technische universität
dortmund

Essays on the Application of Machine Learning in Finance

Dissertation zur Erlangung des akademischen Grades eines
Doctor rerum politicarum
der Fakultät für Wirtschaftswissenschaften
der Technischen Universität Dortmund

PATRICK SPITZER

JULY 13, 2026

Table of Contents

1 Introduction 1

1.1 Summaries of the Essays 3

1.2 Publication Details 6

2 Evaluating Data-Driven Risk Estimation: A Practical Assessment of Volatility Forecasting under Stress Conditions 10

2.1 Introduction 10

2.2 Literature Review 12

2.3 Methodology 13

2.4 Results 19

2.5 Discussion 23

2.6 Conclusion and Future Research 26

3 Bridging the AI Governance Gap: Systematic Weaknesses and Targeted Solutions in Financial Model Risk Management 28

3.1 Introduction 28

3.2 Evolving Regulatory Landscape and Literature Review 30

3.3 Methodology 33

3.4 Enhancing AI Model Risk Management: Weaknesses and Recommendations 38

3.5 Case Studies 55

3.5 Conclusion 61

4 Inherent Model Risk Assessment for AI-Based Models: Integrating AI Risk Drivers into Model Risk Management 63

4.1 Introduction 63

4.2 Literature Review 65

4.3 Methodology 66

4.4 AI-specific risk drivers of inherent model risk 68

4.5 Translation Framework for AI-Specific Drivers in Model Risk Tiering 70

4.6 Use Cases 77

4.7 Conclusion and Future Research 79

References 81

List of Figures

Figure 2.1: Daily actual volatility for AAPL, EUR/USD, and Gold. 14

Figure 2.2: Actual vs. forecasted volatility for AAPL 20

Figure 2.3: Visual representation of the VaR backtesting results for Gold Price. 22

Figure 3.1: High-level Process of multi-stage qualitative evidence synthesis for identifying AI governance weaknesses and recommendations. 33

Figure 3.2: PRISMA flow diagram depicting steps of literature search and selection for included studies. 36

Figure 4.1: PRISMA flow diagram for MRM tiering literature, including search query, eligibility criteria, exclusion reasons, and final retained sample. 67

Figure 4.2: PRISMA flow diagram for AI risk and responsible-AI literature, including search query, screening criteria, exclusion reasons, and final retained sample. 67

List of Tables

Table 2.1: Set of configurations for the hyperparameters of the LSTM and xLSTM networks. 16

Table 2.2: Set of configurations for the hyperparameters of the Autoformer. 17

Table 2.3: Out-of-sample forecasting performance of all models for AAPL, gold price, and the EUR/USD exchange rate across all three periods and assets. 19

Table 2.4: Out-of-sample DTW results of all models for AAPL, gold price, and the EUR/USD exchange rate across all three periods and assets. 20

Table 2.5: Out-of-sample TDI results of all models for AAPL, gold price, and the EUR/USD exchange rate across all periods and assets. 21

Table 2.6: Numbers of VaR exceedances at the 1% significance level of all models for AAPL, gold price, and the EUR/USD exchange rate across all periods and assets. 22

Table 3.1: Eligibility Criteria for Full-Text Review. 36

Table 4.1: Indicative sub-dimensions and classification criteria for inherent model risk assessment. 74

Table 4.2: Indicative Materiality–Complexity Matrix for Inherent Model Risk Tiering, with Tier 1 representing the highest governance priority. 76

Acknowledgements

Embarking on a PhD is a daunting undertaking, even for the boldest and bravest among us. Yet here I am, writing my way through the final stages of this life-changing journey. Pursuing the degree while working full-time has made the path all the more demanding. Completing this dissertation has therefore been not only an academic challenge, but also a period of profound personal growth.

First and foremost, I am grateful to my supervisor, Peter N. Posch. One of his thoughts that has stayed with me throughout this journey is his description of a PhD as a “rough journey” rather than a one-shot effort. His unwavering belief in my abilities and academic potential encouraged me to persevere, even during the most difficult moments. I also benefited greatly from his remarkable ability to engage with a wide range of topics and to provide consistently thoughtful and insightful feedback.

I would like to thank the finance research group (Kira, Alex, Cathrin, Selin, and Kevin) at TU Dortmund University. The years we spent together, whether at the university, during PhD days, or in the CrossFit box, were both professionally enriching and personally great fun.

My sincere thanks also go to my colleagues Henning, Joachim, and Rüdiger, who always had my back when things became particularly demanding, especially during revision rounds and tight deadlines. Your support, flexibility, and willingness to step in made a real difference and helped me keep moving forward.

Completing this dissertation required far more than academic and professional support. I am especially thankful to my friends, who listened, advised, encouraged, and, at times, patiently endured the ups and downs of this process—among them Jana, Franzi, Insa, Miguel, Sarah, and Emma. Listing everyone individually would risk turning these acknowledgements into an Oscar acceptance speech. Nevertheless, your support has meant more than I can express here.

Finally, and most importantly, I would like to thank my mum, Christoph, Sebastian, Petra, Heinz, and Thomas, for their unwavering support and encouragement. Without you, this dissertation would not have been possible. Despite the ever-growing mountain of papers that surrounded me over the years, I somehow made it through.

