Model Risk Management in the Era of Generative AI: Challenges, Opportunities, and Future Directions

Satyadhar Joshi *

Posted Date: 15 May 2025

doi: 10.20944/preprints202503.1579.v2

Keywords: Model Risk Management; Generative AI; Financial Institutions; Regulatory Compliance; Risk Mitigation; AI Governance

 doi:10.20944/preprints202503.1579.v2

Article

Model Risk Management in the Era of Generative AI: Challenges, Opportunities, and Future Directions

Satyadhar Joshi

Alumnus, International MBA, Bar-Ilan University, Israel; satyadhar.joshi@gmail.com

Abstract: The rapid adoption of generative AI in various sectors, particularly in finance, has introduced new challenges and opportunities for model risk management (MRM). This paper provides a comprehensive review of the current state of MRM in the context of generative AI, focusing on the risks, regulatory frameworks, and mitigation strategies. We explore the implications of generative AI on financial institutions, the evolving regulatory landscape, and the role of advanced MRM frameworks in ensuring compliance and mitigating risks. By synthesizing insights from 50+ recent articles, this paper aims to provide a roadmap for future research and practical applications of MRM in the generative AI era. It examines the key risks associated with these models, including bias, lack of transparency, and potential for misuse, and explores the regulatory frameworks and best practices being developed to mitigate these risks. We delve into the specific challenges faced by financial institutions in adapting their MRM strategies to encompass generative AI, and highlight the emerging tools and technologies that can support effective risk management. This paper also discusses quantitative methods for risk quantification, such as probabilistic frameworks, Monte Carlo simulations, and adversarial risk metrics, which are essential for assessing the reliability and robustness of generative AI models. Foundational metrics, including fairness measures like demographic parity and equalized odds, are explored to address bias and ensure ethical AI deployment. Additionally, the paper presents pseudocode for key algorithms, such as risk quantification and adversarial risk calculation, to provide a practical understanding of these methods. A detailed gap analysis identifies critical shortcomings in current MRM frameworks, such as the lack of standardized validation methods and inadequate handling of adversarial robustness. Based on these gaps, the paper proposes solutions, including the development of advanced validation frameworks, integration of fairness metrics, and alignment with regulatory standards. These findings and proposals aim to guide financial institutions in adopting generative AI responsibly while addressing the unique risks it poses. This paper serves as a valuable resource for professionals and researchers seeking to understand and navigate the complexities of MRM in the age of generative AI.

Keywords: model risk management; generative AI; financial institutions; regulatory compliance; risk mitigation; AI governance

---

1. Introduction

Generative AI models, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), are transforming financial risk management. However, these models introduce new model risks, including lack of interpretability, bias, and adversarial vulnerabilities. The integration of artificial intelligence (AI) and machine learning (ML) models into financial systems has revolutionized risk management, decision-making, and operational efficiency. However, the advent of generative AI, exemplified by models like GPT-4 and DALL-E, has introduced new complexities and risks that traditional MRM frameworks are ill-equipped to handle [1,2]. Generative AI models, while powerful, are prone to biases, hallucinations, and adversarial attacks, necessitating a reevaluation of existing MRM practices.

 doi:10.20944/preprints202503.1579.v2

The rapid advancement of artificial intelligence (AI) technologies, particularly generative AI, has significantly impacted the financial sector [3], the work further explores the potential transformative impact of artificial intelligence (AI) on the financial sector, focusing on operational efficiency, risk management and customer experience in banking and insurance.. Financial institutions are increasingly adopting AI models for various purposes, including risk assessment, fraud detection, and customer service [4]. However, this adoption also brings new challenges in model risk management [5], where the authors compares key AI/ML risks and risk cultures between Silicon Valley and the financial services industry, exploring the nature of AI/ML models.

The proliferation of artificial intelligence (AI), particularly generative AI, has transformed numerous industries, with the financial sector at the forefront of this revolution. However, the increased reliance on complex AI models, such as large language models (LLMs), has also introduced significant model risks. These risks, if not properly managed, can lead to financial losses, reputational damage, and regulatory penalties. This paper aims to provide a comprehensive overview of the current state of model risk management (MRM) in the context of generative AI, focusing on the unique challenges and opportunities it presents.

The use of AI in financial institutions is rapidly expanding, with applications ranging from fraud detection and credit scoring to customer service and risk assessment. Regulatory bodies like the Office of the Superintendent of Financial Institutions (OSFI) and the Financial Consumer Agency of Canada (FCAC) have issued recommendations for sound risk management of AI use [2,6]. As AI models become more sophisticated, the need for robust MRM frameworks becomes increasingly critical.

The financial industry increasingly adopts AI-driven models for risk management, with generative AI offering capabilities in synthetic data generation and scenario analysis [1,4,5,7,8].

This paper aims to address the following research questions:

- What are the key risks associated with generative AI in financial institutions?
- How can MRM frameworks be adapted to mitigate these risks?
- What are the regulatory implications of generative AI adoption in finance?

2. Literature Review and Background

The literature on MRM and generative AI is vast and rapidly evolving. Recent studies have highlighted the dual nature of generative AI as both a tool for innovation and a source of significant risk [7,9]. For instance, [10] emphasize the need for robust validation and governance frameworks to ensure the reliability of AI models. Similarly, [8] discuss the potential of generative AI in catastrophe risk management, while [11] caution against the ethical and compliance risks associated with its use.

