparency Index." [Online]. Available: *http://arxiv.org/abs/2407.12929*

[41] Google Cloud, "Generative AI on vertex AI: deprecations." Available: *https://cloud.google.com/vertex-ai/generative-ai/docs/deprecations*, 2025.

[42] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards; Technology, Gaithersburg, MD, NIST Special Publication 1270, 2023. Accessed: Oct. 09, 2025. [Online]. Available: *https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf*

[43] Basel Committee on Banking Supervision, "Principles for operational resilience." Bank for International Settlements, 2021. Available: *https://www.bis.org/bcbs/publ/d516.htm*

[44] S. Kapoor et al., "On the Societal Impact of Open Foundation Models." Accessed: Nov. 17, 2025. [Online]. Available: *http://arxiv.org/abs/2403.07918*

[45] A. Scher, "Observations About LLM Inference Pricing," MIRI Technical Governance Team, Mar. 2025. Accessed: Nov. 13, 2025. [Online]. Available: *https://techgov.intelligence.org/blog/observations-about-llm-inference-pricing*

[46] A. Ajibode, A. A. Bangash, F. R. Cogo, B. Adams, and A. E. Hassan, "Towards Semantic Versioning of Open Pre-trained Language Model Releases on Hugging Face." Accessed: Nov. 13, 2025. [Online]. Available: *http://arxiv.org/abs/2409.10472*

[47] B. Laufer, H. Oderinwale, and J. Kleinberg, "Anatomy of a Machine Learning Ecosystem: 2 Million Models on Hugging Face." Accessed: Nov. 13, 2025. [Online]. Available: *http://arxiv.org/abs/2508.06811*

[48] D. McDuff et al., "On the Standardization of Behavioral Use Clauses and Their Adoption for Responsible Licensing of AI." Accessed: Nov. 13, 2025. [Online]. Available: *http://arxiv.org/abs/2402.05979*

[49] N. Suggs and P. Venables, "Shared fate: Protecting customers with generative AI indemnification." Accessed: Nov. 13, 2025. [Online]. Available: *https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification*

[50] "LMArena leaderboard overview." Accessed: Nov. 13, 2025. [Online]. Available: *https://lmarena.ai/leaderboard*

[51] Prudential Regulation Authority, "Outsourcing and third party risk management," Prudential Regulation Authority, Bank of England, Supervisory Statement SS2/21, Nov. 2025. Accessed: Nov. 13, 2025. [Online]. Available: *https://www.bankofengland.co.uk/prudential-regulation/publication/2021/march/outsourcing-and-third-party-risk-management-ss*

[52] D. Sculley et al., "Hidden technical debt in machine learning systems," in Advances in neural information processing systems (NeurIPS), 2015. Available: *https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf*

[53] A. D'Amour et al., "Underspecification presents challenges for credibility in modern machine learning," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 118, no. 15, 2020. Available: *https://arxiv.org/abs/2011.03395*

[54] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in *Proceedings of the 58th annual meeting of the association for computational linguistics (ACL)*, 2020. Available: *https://arxiv.org/abs/2005.04118*

[55] R. Cantini, A. Orsino, M. Ruggiero, and others, "Benchmarking adversarial robustness to bias elicitation in large language models: Scalable automated assessment with LLM-as-a-judge," *Machine Learning*, vol. 114, p. 249, 2025, doi: 10.1007/s10994-025-06862-6.

[56] N. G. Packin, "Consumer Finance and AI: The Death of Second Opinions?" NYUJ Legis. *& Pub. Pol'y,* vol. 22, p. 319, 2019, Accessed: Mar. 11, 2025. [Online]. Available: *https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nyulpp22&section=10*

[57] N. Kordzadeh and M. Ghasemaghaei, "Algorithmic bias: Review, synthesis, and future research directions," *European Journal of Information Systems*, vol. 31, no. 3, pp. 388–409, May 2022, doi: 10.1080/0960085X.2021.1927212.

[58] A. Bajracharya, U. Khakurel, B. Harvey, and D. B. Rawat, "Recent Advances in Algorithmic Biases and Fairness in Financial Services: A Survey," in *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 1*, vol. 559, K. Arai, Ed., Cham: Springer International Publishing, 2023, pp. 809–822. doi: 10.1007/978-3-031-18461-1_53.

[59] D. Sengar, "Implications of Algorithmic Bias in Financial Services: A Survey of Sources and Levels of Algorithmic Bias Contributing to Social Implications," in *Revolutionizing the Global Stock Market: Harnessing Blockchain for Enhanced Adaptability*, IGI Global, 2024, pp. 60–82. Accessed: Mar. 11, 2025. [Online]. Available: *https://www.igi-global.com/chapter/implications-of-algorithmic-bias-in-financial-services/344541*

[60] K. Okamura and S. Yamada, "Adaptive trust calibration for human-AI collaboration," PLOS *ONE*, vol. 15, no. 2, p. e0229132, Feb. 2020, doi: 10.1371/journal.pone.0229132.

[61] H. Mahmud, A. K. M. N. Islam, S. I. Ahmed, and K. Smolander, "What influences algorithmic decision-making? A systematic literature review on algorithm aversion," *Technological Forecasting and Social Change*, vol. 175, p. 121390, Feb. 2022, doi: 10.1016/j.techfore.2021.121390.

[62] E. Goh et al., "Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial," *JAMA Network Open*, vol. 7, no. 10, p. e2440969, Oct. 2024, doi: 10.1001/jamanetworkopen.2024.40969.

[63] C. Zhai, S. Wibowo, and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review," *Smart Learn. Environ.*, vol. 11, no. 1, p. 28, Jun. 2024, doi: 10.1186/s40561-024-00316-7.

[64] M. Gerlich, "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking," *Societies*, vol. 15, no. 1, 1, p. 6, Jan. 2025, doi: 10.3390/soc15010006.

[65] C. Natali, L. Marconi, L. D. Dias Duran, M. Miglioretti, and F. Cabitza, "AI-Induced Deskilling in Medicine: A Mixed Method Literature Review for Setting a New Research Agenda." Accessed: Mar. 11, 2025. [Online]. Available: *https://papers.ssrn.com/abstract=5166364*

[66] H. Wen and F. Khan, "A risk-based model for human-artificial intelligence conflict resolution in process systems," *Digital Chemical Engineering*, vol. 13, p. 100194, Dec. 2024, doi: 10.1016/j.dche.2024.100194.

The Alan Turing Institute

turing.ac.uk

@turinginst