1 Introduction

Quantitative models have become essential to financial institutions. They support core activities such as risk assessment, capital planning, pricing, portfolio management, fraud detection, and regulatory reporting. As data availability and computational capacity have grown, financial institutions have complemented traditional econometric and statistical methods with machine learning (ML) and artificial intelligence (AI)-based approaches (Buchanan, 2019; Machado, 2026). These methods promise greater flexibility, stronger predictive performance, and the ability to capture nonlinear patterns in complex data environments (Kristjanpoller et al., 2014).

At the same time, the growing use of AI-based models raises important questions for financial risk management and model governance (Aldasoro et al., 2025; Baeza-Yates & Fayyad, 2024). In risk-sensitive applications, predictive accuracy alone is not sufficient. Models must also be reliable under stress, interpretable or at least adequately explainable, robust to changing market and data conditions, and capable of being governed within existing control frameworks. This is particularly important in financial institutions because model outputs may directly affect customers, supervisory reporting, and strategic decisions.

The adoption of AI therefore creates a dual challenge. First, financial institutions must determine whether advanced data-driven models actually improve risk estimation in practice, especially during periods of market stress when model errors are most consequential. Second, they need model risk management (MRM) frameworks that translate AI-specific characteristics, including opacity, adaptive behaviour, and data dependence, into operational requirements for classification, validation, monitoring, and oversight. The central question is consequently not only whether AI-based models can produce more accurate outputs, but also whether they can be used and governed in a manner that is consistent with financial stability, regulatory compliance, and responsible decision-making. Against this background, the dissertation addresses the following overarching research question:

How can financial institutions evaluate and govern AI-based models when gains in predictive performance may be accompanied by greater model risk and additional governance and regulatory challenges?

The dissertation answers this question through three related sub-questions:

1. How do AI-based models perform relative to established econometric approaches under normal and stressed market conditions?

2. Which weaknesses arise when traditional MRM frameworks are applied to AI-based models across the model lifecycle, and how should those frameworks be enhanced?
3. How can the inherent risk of AI-based models be assessed and incorporated into model-risk tiering?

This dissertation addresses these questions through three self-contained essays that link financial risk modelling with AI governance and MRM. Although each study can be read independently, together they examine how increasingly complex and data-driven models should be evaluated and governed in financial institutions.

The essays use complementary methodological approaches. Chapter 2 compares econometric and AI-based forecasting models and considers how their performance affects Value-at-Risk outcomes. Chapter 3 synthesises regulatory and academic evidence to identify weaknesses in the governance of AI-based models and proposes lifecycle-based improvements. Chapter 4 develops a framework for incorporating AI-specific risk drivers into inherent model risk assessment and tiering.

By linking empirical model evaluation, lifecycle governance, and inherent-risk classification, the dissertation brings together dimensions that are often examined separately. It shows that statistical performance alone cannot establish whether a model is suitable for institutional use. Suitability also depends on whether the model can be effectively validated, challenged, and prioritised within the governance framework. The dissertation therefore supports a proportionate approach in which governance requirements reflect the consequences and probability of failure. In doing so, it contributes to research on financial ML, AI governance, and MRM.

1.1 Summaries of the Essays

Chapter 2 investigates whether traditional econometric and ML models remain reliable when financial markets move from stable conditions into periods of severe stress. The analysis focuses on volatility forecasting, a core input for portfolio management, Value-at-Risk calculations, and stress testing. However, forecasting volatility becomes particularly difficult during crises, when markets are characterised by abrupt regime shifts, heavy-tailed shocks, and rapid changes in market dynamics. Models that perform adequately during normal periods may fail precisely when accurate risk estimation is most important.

The essay compares an established econometric benchmark with several ML approaches. Specifically, it evaluates GARCH (Bollerslev, 1986), LSTM-based models (Hochreiter & Schmidhuber, 1997), the extended LSTM architecture (Beck et al., 2024), hybrid GARCH-LSTM models (Kristjanpoller et al., 2014; Kumar et al., 2025), and transformer-based architectures (H. Wu et al., 2021). The empirical analysis uses daily financial market data from equity, commodity, and foreign exchange markets covering periods of significant turbulence, specifically the COVID-19 crisis. The essay assesses model performance across different market regimes using not only conventional point forecast accuracy but also time-dependent measures that capture the shape and timing of volatility forecasts (Le Guen & Thome, 2019; Sakoe & Chiba, 1978). In addition, the essay analyses the implications of model choice for Value-at-Risk backtesting and regulatory suitability.

The results show that ML models can improve certain dimensions of forecasting performance. Autoformer achieves strong point forecasting accuracy, while the extended LSTM architecture captures volatility trajectories and the timing of volatility changes particularly well. However, the findings also reveal a critical limitation for risk management applications. ML models may generate higher Value-at-Risk exceedance rates during stress periods, whereas GARCH produces comparatively conservative VaR estimates and fewer exceedances in several of the analysed asset-period settings. The essay thus identifies a fundamental trade-off between predictive accuracy and risk adequacy. While advanced models may improve forecast precision, traditional econometric models may remain useful as benchmarks or complementary models because of their transparency and stability.

The contribution is fourfold. First, it benchmarks traditional econometric, ML, and hybrid models under real-world stress scenarios. Second, it applies recent deep learning architectures to financial volatility forecasting. Third, it extends the evaluation of volatility forecasts beyond point error metrics by incorporating trajectory- and timing-sensitive performance measures.

Fourth, it explicitly links empirical model performance to regulatory and MRM considerations. The findings therefore show that model evaluation in financial institutions should not be based solely on statistical accuracy but must also account for conservatism, explainability, operational stability, and supervisory acceptability.