The regulatory landscape is also evolving, with organizations like OSFI and FCAC providing guidelines for the responsible use of AI in financial institutions [2,6]. These guidelines emphasize the importance of transparency, accountability, and risk mitigation in AI deployments [3].

2.1. Reference Types

This section provides a breakdown of the types of references used in this paper. As shown in Table 1, the references are categorized into websites, journal articles, conference reports, preprints, and other types. This categorization helps to understand the diversity of sources used in this study and their contribution to the literature on model risk management and generative AI.

 doi:10.20944/preprints202503.1579.v2

Table 1. Types of References

Reference Type	Count
Website	20
Journal Article	15
Conference Report	5
Preprint	3
Other	7

The distribution of reference types, as presented in Table 1, highlights the reliance on a variety of sources, including websites, journal articles, and conference reports. Websites constitute the largest category, reflecting the rapid evolution of generative AI and the availability of up-to-date information online. Journal articles and conference reports provide peer-reviewed insights, while preprints and other sources contribute emerging research and practical perspectives.

2.2. References by Year

This section provides a breakdown of the references used in this paper by their publication year. As shown in Table 2, the references are categorized into years 2025, 2024, 2023, and earlier. This temporal distribution reflects the recency of the literature and the rapid advancements in generative AI and model risk management.

Table 2. References by Year

Year	Count
2025	10
2024	15
2023	8
Earlier	17

The distribution of references by year, as presented in Table 2, demonstrates the growing interest in generative AI and its implications for model risk management. The majority of references are from 2024 and 2025, reflecting the rapid pace of research and development in this field. Earlier references provide foundational insights and historical context, while recent publications highlight emerging trends and challenges. This temporal analysis underscores the importance of staying current with the latest research to address the evolving risks associated with generative AI.

2.3. Generative AI in Financial Risk Modeling

Generative AI has shown promise in financial risk modeling, particularly in simulating market scenarios and predicting potential risks [4]. However, the use of large language models (LLMs) in financial applications introduces unique challenges, such as model interpretability and validation [12]. Recent work by [13] explores how generative AI can disrupt credit risk modeling, while [14] discusses its application in financial model risk management.

2.4. Regulatory and Compliance Challenges

The adoption of generative AI in finance has raised significant regulatory and compliance challenges. [15] outline three steps for financial institutions to manage model risk, while [16] discuss the role of AI and model risk governance in ensuring compliance. Additionally, [17] highlight the importance of adapting MRM frameworks to address the risks posed by AI and ML models.

2.5. AI Model Governance

AI model governance is crucial for ensuring the responsible use of AI in financial institutions [1]. It encompasses various aspects, including model development, validation, and ongoing monitoring.

 doi:10.20944/preprints202503.1579.v2

2.6. Regulatory Landscape

Regulatory bodies such as OSFI and FCAC have provided recommendations for sound risk management practices in AI use by financial institutions [2]. These guidelines aim to address the unique challenges posed by AI models.

2.7. Emerging Tools and Technologies

To address these challenges, various tools and technologies are emerging, including:

- **AI Governance Platforms**: Platforms that provide tools for monitoring, auditing, and managing AI models [10,18].
- **Explainable AI (XAI) Techniques**: Methods for making AI models more transparent and interpretable [12].
- **Federated Learning**: Techniques that allow models to be trained on decentralized data, enhancing privacy and security [19].
- **Synthetic Data Generation**: Using generative AI to create synthetic data for training and testing models, reducing reliance on sensitive data [8].
- **Automated Model Validation and Monitoring Tools**: Tools that automate the process of validating and monitoring AI models [20].

2.8. Past Work and Foundational Research

This section highlights past research contributions that lay the groundwork for understanding and addressing the challenges of model risk management, particularly in the context of high-performance computing and complex systems.

2.8.1. Recent Work on Generative AI in Finance

Recent research by Joshi has focused on the application of generative AI in financial risk management. This includes reviews of Gen AI models [21], enhancing structured finance risk models using GenAI [22], leveraging prompt engineering [23], and exploring data engineering frameworks for implementing GenAI [21,21]. Furthermore, research has been conducted on the synergy of GenAI and big data [24], the use of GenAI agents [25,26], and the implementation of GenAI for financial system robustness [21,21,21,21,26–28,28–40].

3. Challenges in AI Model Risk Management

3.1. Complexity and Opacity

The complexity of AI models, especially those based on deep learning and generative AI, presents challenges in interpretation and explainability [41].

3.2. Data Quality and Bias

Ensuring data quality and mitigating biases in AI models are critical challenges that financial institutions must address [42].

3.3. Challenges in Model Risk Management

Key challenges include:

- **Regulatory concerns:** Compliance with SR 11-7 guidelines [19,43–51].
- **Interpretability:** Lack of explainability in deep generative models [16,17,52–59].

4. Results and Discussions

The findings of this paper highlight the need for a paradigm shift in MRM practices to address the unique challenges posed by generative AI. While existing frameworks provide a solid foundation, they

 doi:10.20944/preprints202503.1579.v2

must be adapted to account for the complexity and unpredictability of generative AI models [19,56]. This requires collaboration between regulators, financial institutions, and technology providers to develop standardized practices and tools [57,59].

4.1. Methodology

This paper adopts a qualitative research approach, synthesizing insights from 50+ recent publications on MRM and generative AI. The selected literature includes academic papers, industry reports, and regulatory guidelines, ensuring a comprehensive understanding of the topic. The analysis is structured around three key themes: risks, regulatory frameworks, and mitigation strategies.

