Model Risk Management (MRM) – Scaling AI within Compliance Requirements

Capgemini invent

Contents

Executive summary

The banking sector is undergoing a significant transformation due to the rapid adoption of artificial intelligence (AI) and generative AI (Gen AI) technologies. These systems, including large language models (LLMs), offer substantial benefits such as enhanced customer experience through personalized interactions and chatbots, and improved operational efficiency via automation of tasks like document analysis and compliance checks. Additionally, Gen AI is being explored for complex functions like risk management, fraud detection, credit underwriting, and regulatory compliance, potentially contributing up to $340 billion annually to the banking sector through productivity gains.

However, deploying Gen AI introduces new risks. These models often operate as **"black boxes"**, making their decision-making processes opaque and challenging traditional validation methods. Risks include generating incorrect outputs ("hallucinations"), amplifying biases, ensuring data privacy and security, maintaining robustness against attacks, and ensuring accountability for outputs. Moreover, these tools are predominantly developed by third-party vendors, making access to the code difficult.

In the U.S. banking sector, the Federal Reserve's SR 11-7 guidance remains the key framework for model risk management (MRM). Although SR 11-7 predates modern AI and Gen AI, its principles are still applicable, broadly defining a model to include many AI/machine learning (ML) and Gen AI applications. It is recommended to align EU AI Act requirements with existing SR 11-7 dimensions that MRM teams already assess through their standard framework. Nonetheless, it is crucial for banks to correctly capture the model risk inherent in the use of these new technologies.

02

The challenge: Circumvent the "black box" nature of Gen AI models

The term "black box" refers to models where the internal mechanisms—how inputs are transformed into outputs—are opaque or exceedingly complex for humans to comprehend. Although banking has long managed vendor models with black box characteristics, particularly in areas such as anti-money laundering (AML) and fraud detection, Gen AI models, especially large language models (LLMs) built on deep learning architectures like transformers, significantly exacerbate this challenge.

Several factors contribute to this amplification:

- *Extreme complexity:* Gen AI models often involve billions of parameters and intricate neural network architectures, rendering their decision-making processes inherently inscrutable even to their developers. Traditional methods of reviewing model code or logic become impractical or impossible.
- *Vast, unstructured training data:* These models are typically trained on massive, diverse datasets scraped from the internet, including text, images, and code. Understanding the specific data points influencing any given output is extremely difficult, as this data often contains inherent biases, inaccuracies, or even harmful content.
- *Emergent capabilities and non-determinism:* Gen AI models can exhibit unexpected capabilities that are not explicitly programmed, and their outputs can be probabilistic rather than deterministic, meaning the same input might not always produce identical output. This variability complicates traditional validation approaches focused on consistent, predictable results.

- ***Generative nature:*** Unlike predictive models that output scores or classifications, Gen AI creates new content (text, code, images). Validating the quality, relevance, coherence, factual accuracy, and safety of this generated content requires different techniques and metrics than those used for assessing predictive accuracy.
- ***Limited transparency of foundation models:*** Many Gen AI applications leverage pre-trained foundation models (FMs) developed by third parties (vendors or open source communities). Access to the developmental details, training data, and internal workings of these FMs is often limited, posing significant challenges for conceptual soundness evaluation.

These characteristics directly challenge the core elements of SR 11-7 validation. Evaluating conceptual soundness becomes difficult without transparency into the model's design and data. Outcome analysis requires new metrics beyond traditional accuracy measures to assess generated content quality, safety, and relevance. Ongoing monitoring must track novel risks like hallucination drift or emergent biases. The lack of explainability hinders the ability to understand why a model produces a certain output, complicating risk assessment, bias detection, and justification to stakeholders and regulators.

Therefore, it is necessary to adapt the traditional MRM framework to enable firms to capture the model risk inherent in the use of these new models.

03

Governance’s role: Adapting Model Risk Management (MRM) framework to AI/Gen AI models’ specifics

Understanding the Scope of Model Risk Management

The primary distinguishing factor between a **“non model”** and a **“model”** is the uncertainty resulting from assumptions made by the developer. When decisions are based on assumptions that introduce uncertainty into the reasoning process, the categorization shifts to “model”, necessitating validation according to SR11-7 requirements.