Chapter 3 turns from model performance to the governance structures required for AI-based models. Although financial institutions already operate established MRM frameworks (Board of Governors of the Federal Reserve System, 2011), many of these frameworks were designed for models with relatively stable inputs, assumptions, and behaviours. AI-based models challenge this design because they may be opaque, adaptive, non-deterministic, dependent on complex or unstructured data, embedded in broader systems, or supplied by third-party providers (Baeza-Yates & Fayyad, 2024).

The chapter examines how traditional MRM frameworks can be adapted to these characteristics. It identifies recurring weaknesses across the model lifecycle, including model identification and inventory, third-party management, validation, and post-deployment controls. AI models may blur the boundaries of what qualifies as a model, involve stakeholders beyond traditional risk functions, require validation that combines outcome-based and process-based evidence, and call for monitoring that accounts for drift, retraining, and changes in use.

In response, the essay proposes targeted enhancements that remain consistent with established MRM principles. These include clearer model definitions, stronger inventory practices, AI-specific tiering criteria, adapted validation methods, improved monitoring and change management, and dedicated governance arrangements for third-party AI systems. Its main contribution is a lifecycle-based taxonomy that interprets these weaknesses as misalignments between AI-specific characteristics and the assumptions embedded in established governance processes. The resulting recommendations translate high-level AI-governance principles into concrete governance measures.

Chapter 4 focuses on a central element of model governance: inherent model risk assessment. In many financial institutions, MRM relies on tiering frameworks (Bank of England, 2023b) that classify models according to their relative risk (Kiritz et al., 2019). These classifications determine the intensity of governance and validation. Inherent model risk assessment is therefore a key mechanism for allocating resources across model inventories.

AI-based models create difficulties for this tiering logic. Existing AI risk frameworks often classify risks at the level of AI systems, use cases, or broader societal impacts (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024). By contrast, MRM tiering requires model-level

and inventory-compatible assessments. This creates a mismatch between the way AI risks are commonly described and the way financial institutions need to classify models for governance purposes. AI-related risks such as opacity, bias, lack of explainability, data drift, and behavioural instability must therefore be translated into dimensions that are compatible with established inherent risk assessment.

The essay addresses this problem by developing a conceptual and design-oriented framework for integrating AI-specific and system-level risk drivers into model-level inherent risk assessment. It argues that AI does not require a separate definition of inherent model risk. The relevant risk remains the potential for adverse consequences arising from incorrect or misused model outputs before the effect of mitigating controls. What changes is how this risk materialises, how observable it is, and how reliably it can be assessed. AI-specific risk drivers may increase materiality when they amplify the severity, legal relevance, stakeholder impact, or propagation of adverse consequences. They may increase complexity when they raise the likelihood of failure, reduce assessability, or make validation and challenge more difficult.

The analysis treats the model-in-use as the relevant unit of assessment. This perspective is important because the risk of an AI model does not depend on model architecture alone. The same technical model may have different inherent risk depending on its deployment context, decision pathway, degree of automation, affected stakeholders, and connection to regulated processes. For example, a model used internally as an advisory tool may have a different risk profile from a similar model used directly in customer-facing or regulatory processes.

The third essay contributes a translation framework that incorporates these risk drivers into existing MRM tiering. By mapping them to materiality and complexity, the framework supports consistent prioritisation across statistical, ML, and generative AI models within a unified inventory. It thereby extends established model-risk practices to AI applications without creating a separate and potentially fragmented governance system.

1.2 Publication Details

Paper 1 (Chapter 2):

Evaluating Data-Driven Risk Estimation: A Practical Assessment of Volatility Forecasting under Stress Conditions

Authors:

Patrick Spitzer and Peter N. Posch

Abstract:

This study provides a comparative analysis of machine learning and traditional econometric volatility forecasting models' performance during market crises, with direct implications for Value-at-Risk (VaR) calculations and regulatory compliance. Using daily data from the equity, commodity, and foreign exchange market spanning the COVID-19 pandemic, we demonstrate that transformer-based architectures (Autoformer) achieve superior point forecasting accuracy, while the extended LSTM (xLSTM) better captures volatility trajectories and more precisely aligns with the timing of volatility changes. However, several machine-learning models exhibit critical limitations for risk management applications because they produce substantially more VaR exceedances than GARCH during stress periods. We identify a fundamental trade-off: while deep learning models improve forecast precision, GARCH-type models maintain regulatory acceptability through conservative risk estimates despite higher capital charges. The findings advocate for hybrid governance frameworks that combine machine learning's adaptive capabilities with traditional models' operational stability, providing actionable insights for financial institutions navigating regulatory requirements.

Publication Details:

Working Paper. Under review at the *International Review of Economics and Finance*

Paper 2 (Chapter 3):

Bridging the AI Governance Gap: Systematic Weaknesses and Targeted Solutions in Financial Model Risk Management

Authors:

Patrick Spitzer and Peter N. Posch

Abstract:

Financial institutions increasingly deploy artificial intelligence (AI) models for critical decisions, yet traditional Model Risk Management (MRM) frameworks struggle to address AI-specific risks, including opacity, adaptive behaviour, and fairness, conduct, and accountability concerns. Based on a structured review of regulatory guidance, academic literature, and industry commentary, this study identifies twelve recurrent weaknesses in current AI model governance across MRM lifecycle domains, including governance and organisation, scope and model inventory, third-party models, model development and validation, deployment and monitoring, and change management. Key gaps include inadequate stakeholder identification due to the expanded scope of model use beyond traditional risk functions, insufficient AI-specific validation methods requiring hybrid outcome- and process-based approaches, and monitoring frameworks that are insufficiently adapted to dynamic AI models. We propose targeted governance enhancements aligned with emerging regulatory frameworks and supervisory guidance, including the EU AI Act, the European Central Bank's Guide to internal models, and the United Kingdom Prudential Regulation Authority's Supervisory Statement SS1/23. These enhancements include clarifying model definitions, incorporating AI-specific tiering criteria, and establishing vendor governance protocols. Three case illustrations demonstrate context-sensitive application across market risk modelling, enterprise chatbot deployment, and strategic AI use-case discovery. These findings bridge the gap between high-level regulatory principles and lifecycle-level MRM implementation, offering financial institutions a structured basis for strengthening AI-related MRM while maintaining regulatory compliance and operational resilience.