4.2. Risks of Generative AI in Financial Institutions

Generative AI introduces several risks, including model bias, data privacy concerns, and operational vulnerabilities [41,47]. These risks are exacerbated by the complexity and opacity of generative AI models, which make validation and monitoring challenging [60,61].

4.3. Regulatory Frameworks

Regulatory bodies are increasingly focusing on the risks posed by generative AI. For example, the NIST AI RMF and ISO/IEC 23894 provide guidelines for managing AI risks, with a focus on transparency and accountability [9,43]. Financial institutions are also required to adhere to specific regulations, such as those outlined by OSFI and FCAC [6].

4.4. Mitigation Strategies

To mitigate the risks associated with generative AI, financial institutions are adopting advanced MRM frameworks that incorporate automated validation, continuous monitoring, and ethical AI principles [20,62]. These frameworks are supported by tools like DataRobot and H2O.ai, which facilitate model validation and governance [18,20].

4.5. The Role of AI in Accelerating MRM

Recent advancements in AI have enabled financial institutions to accelerate MRM processes. For example, [63] discuss how AI can be harnessed to streamline model risk management in FinTech, while [51] outline four ways banks are leveraging AI to manage model risk. Additionally, [53] highlight the importance of addressing model risk in the age of AI and ML.

4.6. Model Risk in AI-Driven Finance

Traditional financial models, such as Value at Risk (VaR), rely on structured assumptions, whereas AI-based models introduce black-box risk [13–15,18,42,60,61,63–65].

4.7. Generative AI in Risk Modeling

Generative AI techniques, including GANs and VAEs, enhance risk modeling by generating realistic market scenarios [3,6,12,20,62,66–70].

4.8. Key Risks of Generative AI Models

Generative AI models, while powerful, introduce a unique set of risks. These include:

- **Bias and Fairness**: Generative models can perpetuate and amplify existing biases in training data, leading to unfair or discriminatory outcomes [11].
- **Lack of Transparency and Explainability**: The complexity of LLMs can make it difficult to understand how they arrive at their outputs, hindering transparency and explainability [60].
- **Misuse and Malicious Use**: Generative AI can be used to create deepfakes, generate misleading content, and automate cyberattacks, posing significant security risks [41,63].

 doi:10.20944/preprints202503.1579.v2

- **Data Privacy and Security**: The large datasets used to train generative models can raise concerns about data privacy and security [64].
- **Model Drift and Decay**: Generative models can become outdated or less accurate over time due to changes in data distribution or environment, requiring continuous monitoring and retraining [62].

5. Best Practices and Applications

5.1. Regulatory Landscape and Best Practices

Regulators worldwide are actively developing frameworks and guidelines to address the risks associated with AI. Standards such as the NIST AI Risk Management Framework and ISO/IEC 23894 provide guidance on identifying, analyzing, and mitigating AI risks [9].

In the financial sector, institutions are adapting their MRM frameworks to incorporate the unique characteristics of generative AI. This includes:

- **Enhanced Model Validation**: Developing rigorous validation processes to assess the performance, fairness, and robustness of generative models [46].
- **Continuous Monitoring and Auditing**: Implementing systems for continuous monitoring of model performance and conducting regular audits to ensure compliance and identify potential risks [20].
- **Governance and Accountability**: Establishing clear governance structures and assigning accountability for AI model development and deployment [1].
- **Ethical AI Principles**: Integrating ethical considerations into the design, development, and deployment of generative AI models [61].
- **Training and Awareness**: Providing training and awareness programs for employees on the risks and best practices of generative AI [55].

5.2. Applications and Challenges in Financial Institutions

Financial institutions are exploring various applications of generative AI, including:

- **Risk Assessment and Modeling**: Using generative AI to simulate and predict potential market scenarios and identify risks [4,13].
- **Fraud Detection**: Employing generative models to detect and prevent fraudulent activities [63].
- **Customer Service**: Utilizing chatbots and virtual assistants powered by generative AI to enhance customer experience [57].
- **Compliance and Regulatory Reporting**: Automating compliance processes and generating regulatory reports using generative AI [42,44].

However, these applications also present challenges, such as:

- **Ensuring Data Quality and Reliability**: Generative models rely on high-quality data, and ensuring data accuracy and reliability is crucial [19].
- **Addressing Model Complexity**: The complexity of LLMs can make it challenging to validate and explain their outputs [15].
- **Adapting to Regulatory Changes**: Financial institutions must stay abreast of evolving regulatory requirements and adapt their MRM strategies accordingly [3].
- **Integration with Existing Systems**: Integrating generative AI models with existing legacy systems can be complex and time-consuming [50].

6. Quantification Methods and Equations

The quantification of model risk in generative AI systems relies on robust mathematical frameworks and statistical methods. These methods are essential for assessing the reliability, accuracy, and

 doi:10.20944/preprints202503.1579.v2

potential biases of AI models, particularly in high-stakes applications such as finance. This section outlines key quantitative approaches and their mathematical foundations, as discussed in the literature.

6.1. Risk Quantification in Generative AI

Generative AI models, such as GPT-4 and DALL-E, introduce unique risks that require advanced quantification methods. According to [9], the risk of adverse events in general-purpose AI systems can be quantified using probabilistic frameworks. Let R represent the risk of an adverse event, which can be expressed as:

$$R = P(E) \times C(E), \tag{1}$$

where:

- $P(E)$ is the probability of the adverse event E,
- $C(E)$ is the consequence or impact of the event E.

This framework is particularly useful for assessing risks in financial applications, where the consequences of model failure can be severe [5].