Regardless of whether it is a simple regression or an exotic derivative pricing model, banks are required to validate all their models under SR11-7 requirements. This also applies to new AI and Gen AI use cases developed recently, as they all depend on a set of assumptions that need to be checked.

Reinforcing Governance Components to Cover AI/ Gen AI Specific Model Risk

While specifics of AI/Gen AI models are included in the current model risk management framework due to their categorization as models, it is important to establish certain prerequisites for their validation.

The enhancement of model risk governance should cover the following items:

- ***Policies and Procedures:*** MRM policies and procedures should be reviewed regularly to address specifics and validation requirements of AI models. Regulations around AI models need to be monitored closely.
- ***Roles and Responsibilities:*** A strong governance structure requires clearly defined roles and responsibilities for every party involved in the AI model lifecycle, including model owners, developers, implementers, end-users, data stewards, and MRM as the independent validation function.
- ***Model Inventory:*** Specific rules should be established to identify AI/Gen AI models, ensuring they are recorded and easily identifiable within the model inventory. There should also be specific validation rules for AI models regarding frequency. Additionally, the tiering of Gen AI models may need refinement to integrate new weights/scales based on dimensions already used (usage, complexity).
- ***Documentation:*** Comprehensive documentation and relevant model literature are essential for independent parties to understand AI models beyond the "black box" effect. Documentation should include details on the model's purpose and use, design, data sources, development process, implementation details, testing results, validation findings, limitations, weaknesses, and controls.
- ***Oversight from Board of Directors and Senior Management:*** Active oversight from the board and senior management is crucial for effective Gen AI model risk management.

Validation’s role: Adapting their validation toolkit to provide a relevant assessment of Gen AI model risk

Key considerations within the validation framework

As stated by the SR 11-7 guidance, an effective validation framework should include three core elements:

1. Evaluation of conceptual soundness

Especially when using third party models, the conceptual soundness on Gen AI models needs augmentation. It involves a thorough review of literature and documentation, rigorous assessment of data used for fine tuning systems, validation of the rationale for customization, benchmarking against simpler models to justify complexity, and evaluating explainability.

2. Outcomes analysis

The analysis of the Gen AI models output requires metrics tailored to the specific Gen AI task to assess relevance, coherence, factual accuracy, fluency, toxicity, and fairness. It includes robustness testing, bias and fairness testing, hallucination and toxicity detection, and human evaluation to calibrate automated metrics and assess qualitative aspects of the output.

3. Ongoing monitoring

The ongoing monitoring plan of Gen AI models needs to be enhanced and will need to include new KPIs (compared with more traditional models) to measure hallucination rates, toxicity levels, output relevance drift, and user feedback patterns. It also involves monitoring input/output drift, performance of safety filters, and using automated tools with human oversight.

Validation Toolkit

The validation of AI/Gen AI models requires specialized toolkits of techniques and metrics that address the specific characteristics of those models. We propose a set of evaluations and techniques available to MRM Validation teams to conduct their risk assessment.

1. Evaluation of conceptual soundness

- *Input management & pre-processing:* Gen AI's use of large datasets, which may include sensitive customer information or proprietary data, poses risks related to data quality, bias introduction, security breaches, and compliance with data privacy regulations like GDPR and CCPA. It is important to ensure the integrity and security of data used for training, fine-tuning, and inference.