Publication Details:

Working Paper. Revised after minor revision and re-submitted to the *Journal of Banking Regulation*

Paper 3 (Chapter 4):

Inherent Model Risk Assessment for AI-Based Models: Integrating AI Risk Drivers into Model Risk Management

Authors:

Patrick Spitzer and Peter N. Posch

Abstract:

Purpose

This paper examines how artificial intelligence (AI)-specific and system-level risk drivers can be incorporated into inherent model risk assessment and tiering in financial institutions. It addresses the mismatch between AI risk frameworks, which operate at system or use-case level, and model risk tiering, which requires model-level, inventory-compatible classifications.

Design/methodology/approach

The paper adopts a framework-building approach based on an integrative review and qualitative synthesis of model risk management (MRM), AI risk assessment literature, and regulatory guidance. It identifies AI-specific risk manifestations relevant to inherent model risk, maps them to materiality and complexity, and illustrates application through banking use cases.

Findings

The paper argues that AI does not require a separate definition of inherent model risk but changes how risk manifests and how reliably it can be assessed. AI-specific drivers affect materiality where they increase consequence severity, legal relevance, or stakeholder impact, and complexity where they increase the probability of model failure or reduce assessability.

Research limitations/implications

The framework is conceptual and illustrated through banking use cases rather than validated on model inventories. Future research should test calibration, inter-rater reliability, and institutional usability.

Practical implications

The framework helps institutions integrate AI-specific risk drivers into MRM tiering and allocate proportionate validation and governance resources without creating a parallel AI risk taxonomy.

Originality

The paper develops a mechanism translating AI-specific and system-level risk drivers into MRM tiering logic across traditional, machine learning, and generative AI models.

Publication Details:

Working Paper. Submitted to the *Journal of Financial Regulation and Compliance*

2 Evaluating Data-Driven Risk Estimation: A Practical Assessment of Volatility Forecasting under Stress Conditions

2.1 Introduction

Forecasting financial market volatility is a fundamental task in risk management and asset pricing. Volatility is widely recognised as a proxy for financial risk, and its sudden surges during market crises can severely impact financial stability. Therefore, accurate volatility forecasts are critical inputs for portfolio allocation, derivative pricing, Value-at-Risk (VaR) calculations, and stress testing frameworks.

However, forecasting volatility remains a challenging endeavour, particularly under stress conditions characterised by abrupt regime shifts, heavy-tailed shocks, and rapid changes in market dynamics. Traditional forecasting models that perform reasonably well during normal periods may fail to anticipate significant volatility changes during crises, exposing institutions to underestimated risks precisely when resilience is most critical. The 2008 global financial crisis and the 2020 COVID-19 pandemic revealed significant shortcomings in existing risk models, reigniting interest in developing more robust forecasting frameworks.

The emergence of machine learning (ML) and deep learning techniques offers new opportunities to enhance volatility forecasting. Unlike traditional parametric models, ML methods can flexibly capture nonlinear patterns and regime shifts, potentially improving predictive performance in complex environments. Yet, concerns about interpretability, stability, and operational deployment remain obstacles to their widespread adoption in risk-sensitive industries.

This paper presents a comprehensive empirical assessment of volatility forecasting models under stress conditions, with a particular focus on comparing an established econometric benchmark with ML approaches. Specifically, we evaluate the performance of a GARCH model, Long Short-Term Memory (LSTM) networks, including an extended LSTM architecture (xLSTM), and hybrid GARCH-LSTM models using daily squared logarithmic returns of financial data covering periods of significant turbulence, such as the COVID-19 crisis, to test robustness under structural breaks and heavy-tailed returns. To that end, we assess model

accuracy not only using forecast point error metrics (mean squared error (MSE)) and regulatory backtesting of VaR but also time-dependent performance measures that capture the shape and temporal alignment between predicted and actual volatility trajectories. Furthermore, we discuss practical considerations such as model adaptability, interpretability, and implementation challenges. These questions are not only methodological but also carry international relevance, as volatility models underpin capital regulation and systemic risk management. Given the global nature of financial institutions and cross-market spillovers, assessing the regulatory suitability of ML models is crucial for both international banking supervision and financial stability.

Our contributions are fourfold. First, we benchmark traditional econometric and ML models side-by-side under real-world stress scenarios, providing insights into their relative strengths and weaknesses across stable, crisis, and recovery regimes. Prior work has compared GARCH-type and neural models or proposed hybrids, and some studies situate results within regulatory VaR backtesting (e.g., Pérez-Hernández et al., 2024). We complement this literature by explicitly structuring the evaluation across distinct volatility regimes, thereby stressing model stability under real-world conditions.