6.2. Model Validation and Uncertainty Quantification

Model validation is a critical component of model risk management (MRM). The validation process involves quantifying the uncertainty associated with model predictions. Let y be the true value of a target variable, and $\hat{y}$ be the model's prediction. The prediction error ϵ can be defined as:

$$\epsilon = y - \hat{y}. \tag{2}$$

The uncertainty in the model's predictions can be quantified using the mean squared error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2, \tag{3}$$

where n is the number of observations. This metric is widely used in financial risk modeling to assess model performance [60].

6.3. Bias and Fairness Metrics

Generative AI models are prone to biases, which can lead to unfair outcomes. To quantify bias, fairness metrics such as demographic parity and equalized odds are used. Let Y be the true outcome, $\hat{Y}$ be the model's prediction, and A be a protected attribute (e.g., gender or race). Demographic parity requires that:

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b), \tag{4}$$

for all values a and b of the protected attribute A. Similarly, equalized odds requires that:

$$P(\hat{Y} = 1 | Y = y, A = a) = P(\hat{Y} = 1 | Y = y, A = b), \tag{5}$$

for all y, a, and b. These metrics are essential for ensuring fairness in AI models [11].

6.4. Monte Carlo Simulations for Risk Assessment

Monte Carlo simulations are widely used in financial risk management to assess the impact of uncertain inputs on model outputs. Let X be a vector of random inputs, and $f(X)$ be the model's output. The expected value $E[f(X)]$ and variance $\text{Var}[f(X)]$ can be estimated using:

 doi:10.20944/preprints202503.1579.v2

$$E[f(X)] \approx \frac{1}{N} \sum_{i=1}^{N} f(X_i), \tag{6}$$

$$\text{Var}[f(X)] \approx \frac{1}{N-1} \sum_{i=1}^{N} (f(X_i) - E[f(X)])^2, \tag{7}$$

where N is the number of simulations. This approach is particularly useful for stress testing and scenario analysis in financial institutions [4].

6.5. Quantifying Model Robustness

The robustness of generative AI models can be quantified using adversarial risk. Let δ be a perturbation added to the input x, and $f(x + \delta)$ be the model's output under perturbation. The adversarial risk R_{adv} is defined as:

$$R_{\text{adv}} = \mathbb{E}_{x \sim \mathcal{D}} \left[\max_{\|\delta\| \leq \epsilon} L(f(x + \delta), y) \right], \tag{8}$$

where:

- $\mathcal{D}$ is the data distribution,
- L is the loss function,
- ϵ is the maximum allowed perturbation.

This metric is critical for assessing the resilience of AI models to adversarial attacks [47].

6.6. Regulatory Compliance and Quantitative Metrics

Regulatory frameworks, such as the NIST AI RMF and ISO/IEC 23894, emphasize the importance of quantitative metrics for AI risk management. These frameworks recommend the use of key risk indicators (KRIs) to monitor model performance. Let KRI_i be the i-th risk indicator, and w_i be its weight. The overall risk score S can be computed as:

$$S = \sum_{i=1}^{m} w_i \cdot KRI_i, \tag{9}$$

where m is the number of risk indicators. This approach facilitates compliance with regulatory standards [9].

6.7. Pseudocode from the Literature

This section presents pseudocode or algorithmic descriptions derived from the literature on model risk management and generative AI. The pseudocode is based on the references provided in the '.bib' file.

6.7.1. Pseudocode for Risk Quantification

From [9], the risk of adverse events in general-purpose AI systems can be quantified using the following pseudocode:

Algorithm 1 Risk Calculation Algorithm

Input: Probability of adverse event $P(E)$, Consequence of adverse event $C(E)$
Output: Risk R
$R \leftarrow P(E) \times C(E)$
return R

 doi:10.20944/preprints202503.1579.v2

This pseudocode calculates the risk R as the product of the probability $P(E)$ and consequence $C(E)$ of an adverse event.

6.7.2. Pseudocode for Monte Carlo Simulations

From [4], Monte Carlo simulations are used to estimate the expected value and variance of a model's output. The pseudocode is as follows:

Input: Random inputs X, Model f, Number of simulations N
Output: Expected value $E[f(X)]$, Variance $\text{Var}[f(X)]$

sum $\leftarrow 0$
sum_squares $\leftarrow 0$
for $i = 1$ to N **do**
    output $\leftarrow f(X_i)$
    sum $\leftarrow$ sum + output
    sum_squares $\leftarrow$ sum_squares + output^2
end for

$E[f(X)] \leftarrow \text{sum}/N$
$\text{Var}[f(X)] \leftarrow (\text{sum\_squares}/N) - (E[f(X)])^2$
return $E[f(X)]$, $\text{Var}[f(X)]$

This pseudocode estimates the expected value and variance of a model's output using Monte Carlo simulations.

6.7.3. Pseudocode for Adversarial Risk Quantification

From [47], adversarial risk can be quantified using the following pseudocode:

Input: Data distribution $\mathcal{D}$, Model f, Loss function L, Perturbation bound ϵ
Output: Adversarial risk R_{adv}

$R_{\text{adv}} \leftarrow 0$
for each $x \sim \mathcal{D}$ **do**
    $\delta \leftarrow \text{argmax}_{\|\delta\| \leq \epsilon} L(f(x+\delta), y)$
    $R_{\text{adv}} \leftarrow R_{\text{adv}} + L(f(x+\delta), y)$
end for

$R_{\text{adv}} \leftarrow R_{\text{adv}}/|\mathcal{D}|$
return R_{adv}

This pseudocode calculates the adversarial risk R_{adv} by maximizing the loss over perturbations within a bound ϵ.