  Techniques:

  - *Data Quality Assessment:* Implementing thorough processes to verify data accuracy, completeness, consistency, timeliness, and relevance to the intended use case.
  - *Data Lineage and Provenance:* Establishing clear tracking of data origins, transformations, and usage throughout the model lifecycle.
  - *Privacy-preserving Techniques (PPTs):* Using methods to protect sensitive information.
    - Data Anonymization/Pseudonymization/Masking to remove or replace personally identifiable information (PII).
    - Encryption to protect data (both at rest and in transit).
    - Synthetic Data Generation to create artificial data that mimics the statistical properties of real data but contains no actual sensitive information.
    - PII Detection and Filtering by using automated tools to identify and remove sensitive information from datasets or user inputs/outputs.
  - *Security Controls:* Implementing robust technical security measures.
    - Access Controls and Authentication to ensure only authorized personnel and systems can access models and data.
    - Secure Development Lifecycle to integrate security practices throughout model development.
    - Threat Detection and Response by implementing systems to detect and respond to attacks or anomalous activity.
    - Sandboxing to isolate model execution environments to limit potential damage from breaches.
    - Adherence to Frameworks: Utilizing established security frameworks (NIST AI Risk Management Framework).
  - *Data Governance:* Establishing and enforcing clear policies, standards, roles, and responsibilities for data handling, access, usage, quality, security, and privacy compliance across the organization.

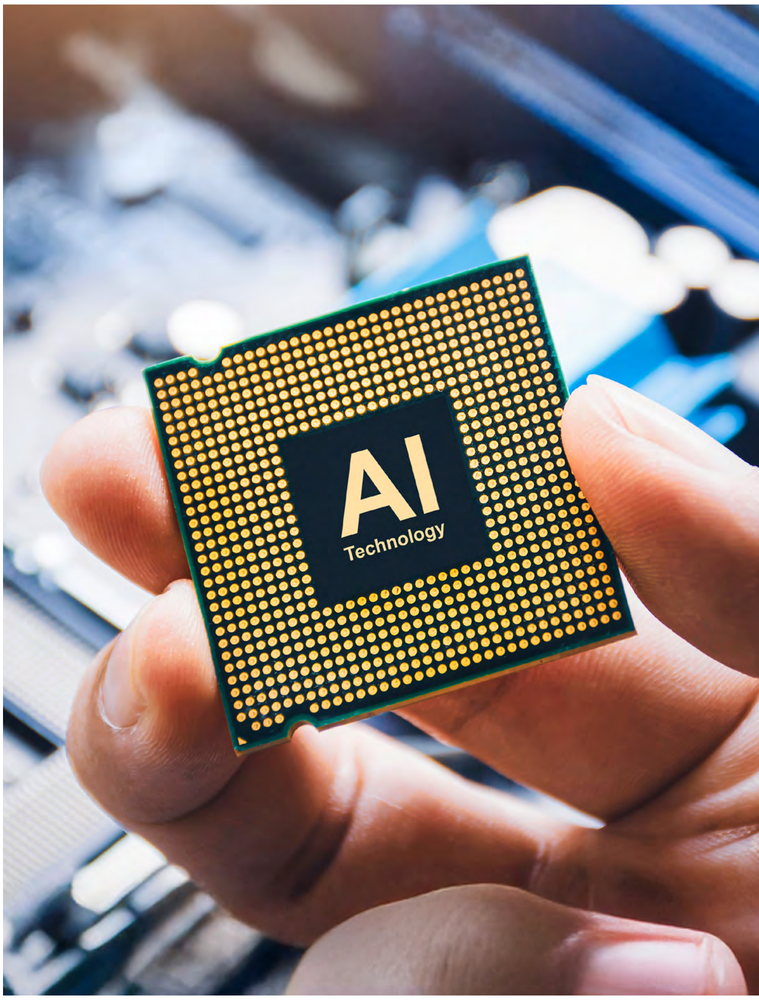
- *Core LLM & Augmentation:* Ensuring consistent and reliable model performance under diverse conditions, including noisy inputs, edge cases, and deliberate manipulation (adversarial attacks).

  Techniques:

  - *Input Perturbation Testing:* Evaluating model sensitivity by making small changes to input prompts to see if the output meaning remains stable.
  - *Stress Testing:* Assessing model performance under extreme or unexpected conditions.