Second, we provide the first application of the xLSTM architecture to volatility forecasting. To our knowledge, xLSTM has not been applied to risk estimation in finance. Existing studies discuss ML models, such as LSTM, and traditional econometric models, but not the latest advances in extended memory architectures. Our study addresses this gap.

Third, we go beyond conventional evaluation metrics by incorporating trajectory- and timing-sensitive measures (Dynamic Time Warping (DTW) and Time Distortion Index (TDI)) alongside MSE and VaR backtesting. While previous studies have typically relied on point accuracy and exceedance tests, we demonstrate that DTW and TDI add complementary insights into the timing and shape of volatility responses, particularly under crisis conditions.

Fourth, we explicitly link these empirical findings to regulatory and model risk management perspectives. Unlike most forecasting papers that focus on predictive gains alone, we emphasise the trade-off between accuracy and conservatism, showing why GARCH remains regulation-friendly despite lower accuracy. This regulatory framing responds to the growing Trustworthy Artificial Intelligence agenda (Giudici, 2024), thereby embedding the results in the broader discussion of governance-ready machine learning.

Taken together, these contributions extend the literature by introducing a new architecture, broadening the evaluation framework, and embedding results within the debate on regulatory compliance and model governance.

The remainder of this paper is organised as follows. Section 2 reviews the relevant literature on volatility forecasting models. Section 3 describes the methodology and experimental design. Section 4 presents empirical results. Section 5 discusses practical implications, and Section 6 concludes.

2.2 Literature Review

Volatility forecasting has been extensively studied, with models evolving from traditional econometric approaches to contemporary ML-based methods. Foundational contributions include the ARCH model (Engle, 1982) and its generalisation, GARCH (Bollerslev, 1986), which capture volatility clustering and persistence and remain central to both academia and practice. Extensions such as EGARCH (Nelson, 1991) and GJR-GARCH (Glosten et al., 1993) account for asymmetric responses to shocks. Surveys (Andersen et al., 2003; Poon & Granger, 2003) emphasise the broad applicability of GARCH-type models but also note their limits in handling abrupt structural breaks and extreme events.

The development of ML methods introduced flexible, data-driven alternatives (Díaz et al., 2024; Gunnarsson et al., 2024; Petrozziello et al., 2022; Qiu et al., 2024). Neural networks, particularly LSTM (Hochreiter & Schmidhuber, 1997), show promise in capturing nonlinear patterns and long-term dependencies (Kristjanpoller et al., 2014; Yijun Wang et al., 2023). Empirical studies (A. Yang et al., 2024; Zhao et al., 2024) demonstrate that LSTM-based models often outperform GARCH in periods of stress, though concerns persist about overfitting, data intensity, and opacity relative to econometric approaches.

Recent advances have extended LSTM design. The xLSTM architecture (Beck et al., 2024) incorporates exponential gating and enhanced memory cells to improve long-sequence learning and training stability. While successful in fields such as natural language processing, its application to financial volatility forecasting remains largely unexplored.

Meanwhile, Transformer architectures (Vaswani et al., 2017) provide an alternative by modelling long-range dependencies via self-attention. Although vanilla Transformers may underperform with financial data (Zeng et al., 2023), specialised models, such as Autoformer (H. Wu et al., 2021), Informer (H. Zhou et al., 2021), Crossformer (Z. Zhang et al., 2024), and

iTransformer (Y. Liu et al., 2023) address challenges of periodicity, sparsity, and cross-variable dependence. Autoformer, in particular, has shown strong results in time-series competitions and is well-suited to financial forecasting.

Despite these advances, evidence on ML superiority remains mixed. The M4 competition (Makridakis et al., 2020) found pure ML approaches often underperformed statistical benchmarks, while Elsayed et al. (2021) show simpler algorithms like Gradient Boosted Trees can outperform deep networks when data are noisy or limited. These findings suggest complexity does not guarantee performance gains and that effectiveness depends on data characteristics.

In response, hybrid models that combine econometric structure with ML flexibility have gained traction (Kristjanpoller et al., 2014; Kumar et al., 2025; Yuejing Wang et al., 2024). By integrating GARCH outputs into neural architectures, these approaches leverage interpretability and domain grounding alongside adaptability, often achieving superior performance in volatile conditions (Kumar et al., 2025; A. Yang et al., 2024).

Overall, the literature shows that while traditional models offer theoretical grounding, and ML models bring flexibility and adaptability, both face challenges under extreme conditions. The emerging consensus highlights hybrids as promising. Yet significant gaps remain: little is known about how newer architectures such as extended memory networks or Transformer-based models perform under stress, or how model classes compare when evaluated not only by predictive accuracy but also by robustness and regulatory suitability.

Against this backdrop, our study evaluates traditional, ML, and hybrid volatility forecasting models under real-world stress conditions, introducing new architectures and metrics while explicitly embedding the analysis in a regulatory and governance context.

2.3 Methodology

2.3.1 Data and Market Regimes

We base our analysis on daily financial time series data spanning over a decade, allowing us to capture both routine market behaviour and extreme events. We employ daily price data for three representative assets: AAPL (Apple Inc. stock), EUR/USD exchange rate, and gold price. These assets represent a mix of equity, foreign exchange, and commodity markets, respectively, helping to generalise our findings. The sample period spans from 12 May 2012 to 16 Dec 2022. From these prices, we compute daily log returns, which serve as the basis for volatility

estimation and VaR calculation. For the out-of-sample accuracy evaluation, we use squared log returns as a proxy for volatility. Throughout the analysis, the term “actual volatility” refers to these squared log returns. Data were obtained from Datastream.