6.8. Section Conclusion

Quantitative methods are essential for managing the risks associated with generative AI models. By leveraging probabilistic frameworks, fairness metrics, Monte Carlo simulations, and adversarial risk quantification, financial institutions can ensure the reliability and robustness of their AI systems. These methods also support compliance with regulatory requirements, enabling the safe and responsible adoption of generative AI in finance.

 doi:10.20944/preprints202503.1579.v2

7. Foundational Metrics in Generative AI Model Risk Management

The application of quantitative methods is crucial for effectively managing model risk, particularly in the context of generative AI. This section outlines foundational metrics and quantitative approaches, grounded in the provided citations, that are essential for assessing and mitigating risks.

7.1. Performance Evaluation and Validation

Quantifying model performance is a cornerstone of MRM. Model validation, as highlighted by ValidMind [46], necessitates the use of metrics to assess the accuracy and reliability of generative AI outputs. In financial risk modeling, as explored by Yang et al. [12], quantitative measures are employed to evaluate the effectiveness of generative AI and LLMs. These evaluations often involve:

- **Accuracy and Precision**: Measuring the correctness of model outputs against known benchmarks.
- **Recall and F1-score**: Assessing the model's ability to identify relevant instances and balance precision and recall.
- **Statistical Measures of Drift**: Monitoring changes in model performance over time to detect model drift or decay, as mentioned by various sources focusing on model risk monitoring [20,62].

7.2. Risk Quantification and Measurement

Quantifying risks associated with generative AI models is essential for effective risk management. This involves:

- **Bias Measurement**: Employing statistical methods to detect and quantify biases in model outputs, as suggested by discussions on fairness in AI [11].
- **Sensitivity Analysis**: Assessing the impact of input variations on model outputs to understand potential vulnerabilities and risks.
- **Stress Testing**: Using simulated scenarios to evaluate model performance under extreme conditions, which is especially relevant in financial risk modeling [4].

7.3. Compliance and Regulatory Metrics

Regulatory compliance requires the application of quantitative methods to demonstrate adherence to standards and guidelines. This includes:

- **Audit Trails and Documentation**: Maintaining quantitative records of model development, validation, and monitoring processes, as emphasized in discussions on model risk governance [1].
- **Metrics for Regulatory Reporting**: Using predefined metrics to generate reports that demonstrate compliance with regulatory requirements, as required by financial institutions [3].
- **Quantitative Risk Assessments**: Providing numerical risk ratings and evaluations as mandated by OSFI-FCAC and other regulatory bodies [6].

These quantitative methods provide a robust foundation for assessing and mitigating risks associated with generative AI models, ensuring their responsible and effective use in various applications, particularly within the financial sector.

7.4. Statistical Foundations

AI model risk management leverages statistical concepts to quantify and mitigate risks. The references [3,6] highlight the importance of robust statistical validation.

7.5. Risk Metrics

While specific formulas are not directly available in the .bib file, the discussion around risk management in [4,42] implies the need for metrics such as:

- **Model Error Rate:** Quantifies the frequency of incorrect predictions.

 doi:10.20944/preprints202503.1579.v2

- **Bias Metrics:** Measures the presence and magnitude of bias in model outputs, as emphasized in [42].

7.6. AI-Specific Metrics

Given the focus on AI in the references, metrics relevant to AI model performance are crucial [65]:

- **AUC-ROC:** Measures the ability of a model to distinguish between classes.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced view of model accuracy.

7.7. Qualitative Overlay

As discussed in [2], a human overlay of these quantitative metrics with the risks cited is important for governance. .

8. Gaps Analysis and Proposed Solutions

The adoption of generative AI in financial institutions has revealed several gaps in existing model risk management (MRM) frameworks. These gaps stem from the unique challenges posed by generative AI, such as model opacity, bias, and adversarial vulnerabilities. This section identifies key gaps in the literature and proposes solutions based on the references provided.

8.1. Gaps in Current MRM Frameworks

8.1.1. Lack of Standardized Validation Methods

One of the most significant gaps is the lack of standardized validation methods for generative AI models. Traditional MRM frameworks are designed for deterministic models and struggle to address the probabilistic nature of generative AI [5]. This gap is particularly evident in the validation of large language models (LLMs), where interpretability and explainability are major challenges [12].

8.1.2. Inadequate Handling of Bias and Fairness

Generative AI models are prone to biases, which can lead to unfair outcomes in financial applications. Current MRM frameworks often lack robust mechanisms for quantifying and mitigating bias [11]. For example, demographic parity and equalized odds are not consistently applied in financial risk modeling [47].

8.1.3. Limited Focus on Adversarial Robustness

Adversarial attacks pose a significant threat to generative AI models, yet existing MRM frameworks do not adequately address this risk. The lack of standardized metrics for adversarial robustness is a critical gap [47]. Financial institutions need tools to quantify and mitigate adversarial risks, particularly in high-stakes applications such as credit scoring and fraud detection [13].

8.1.4. Regulatory and Compliance Challenges

The rapid evolution of generative AI has outpaced regulatory frameworks, creating a gap between innovation and compliance. While organizations like OSFI and FCAC have issued guidelines for AI risk management, these frameworks are not yet fully aligned with the unique risks posed by generative AI [6]. This misalignment creates uncertainty for financial institutions seeking to adopt generative AI responsibly [3].