- *Adversarial Testing:* Actively attacking the model with inputs designed to bypass safety filters or elicit harmful/undesired outputs including:
    - **Jailbreaking/Prompt Injection:** Crafting prompts to trick the model into violating safety policies or generating prohibited content.
    - **Frameworks and Benchmarks:** Using structured adversarial testing approaches and comparing robustness against known attack benchmarks.
- *Security Testing:* Evaluating the model and its infrastructure for vulnerabilities related to data security, access controls, specific AI threats, membership inference, and model theft.
- *Consistency and Reliability:* Evaluating the consistency of model outputs across runs and data subsets or time periods and ensuring reproducibility.

- *Explainability & Logging:* Addressing the “black box” nature of complex Gen AI models to meet regulatory expectations, provide clear reasons for the model’s decisions, and enable users’ understanding of how the model arrived at a particular outcome.



Techniques:

- *Model-agnostic Local Explanations (LIME):* LIME works by perturbing individual inputs and fitting a simpler, interpretable model to the local region around that input’s prediction. This explains why a specific prediction was made by highlighting influential input features (e.g., words in text). It’s useful for understanding individual outcomes but can be sensitive to perturbation settings and may lack consistency.
    - **Inherently Interpretable Models:** Where feasible and performance is adequate, favoring models that are naturally easier to understand, such as linear/logistic regression, decision trees, or rule-based systems.
    - **Proxy/Surrogate Models:** Training a simpler, interpretable model to mimic the behavior of the complex black-box model. The simpler model’s logic can then be analyzed as an approximation.
- *Attention Mechanisms:* For models based on architectures like transformers, visualizing attention weights can show which parts of the input sequence the model focused on when generating a specific part of the output.
- *Feature Engineering:* Creating input features that are more interpretable to humans can aid understanding.
- *Emerging Techniques:* Research is ongoing into methods like self explanation (prompting the model to explain its reasoning) and advanced visualization techniques.
- *Limitations:* Post-hoc explanation methods like LIME provide approximations of the model’s behavior and may not fully capture the intricate non-linear interactions within deep learning models. Their utility for very large foundation models might be limited.

2. Outcomes analysis and ongoing monitoring

We merge both topics as they can be assessed at initiation and on a continuous basis through an Ongoing Monitoring plan.

- *Output processing & Guardrails:* Gen AI models can inherit and amplify biases from their training data, leading to unfair outcomes in applications.

Techniques:

- *Bias Detection:*
  - Guardrail models to detect biases (e.g., gender, race).
  - Evaluate performance on benchmark datasets to reveal biases.
  - Subgroup analysis to analyze performance metrics for different demographic groups.
  - Adversarial Testing/Red Teaming to uncover hidden biases.
  - Audit of the training data for representational biases.
- *Fairness Metrics:* Selection of appropriate statistical fairness metric based on context.
- *Mitigation Strategy:*
  - Pre-processing: Adjust training data to reduce bias.
  - In-processing: Modify algorithms to include fairness constraints.
  - Post-processing: Adjust outputs to achieve fairness goals.
  - Human Review: Ensure human oversight to review and correct biases.
- *Toxicity Detection:* Score outputs for various categories of toxicity (e.g., harassment, hate speech) using specialized classifier or guardrails and set acceptable thresholds based on use case and risk appetite.

- *Explainability & Logging:* Evaluating if the generated content is fit for its intended purpose, as traditional accuracy measures are not sufficient to address the subjective and variable nature of Gen AI outputs.

Techniques:

- *Relevance and Coherence:* Evaluate if the output directly addresses the prompt (relevance) and flows logically (coherence).
- *Accuracy/Factual Consistency (Hallucination Detection):* Detect outputs inconsistent with provided source information or established facts. Methods include Natural Language Inference (NLI), Self-check Methods, Fact-checking/Verification Methods, and Human Evaluation.
- *Fluency and Readability:* Assess the grammatical correctness and naturalness of the generated language.
- *Task-specific Metrics:* Tailor evaluation metrics to the specific use case.
- *Human Oversight:* Structured human evaluation on a sample basis ensures the Gen AI model goals are met.