To determine the impact of the performance of volatility forecasting models under different market conditions, we segment the data into three overlapping sub-periods. The first period ranges from 12.05.2012 to 17.12.2019, representing a relatively stable market environment before the pandemic as no sudden regime shifts are observed, making it a suitable baseline for non-stressed conditions. The second period covers 12.05.2012 to 17.12.2020, thereby including the onset of COVID-19 pandemic outbreak and the resulting sharp volatility spikes. It is suitable to test how models perform when they have experienced a major shock in the training data. The third period ranges from 12.05.2012 to 16.12.2022, capturing a period in which markets began to stabilise as pandemic-related uncertainties gradually receded. It represents mixed conditions: an initial stress followed by recovery. Models trained on this full sample see both the crisis and the recovery phase. For each period, we reserve the last 250 trading days as the backtesting (out-of-sample) window.

Figure 2.1 illustrates the actual volatility across the three periods for all three assets. The figure plots the daily actual volatility, approximated by squared daily returns, during each backtesting window (Pre, During, and Post COVID-19). As shown in Figure 2.1, all assets exhibit occasional volatility outliers, but the COVID-19 crisis period stands out with a significantly higher average volatility level compared to the pre- and post-crisis periods. This confirms that our period segmentation successfully captures distinct volatility regimes: a low-volatility regime, a high-volatility stress regime, and a moderate volatility regime following the stress.

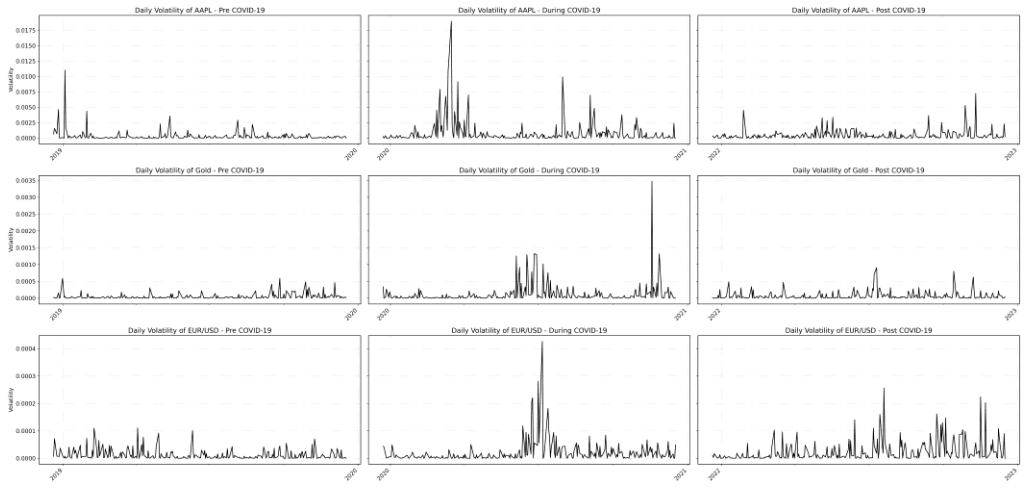

Figure 2.1: Daily actual volatility for AAPL, EUR/USD, and Gold.
This figure shows the daily actual volatility (squared log returns) for AAPL, EUR/USD, and Gold during each backtesting period (Pre, During, Post COVID-19). The During COVID-19 period shows markedly higher volatility spikes than the other periods.

2.3.2 Forecasting Models

We evaluate a diverse set of models for volatility forecasting, which can be categorised into traditional econometric models, ML models, and hybrid approaches that combine the two. All models produce one-step-ahead volatility forecasts and VaR for each of the 250 out-of-sample days using a rolling estimation window. The forecasts are compared with actual outcomes, consistent with the Basel Committee on Banking Supervision (2026) recommendation for VaR backtesting. The ML models are trained separately for each asset and each period up to the start of the backtesting window. The networks are reinitialised and retrained for each scenario. Below we describe each model and any relevant implementation details.

2.3.2.1 GARCH–Traditional Econometric Model

As a benchmark econometric model, we use the GARCH model due to its widespread regulatory and practical adoption for volatility and risk estimation. While more advanced econometric models, such as EGARCH and GJR-GARCH, have been proposed to address asymmetries and tail behaviour in financial volatility, our primary objective is not to optimise within the GARCH family. Instead, we focus on evaluating whether modern ML architecture offers meaningful performance improvements relative to the most widely used and regulatorily accepted baseline. This choice reflects common practice in financial institutions, where GARCH models remain central to VaR calculations and regulatory frameworks. It provides a conceptually transparent, parsimonious benchmark that captures volatility clustering and serves as a grounded reference point for comparison under stress conditions.

In this study, we assume a normal distribution for returns and the parameters are estimated via maximum likelihood for each asset and each training period. The GARCH lag orders for symmetric innovations (p) and lagged volatility (q) were systematically varied within the set {1,2,3}. The mean was modelled using constant mean, specifically we assumed a mean of zero. For the hybrid models, we default to GARCH(1,1) to manage training time and maintain consistency.