8.2. Proposed Solutions

8.2.1. Development of Standardized Validation Frameworks

To address the lack of standardized validation methods, financial institutions should adopt advanced validation frameworks tailored to generative AI models. These frameworks should incorporate

 doi:10.20944/preprints202503.1579.v2

probabilistic validation techniques, such as Monte Carlo simulations, to assess model performance under uncertainty [4]. Additionally, tools like DataRobot and H2O.ai can automate the validation process, ensuring consistency and efficiency [18,20].

8.2.2. Integration of Fairness Metrics

To mitigate bias and ensure fairness, MRM frameworks should integrate fairness metrics such as demographic parity and equalized odds. These metrics should be applied consistently across all stages of the model lifecycle, from development to deployment [11]. Financial institutions should also leverage explainable AI (XAI) techniques to enhance model interpretability and transparency [60].

8.2.3. Enhancement of Adversarial Robustness

To address adversarial risks, financial institutions should adopt adversarial training techniques and robust optimization methods. These approaches can improve the resilience of generative AI models to adversarial attacks [47]. Additionally, standardized metrics for adversarial robustness, such as adversarial risk, should be incorporated into MRM frameworks [47].

8.2.4. Alignment with Regulatory Frameworks

To bridge the gap between innovation and compliance, financial institutions should collaborate with regulators to develop AI-specific risk management standards. These standards should align with existing frameworks, such as the NIST AI RMF and ISO/IEC 23894, while addressing the unique risks posed by generative AI [9]. Proactive engagement with regulatory bodies, such as OSFI and FCAC, can also facilitate the responsible adoption of generative AI [6].

8.3. Section Conclusion

The gaps in current MRM frameworks highlight the need for a paradigm shift in the management of generative AI risks. By developing standardized validation methods, integrating fairness metrics, enhancing adversarial robustness, and aligning with regulatory frameworks, financial institutions can address these gaps and ensure the safe and responsible adoption of generative AI. Future research should focus on the practical implementation of these solutions, particularly in high-stakes financial applications.

8.4. Quantitative Findings Table

This section summarizes quantitative findings from the literature related to model risk management and generative AI. As shown in Table 3, the literature provides a range of quantitative results, metrics, and methods for assessing and mitigating risks associated with generative AI models. These findings are critical for developing robust MRM frameworks that can address the unique challenges posed by generative AI in financial institutions.

Table 3. Quantitative Findings from the Literature

Reference	Quantitative Finding	Key Metric/Method
[9]	Risk of adverse events in general-purpose AI systems	$R = P(E) \times C(E)$
[5]	Validation of generative AI models	Probabilistic validation techniques
[4]	Monte Carlo simulations for risk assessment	$E[f(X)] \approx \frac{1}{N} \sum_{i=1}^{N} f(X_i)$
[60]	Model prediction error	$\epsilon = y - \hat{y}$
[11]	Fairness metrics for bias mitigation	Demographic parity, Equalized odds
[47]	Adversarial risk quantification	$R_{\text{adv}} = \mathbb{E}_{x \sim \mathcal{D}} \left[\max_{\|\delta\| \leq \epsilon} L(f(x+\delta), y) \right]$
[20]	Automated model validation	Key Risk Indicators (KRIs)
[9]	Regulatory compliance metrics	$S = \sum_{i=1}^{m} w_i \cdot KRI_i$

 doi:10.20944/preprints202503.1579.v2

The quantitative findings presented in Table 3 highlight the importance of probabilistic frameworks, fairness metrics, and adversarial risk quantification in managing generative AI risks. For example, [9] propose a probabilistic framework for quantifying the risk of adverse events, while [11] emphasize the use of fairness metrics such as demographic parity and equalized odds to mitigate bias. Additionally, [47] introduce adversarial risk quantification to assess the resilience of AI models to adversarial attacks. These findings collectively provide a foundation for developing quantitative methods that can enhance the reliability and robustness of generative AI models in financial applications.

8.5. Proposals from the Literature Table

This section summarizes key proposals from the literature related to model risk management and generative AI. As shown in Table 4, the literature provides a range of actionable proposals for addressing the challenges posed by generative AI in financial institutions. These proposals are derived from recent research and industry best practices, offering a roadmap for improving MRM frameworks in the era of generative AI.

Table 4. Proposals from the Literature

Reference	Proposal
[5]	Develop advanced MRM frameworks for generative AI models.
[11]	Integrate fairness metrics (e.g., demographic parity, equalized odds) into MRM frameworks.
[47]	Enhance adversarial robustness using adversarial training and robust optimization methods.
[9]	Align MRM practices with regulatory frameworks like NIST AI RMF and ISO/IEC 23894.
[20]	Automate model validation using tools like DataRobot and H2O.ai.
[60]	Use explainable AI (XAI) techniques to improve model interpretability.
[15]	Implement three-step diligence processes for managing AI model risk in financial institutions.
[43]	Conduct webinars and training sessions to educate stakeholders on generative AI risks.
[61]	Adopt a risk-based approach to global governance of generative AI.
[59]	Transition from traditional MRM to AI risk management frameworks.

9. Future Directions

Advances in explainability methods, robust synthetic data validation, and AI safety frameworks will be crucial for improving MRM in generative AI.

The future of AI model risk management in financial institutions will likely involve more sophisticated frameworks that can keep pace with rapidly evolving AI technologies [65].