05

Get set to be MRM-compliant by design

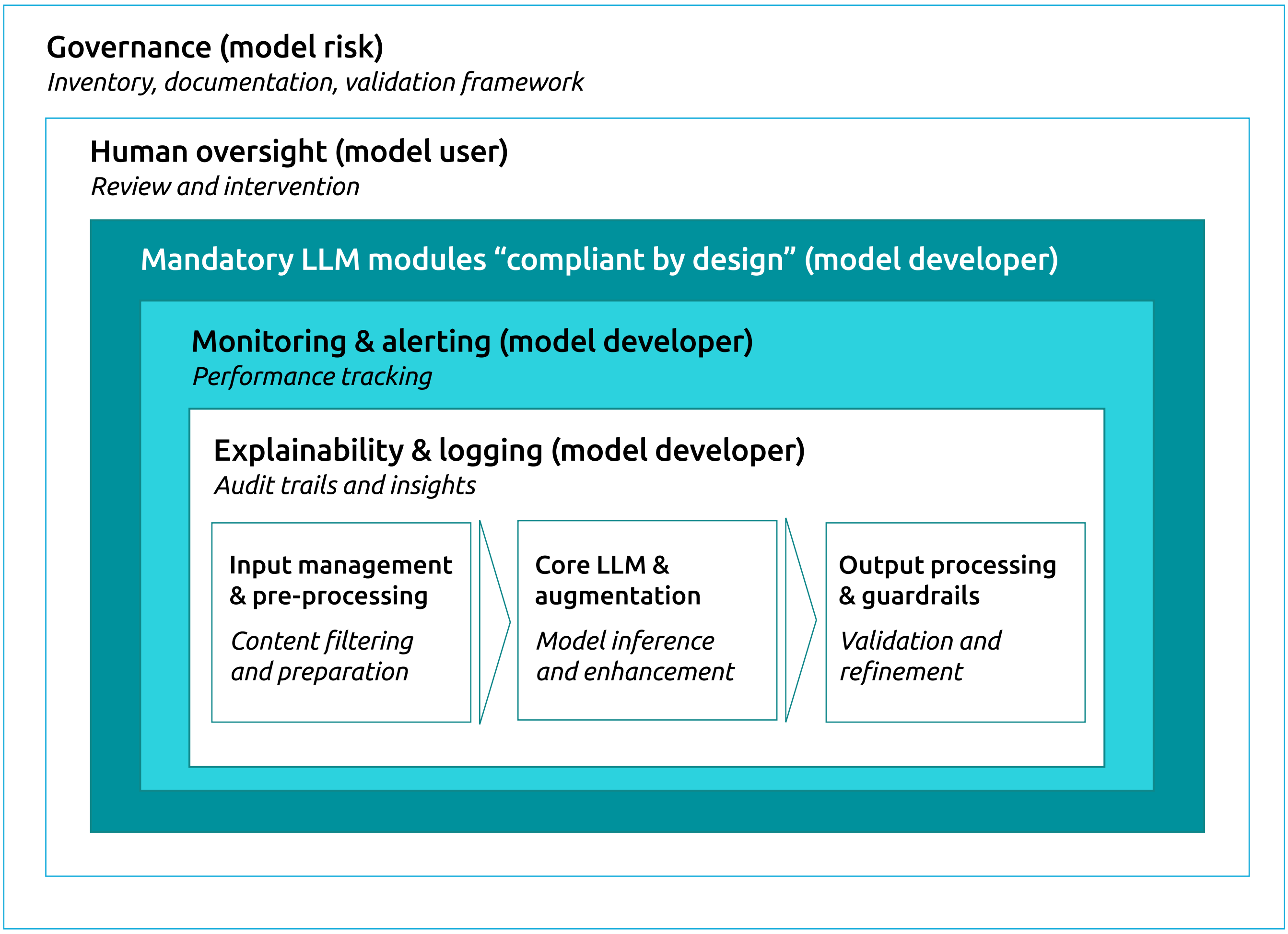
The black box characteristics of Gen AI models make it complex for model risk validation teams to assess the model risk efficiently. Model developers bear significant responsibility to enable model validation teams to assess the risk of such models to the extent possible. Leveraging the EU AI Act dimensions, model developers should develop a functional architecture allowing them to break the "black box" and provide model risk teams with more information, notably regarding the explainability of the outputs generated by the Gen AI model.