2.3.2.2 LSTM–Recurrent Neural Network

We implement an LSTM neural network as well as the extended xLSTM, given that the LSTM architecture has demonstrated strong empirical performance in volatility forecasting, particularly due to its ability to model long-range temporal dependencies and adapt to the non-linear dynamics commonly observed in financial time series (Kristjanpoller et al., 2014; Yuejing

Wang et al., 2024). Both networks take as input a sequence of past squared log returns of fixed length and output the next value. We train both networks by minimising the MSE between their volatility proxy predictions and the actual realised values in the training data. The xLSTM retains the overall recurrence and gating structure of a standard LSTM but incorporates two key improvements: (1) an exponential gating mechanism that provides a more powerful way to modulate information flow, and (2) enhanced memory via scalar and matrix memory cells that allow part of the state to be updated in a linear fashion and part to be updated in parallel (Beck et al., 2024). In essence, the xLSTM cell introduces additional gates/parameters that address known LSTM limitations such as gradient dilution over long sequences and lack of parallelism. In practice, our xLSTM implementation builds on a recent architecture by Alharthi and Mahmood (2024). To find the best-performing model, a comprehensive hyperparameter tuning process was conducted by implementing a grid search strategy. Given computational constraints and the significant training time required due to the complexity of the implemented ML models, we limited the hyperparameter search to a reasonable degree. Therefore, we used only a coarse grid search for the number of layers, the number of neurons, and the sequence length. Other hyperparameters were default values, such as the loss function, epochs, batch sizes, activation functions, and learning rate.

Furthermore, all models were compiled using the Adam optimiser. Table 2.1 summarises the configurations used for hyperparameter tuning of the LSTM and xLSTM networks.

Table 2.1: Set of configurations for the hyperparameters of the LSTM and xLSTM networks**.**

Hyperparameter	**Configuration**
Number of Neurons	[32, 64]
Number of Layers	[2,4]
Learning rate for the Adam optimiser	0.01
Loss function	MSE
Sequence Length	[5,10]
Epochs	20
Batch size	16

2.3.2.3 Autoformer–Transformer-Based Model

We use the Autoformer model (H. Wu et al., 2021) as a representative Transformer-based approach for time-series forecasting as it is specifically designed for long-horizon forecasting by introducing an Auto-Correlation mechanism in place of standard self-attention, aiming to capture seasonality and repeating patterns more effectively. It also decomposes the input series into trend and seasonal components and forecasts them with specialised decoder structures. We chose Autoformer over a vanilla Transformer because prior research indicated better

performance on financial data by explicitly addressing the characteristics of time-series data (Y. Wu, 2023; J. Zhang & Chen, 2022). The model is trained to minimise MSE on the training set. Similar to the LSTM and xLSTM network, we performed a comprehensive grid search strategy and limited the number of hyperparameters. Key hyperparameters are summarised in Table 2.2 and include the number of encoder/decoder layers and heads in the attention mechanism, which we set based on recommendations from the original study.

Table 2.2: Set of configurations for the hyperparameters of the Autoformer.

Hyperparameter	**Configuration**
Number of Heads	[2,4]
Number of Layers in Encoder	[2,4]
Number of Layers in Decoder	[2,4]
Learning rate for the Adam optimiser	0.01
Loss function	MSE
Sequence Length	[5,10]
Epochs	20
Batch size	16

2.3.2.4 Hybrid GARCH–LSTM Models

Following recent literature that advocates model hybridisation (Kim & Won, 2018; W. K. Liu & So, 2020), we include two hybrid approaches: GARCH-LSTM and GARCH-xLSTM. Practically, we first run a GARCH(1,1) on the training data to get an estimated volatility series. The estimated volatility is then fed into the LSTM as an additional input. For xLSTM, the setup is analogous to Hybrid-LSTM.

2.3.3 Value-at-Risk Calculation

We use a parametric VaR approach: for each day in the backtest, the model provides a volatility forecast, which is then used to compute the 1-day 99% VaR. This assumes returns are normally distributed with mean zero for VaR purposes. While this assumption may not hold exactly, it provides a consistent way to compare how different volatility forecasts translate into VaR.

2.3.4 Evaluation Metrics

We evaluate model performance using a comprehensive set of metrics, inspired by Jhin et al. (2024), to capture various aspects of forecast quality. By examining the following metrics, we aim to build a holistic view: MSE tells us about point-wise accuracy, DTW about shape accuracy, TDI about timing, and exceedances about risk management and regulatory implications.

2.3.4.1 Mean Squared Error

To compare the predictive power of each model in forecasting volatility, we mainly rely on the MSE to assess whether there are significant differences in their prediction accuracies. The MSE is the most frequently used criterion in forecasting studies and is robust even when a noisy volatility proxy is employed—it yields the same ranking of competing forecasts using a statistically unbiased volatility proxy (Hansen & Lunde, 2006; Patton, 2011). Because MSE is sensitive to the magnitude of errors, it effectively captures the accuracy of a model's forecasts.

2.3.4.2 Dynamic Time Warping

DTW measures the similarity between two time series that is flexible to timing differences (Sakoe & Chiba, 1978; Vintsyuk, 1972). It determines the optimal alignment between predicted and actual volatility series by allowing flexible time shifts of the time axis. We use DTW to measure how well the shape of the volatility forecast matches the shape of the actual volatility. We compute DTW distance for each forecast vs actual pair in the backtest. DTW is useful as it is less sensitive to small timing offsets. It focuses on the overall trajectory matching. For example, if a model predicts a volatility spike one day late, MSE will treat that harshly, however, DTW still credits that the model did catch the spike shape. Then, the DTW value will be low, signalling good alignment of the two timeseries.