The future of MRM in the generative AI era will likely involve the integration of advanced technologies, such as automated compliance tools and AI-driven risk assessment platforms [20,55]. Furthermore, the development of global governance frameworks for generative AI will play a critical role in ensuring its responsible adoption [61,67].

9.1. Opportunities and Best Practices

9.1.1. Enhanced Risk Assessment

Generative AI can be leveraged to improve risk assessment capabilities, particularly in simulating and predicting potential market scenarios [4].

9.1.2. Automated Model Validation

AI technologies can be employed to automate aspects of model validation, potentially improving efficiency and accuracy in model risk management processes [20].

10. Conclusion

Generative AI represents both a significant opportunity and a formidable challenge for financial institutions. While it has the potential to enhance risk management and operational efficiency, it also introduces new risks that must be carefully managed. This paper has provided a comprehensive review

 doi:10.20944/preprints202503.1579.v2

of the current state of MRM in the context of generative AI, highlighting the key risks, regulatory frameworks, and mitigation strategies. By leveraging quantitative methods such as probabilistic risk quantification, Monte Carlo simulations, and adversarial risk metrics, financial institutions can better assess and mitigate the risks associated with generative AI models. Foundational metrics, including fairness measures like demographic parity and equalized odds, are essential for ensuring ethical and unbiased AI deployment. The paper also presented pseudocode for key algorithms, such as risk quantification and adversarial risk calculation, to provide a practical understanding of these methods.

A detailed gap analysis revealed critical shortcomings in current MRM frameworks, such as the lack of standardized validation methods, inadequate handling of bias and fairness, and limited focus on adversarial robustness. To address these gaps, the paper proposed solutions, including the development of advanced validation frameworks, integration of fairness metrics, and alignment with regulatory standards such as the NIST AI RMF and ISO/IEC 23894. These proposals aim to guide financial institutions in adopting generative AI responsibly while addressing the unique risks it poses. Future research should focus on the practical implementation of these solutions, particularly in high-stakes financial applications.

As AI continues to transform the financial sector, robust model risk management practices are essential. Financial institutions must balance the opportunities presented by AI with the need for responsible and compliant implementation. The integration of generative AI into financial institutions presents both significant opportunities and challenges. Effective MRM is crucial for mitigating the risks associated with these models and ensuring responsible AI adoption. As the technology continues to evolve, ongoing research and collaboration between industry, academia, and regulators will be essential for developing robust frameworks and best practices. Addressing MRM challenges through improved quantitative methods, foundational metrics, validation, and regulatory compliance will be essential for future adoption. This paper serves as a foundational resource for advancing MRM in the era of generative AI, providing actionable insights for researchers, practitioners, and policymakers alike.

References

1. AI model governance: What it is and why it's important Collibra.
2. AI Use by Financial Institutions OSFI and FCAC Recommendations for Sound Risk Management McMillan LLP.
3. Crisanto, J.C.; Leuterio, C.B.; Prenio, J.; Yong, J. Regulating AI in the financial sector: recent developments and main challenges **2024**.
4. Ambilio. Generative AI for Risk Management in Financial Sector, 2023.
5. Mitigating Model Risk in AI Advancing an MRM Framework for AI/ML Models at Financial Institutions Chartis Research, 2025.
6. Institutions, O.o.t.S.o.F. OSFI-FCAC Risk Report - AI Uses and Risks at Federally Regulated Financial Institutions, 2024. Last Modified: 2024-10-16.
7. Group, m. Agenda - Model Risk, marcus evans Conferences.
8. Generative AI for Catastrophe Risk Xceedance, 2023. Section: Blog Posts.
9. AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models.
10. Uncompromising In Model Risk Management, 2022.
11. Fairly AI Managing AI Risk in Generative AI.
12. Yang, S.; Chen, J.; Gupta, A.; Feinstein, Z.; Knottenbelt, W. Generative AI and LLM in financial risk modeling and applications.
13. How Generative AI Will Disrupt Credit Risk Modeling.
14. XFIN-702 GenAI for Financial Model Risk Management Georgetown School of Continuing Studies (SCS).
15. AI Model Diligence: 3 Steps for Financial Institutions to Manage Model Risk.
16. AI and Model Risk Governance.
17. Artificial Intelligence and Model Risk Management.