As model risk management procedures evolve, with a focus on ongoing monitoring (i.e., ensuring models perform as expected), a core element of the validation framework will be the capacity to explain outputs of the models. This means model developers will need to provide such information or risk a potential no-go from model risk teams.

Alongside the model risk framework developed by the company, the model developer can build functional layers that provide the necessary information for model risk validation, such as:

- ***Explainability and logging:*** This layer facilitates model risk management by providing comprehensive documentation and transparency. The detailed logs and explainability tools enable validation teams to thoroughly analyze model behavior, verify compliance with policies, and investigate any questionable outputs, making MRM validation significantly more straight forward.

- *Monitoring and alerting system:* The monitoring and alerting system facilitates straightforward model risk management periodic reviews through comprehensive dashboards, automated performance reporting, and continuous performance monitoring (as per frequency agreed with MRM).
- *Input management & pre-processing:* This layer provides transparency for MRM validation through comprehensive documentation of all pre-processing steps and filtering decisions.
- *Core LLM & augmentation:* Validation teams can validate this layer by reviewing model documentation, evaluating selection criteria, assessing version control protocols, and examining and fine tuning datasets. For RAG (Retrieval Augmented Generation) implementations, knowledge source curation processes and retrieval accuracy can be independently tested to verify information quality and relevance.
- *Output processing and guardrails:* This layer offers strong validation opportunities through direct testing of guardrails against known problematic inputs. MRM teams can independently verify each guardrail component (automated checks, factual verification, sanitization, and policy enforcement) with clear pass/fail criteria and documented examples of intercepted issues.


Governance (model risk)
Inventory, documentation, validation framework
Human oversight (model user)
Review and intervention
Mandatory LLM modules "compliant by design" (model developer)
Monitoring & alerting (model developer)
Performance tracking
Explainability & logging (model developer)
Audit trails and insights
Input management & pre-processing
Content filtering and preparation
Core LLM & augmentation
Model inference and enhancement
Output processing & guardrails
Validation and refinement

Anticipating MRM adjustments to include AI/Gen AI models' specifics

Even though there are no specific AI requirements with regards to model risk, US banks can leverage what has been developed through the **AI Act** enforced in the European Union (EU AI Act). This act provides a defined framework with clear requirements and obligations for specific uses of AI. By aligning each dimension of the EU AI Act with the SR 11-7 dimensions, we ensure that responsible AI development and deployment are integrated into the existing framework of model risk management.

The dimensions below should be considered during the validation of the model, so model developers should get ready and be prepared to adapt to new questions from MRM as per proposed mapping below:

EU AI Act dimension	Corresponding SR 11-7 dimension	Mapping explanation
Fairness Ensure fairness and impartiality in the design, deployment, and impact to prevent bias and discrimination	**Governance** Compliance with internal standards in relation with the monitoring of the model throughout its lifecycle	Model should be developed so that they comply with internal rules, which involves complying to procedures, including non-discrimination and reducing bias
Explainability Provide clear reasons for the model's decisions and enable users' understanding of how the model delivered this particular outcome	**Methodology and design** Explain the specifics of the model design methodology (e.g. methodology characteristics, limitations, behavior)	Developer should be able to explain how the model works and how the results are obtained
Frugality Assess computational costs, memory usage, and energy consumption against performance needs	**Model inputs** Assess the risk related to the quality of raw data, controls, and processing of data	Both topics relate to data management
Drift control Ensure the model remains effective even if data distribution changes	**Methodology and design** Assess the risk associated with the use of a model design methodology (methodology characteristics, limitations, behavior)	Drift control is part of methodology assessment as the choice of the methodology should consider potential changes in data distribution
Data quality Assess the risk related to the quality of raw data, controls, and processing of data	**Model inputs** Assess the risk related to the quality of raw data, controls, and processing of data	Data quality is a sub-dimension of model input assessment under SR 11-7
Robustness Assess model ability to handle challenging scenarios beyond the training data (e.g. noisy data)	**Methodology and design** Assess the risk associated with the use of a model design methodology (methodology characteristics, limitations, behavior)	The choice of the model should be assessed against its ability to provide robust results
Performance Assess the risk related to the model performance in the design phase, or periodically to ensure testing at the end of design phase and monitoring/back-testing results are satisfactory	**Model performance** Assess the risk related to the model performance in the design phase, or periodically to ensure testing at the end of design phase and monitoring/back-testing results are satisfactory	Same terminology - no difference with classical models
Responsibility Assess ethical and accountable use of the model development (development should adhere to ethical guidelines)	**Governance** Assess compliance with internal standards in relation with the monitoring of the model throughout its lifecycle	Developers should follow internal guidelines including ethical guidelines,in compliance with the governance in place

06

Seize the moment to revisit vendor risk policies

The majority of AI solutions utilized in the industry are provided by third-party vendors. Consequently, banks often have limited control over the development of these models, lack sufficient documentation (falling short of SR 11-7 requirements), and face challenges in enforcing methodology revisions or obtaining ongoing monitoring and testing of the model over time. This results in insufficient understanding of the performance of such models.

While banks aim to leverage cutting-edge technologies, they must ensure that appropriate vendor policies are established. This includes not only cybersecurity questionnaires but also involving MRM teams at the appropriate stages to verify that all key requirements are addressed.

As previously explained, the use of potential personal data necessitates comprehensive governance regarding its treatment. Privacy and confidentiality must be effectively managed to prevent PII/ NPI leakage, unauthorized data access, and non-compliance with privacy laws such as GDPR and CCPA.

Conclusion

As financial institutions embrace the transformative potential of Gen AI in KYC and broader risk management functions, they must also confront the unique and complex risks these technologies introduce. The traditional model risk management frameworks, while foundational, require thoughtful adaptation to address the opacity, variability, and ethical considerations inherent to Gen AI systems. By **aligning SR 11-7 principles with the EU AI Act dimensions**, institutions can build a more robust, forward-looking governance framework that ensures fairness, explainability, robustness, and accountability.

Moreover, the integration of specialized validation toolkits – ranging from hallucination detection and fairness audits to adversarial robustness testing and privacy-preserving techniques – enables a more nuanced and effective assessment of Gen AI models.

Ultimately, the successful deployment of Gen AI in regulated environments hinges not only on technical excellence but also on proactive governance, rigorous validation, and a culture of responsible innovation. Institutions that invest in these capabilities today will be better positioned to harness the full value of Gen AI while safeguarding trust, compliance, and resilience.

MRM teams should adapt their current framework to correctly prepare for these new technologies and models to be assessed as thoroughly as they should be. While MRM teams might be ready in terms of knowledge and technical skills, they should prepare for a high volume of new Gen AI solutions to come and be ready to follow the business urge to leverage those technologies.

On the other hand, **Business Lines should** also prepare for new requirements from MRM teams to avoid any blocking situation before benefitting from the full power of these technologies.

Authors

Victor Aubelle

Manager

victor.aubelle@capgemini.com

Patrick Bucquet

Financial Services Lead

patrick.bucquet@capgemini.com

About Capgemini Invent

As the digital innovation, design and transformation brand of the Capgemini Group, Capgemini Invent enables CxOs to envision and shape the future of their businesses. Located in over 30 studios and more than 60 offices around the world, it comprises a 12,500+ strong team of strategists, data scientists, product and experience designers, brand experts and technologists who develop new digital services, products, experiences and business models for sustainable growth.

Capgemini Invent is an integral part of Capgemini, a global business and technology transformation partner, helping organizations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 340,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fueled by its market leading capabilities in AI, generative AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2024 global revenues of €22.1 billion.

Get the future you want | www.capgemini.com /invent

Capgemini invent