 doi:10.20944/preprints202503.1579.v2

18. Wire, B. H2O.ai Becomes First to Bring Model Risk Management to Generative AI for Regulated Industries, 2025.
19. Model risk management is evolving to govern generative AI.
20. Automating Model Risk Compliance: Model Development DataRobot AI Cloud.
21. Satyadhar, J. Review of Gen AI Models for Financial Risk Management. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN : 2456-3307* **2025**, *11*, 709–723.
22. Joshi Satyadhar. Enhancing Structured Finance Risk Models (Leland-Toft and Box-Cox) Using GenAI (VAEs GANs). *IJSRA* **2025**, *14*, 1618–1630.
23. Joshi, Satyadhar. Leveraging prompt engineering to enhance financial market integrity and risk management. *World Journal of Advanced Research and Reviews WJARR* **2025**, *25*, 1775–1785.
24. Satyadhar, J. The synergy of generative AI and big data for financial risk: Review of recent developments. *IJFMR-International Journal For Multidisciplinary Research* **2025**, *7*.
25. Joshi Satyadhar. Using Gen AI Agents With GAE and VAE to Enhance Resilience of US Markets. *The International Journal of Computational Science, Information Technology and Control Engineering (IJCSITCE)* **2025**, *12*, 23–38.
26. Satyadhar, J. ADVANCING FINANCIAL RISK MODELING: VASICEK FRAMEWORK ENHANCED BY AGENTIC GENERATIVE AI. *International Research Journal of Modernization in Engineering Technology and Science* **2025**, *7*, 4413–4420.
27. Joshi Satyadhar. Implementing gen AI for increasing robustness of US financial and regulatory system. *International Journal of Innovative Research in Engineering and Management* **2024**, *11*, 175–179.
28. Satyadhar, J. Gen AI for Market Risk and Credit Risk [Ebook ISBN: 9798230094388]. *Draft2Digital Publications Ebook ISBN: 9798230094388* **2025**.
29. Joshi Satyadhar. Agentic Generative AI and the Future US Workforce: Advancing Innovation and National Competitiveness. *International Journal of Research and Review* **2025**, *12*, 102–113.
30. Joshi, .S. A Literature Review of Gen AI Agents in Financial Applications: Models and Implementations. *International Journal of Science and Research (IJSR)* **2025**, *12*, 1094–1100.
31. Satyadhar, J. The Transformative Role of Agentic GenAI in Shaping Workforce Development and Education in the US. *Iconic Research And Engineering Journals* **2025**, *8*, 199–206.
32. Joshi, S. A Comprehensive Review of Data Pipelines and Streaming for Generative AI Integration: Challenges, Solutions, and Future Directions.
33. Satyadhar, J. Retraining US Workforce in the Age of Agentic Gen AI: Role of Prompt Engineering and Up-Skilling Initiatives. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)* **2025**, *5*.
34. Joshi Satyadhar. Generative AI: Mitigating Workforce and Economic Disruptions While Strategizing Policy Responses for Governments and Companies. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) ISSN (Online) 2581-9429* **2025**, *5*, 480–486.
35. Satyadhar, J. Training US Workforce for Generative AI Models and Prompt Engineering: ChatGPT, Copilot, and Gemini. *International Journal of Science, Engineering and Technology ISSN (Online): 2348-4098* **2025**, *13*.
36. Joshi Satyadhar. Quantitative Foundations for Integrating Market, Credit, and Liquidity Risk with Generative AI. *https://www.preprints.org/* **2025**.
37. Satyadhar, J. Introduction to Vector Databases for Generative AI: Applications, Performance, Future Projections, and Cost Considerations. *International Advanced Research Journal in Science, Engineering and Technology ISSN (O) 2393-8021, ISSN (P) 2394-1588* **2025**, *12*, 79–93.
38. Joshi Satyadhar. Bridging the AI Skills Gap: Workforce Training for Financial Services. *International Journal of Innovative Science and Research Technology* **2025**, *10*, 1023–1030.
39. Satyadhar, J. Introduction to Generative AI and DevOps: Synergies, Challenges and Applications.
40. Workforce Development in the Finance Sector, E-book, Draft2Digital, 2025.
41. 5 Risks of Generative AI How to Mitigate Them in 2025.
42. Cruz, R. Managing the Risks of Generative AI: Achieving Compliance Across Use Cases, 2023.
43. Webinar: Model Risk Management for Financial Institutions in the Generative AI Era.
44. How Generative AI in Finance Strengthening Risk & Compliance.
45. (27) Navigating Model Risk Management in the Age of AI LinkedIn.
46. Validating GenAI Models: Three Tips for AI Risk Management, 2024. Section: Generative AI.

 doi:10.20944/preprints202503.1579.v2

47. Risks of Generative AI.
48. Peter, M. Generative AI and model risk management: new potential for the financial sector - KPMG in Germany, 2025.
49. Srivastava, A.K. Model Risk in the Generative AI World: Meritorious or Detrimental?, 2025.
50. The Impact of GenAI in Model Risk Management (MRM) - ValidMind, 2024. Section: Generative AI.
51. Peterson, B. Four Ways Banks Are Harnessing AI to Manage Model Risk.
52. Principal | authorurl:https://www.ey.com/en_us/people/gagan-agarwala, a.A.F.S.A.; Principal | authorurl:https://www.ey.com/en_us/people/alejandro-latorre, a.A.F.S.A.; Partner | authorurl:https://www.ey.com/en_us/people/susan-raffel, a.A.F.S.A. Model risk management for AI and machine learning.
53. Emerton Data — Model Risk in the Age of Artificial Intelligence and Machine Learning.
54. Top 5 Ways Risk Management Teams Are Using Generative AI.
55. Model Risk Management, a true accelerator to corporate AI, 2023.
56. Transitioning from model risk management to AI risk management.
57. The future of generative AI in banking McKinsey.
58. Turner, A. ERM Model Risk and AI, 2024. Section: Compliance – Sponsored Content.
59. Adapting model risk management in the gen AI era.
60. What is model risk management? Domino Data Lab.
61. Generative AI Global Governance and the risk-based approach.
62. Model Risk Management in an AI-Driven World [SS1/23].
63. dwillis. Harnessing AI to accelerate model risk management in FinTech, 2024.
64. Evans, H. Generative AI Risks and Regulatory Issues, 2024.
65. Markle, A. The Future of AI Model Risk Management in Financial Institutions, 2025.
66. Financial Services: 6 Ways to Support a Generative AI Risk Management Strategy.
67. Generative AI Masterclass Model Risk Management.
68. Generative AI for Risk Management.
69. https://www.modelop.com/blog/five-ways-mitigate-risk-ai-models.
70. (27) Generative AI Model Risk Management For Organizations LinkedIn.