2.3.4.3 Time Distortion Index

As highlighted in Le Guen and Thome (2019), it is essential that the model accurately captures both the underlying shape and timing of the time series. Therefore, we add the TDI to the evaluation framework as it quantifies the degree of temporal distortion needed to align two series, i.e., it measures the average lag or lead of the predictions relative to actual events. A small TDI means the predictions are roughly synchronous with actual changes, whereas a large TDI indicates consistent delays. Therefore, this metric directly quantifies the model's responsiveness to temporal changes.

2.3.4.4 VaR Exceedances

To evaluate the regulatory utility of the models, we conduct VaR backtests on the 99% one-day VaR forecasts from each model. We count the number of days in the 250-day backtest window where the actual loss exceeds the forecasted VaR, which is aligned with standard industry practices and regulatory expectations. According to Basel's traffic light test for 99% VaR, a

lower number of exceedances is better in terms of risk adequacy. However, overly high VaR could be inefficient. So, we interpret exceedances with the understanding that 2.5 exceedances per 250 days is considered ideal under regulatory guidelines.

2.4 Results

We present the out-of-sample performance of each model (GARCH, LSTM, Hybrid-LSTM, xLSTM, Hybrid-xLSTM, Autoformer) for each asset separately across three distinct periods. The results for the out-of-sample volatility forecasts based on the MSE are summarised in Table 2.3. To capture the alignment of volatility dynamics over time, Table 2.4 demonstrates DTW results, while Table 2.5 presents TDI scores. In addition to assessing forecasting accuracy, trajectory, and timing alignment, we evaluate the regulatory suitability of each model through VaR backtesting, with results shown in Table 2.6. The period-specific structure enables a clear assessment of each model's robustness under varying market conditions. While model performance appears broadly consistent across volatility regimes, ML approaches tend to show relative advantages during periods of market stress.

The MSE results in Table 2.3 show that Autoformer maintains the lowest overall forecasting error for most assets across the three subperiods, demonstrating its capacity to capture short-term fluctuations while preserving a broader perspective on the time series. The one notable exception is AAPL during COVID-19, where GARCH slightly outperforms Autoformer. Despite enhancements such as exponential gating, none of the recurrent networks fully match Autoformer's consistent MSE performance, underscoring their relative difficulty in handling sudden, drastic regime shifts.

Table 2.3: Out-of-sample forecasting performance of all models for AAPL, gold price, and the EUR/USD exchange rate across all three periods and assets.

	Pre COVID-19			During COVID-19			Post COVID-19		
Asset	AAPL	Gold Price	EUR/USD	AAPL	Gold Price	EUR/USD	AAPL	Gold Price	EUR/USD
GARCH	7.85E-07	8.10E-09	3.77E-10	**3.75E-06**	9.42E-08	1.69E-09	7.22E-07	1.58E-08	1.46E-09
LSTM	7.97E-07	8.59E-09	5.21E-10	4.31E-06	9.82E-08	1.71E-09	7.30E-07	1.56E-08	1.50E-09
Hybrid-LSTM	8.07E-07	8.26E-09	4.44E-10	4.19E-06	9.85E-08	1.78E-09	7.44E-07	1.67E-08	1.50E-09
xLSTM	1.47E-06	1.64E-08	8.20E-10	5.64E-06	1.82E-07	2.86E-09	1.31E-06	2.66E-08	2.79E-09
Hybrid-xLSTM	1.10E-06	1.22E-08	6.78E-10	4.26E-06	1.34E-07	2.12E-09	1.09E-06	2.33E-08	2.09E-09
Autoformer	**6.89E-07**	**7.89E-09**	**3.51E-10**	4.02E-06	**9.20E-08**	**1.54E-09**	**6.95E-07**	**1.43E-08**	**1.42E-09**

Notes: The table shows the MSE for one-day-ahead volatility forecasts for each period and asset. The best results are marked in bold for each period and asset. For the MSE, lowest values perform best.

While MSE evaluates point forecast accuracy, DTW and TDI allow us to examine the similarity in the shape and timing of volatility paths respectively, especially relevant in stress periods

where forecast lag matters (Table 2.4). With respect to DTW, xLSTM clearly excels by aligning closely with the actual volatility trajectory (Figure 2.2). Here, xLSTM outperformed all other models, achieving the lowest DTW scores. This suggests that xLSTM more accurately matches the dynamic ups and downs of volatility, even when level prediction is not perfect. In contrast, the Autoformer, despite its excellent MSE performance, exhibits larger DTW distances, indicating misalignment with the temporal structure of volatility shifts. The GARCH model also performs relatively poorly in DTW terms, consistently smoothing over rapid changes in volatility, resulting in the underrepresentation of abrupt regime shifts, highlighting its limitations in replicating the trajectory of volatility movements.

Table 2.4: Out-of-sample DTW results of all models for AAPL, gold price, and the EUR/USD exchange rate across all three periods and assets.

	Pre COVID-19			During COVID-19			Post COVID-19		
Asset	AAPL	Gold Price	EUR/USD	AAPL	Gold Price	EUR/USD	AAPL	Gold Price	EUR/USD
GARCH	7.0887	1.2569	0.3247	16.3465	2.9668	0.4352	10.3575	1.8636	0.5687
LSTM	7.7743	1.4075	0.4642	19.0090	2.7463	0.5539	9.8648	1.8546	0.5304
Hybrid-LSTM	6.6924	1.0377	0.2548	18.4682	3.0176	0.5975	9.5802	1.5647	0.5118
xLSTM	**1.4100**	**0.5819**	**0.1809**	**1.3049**	**0.1107**	**0.0229**	**0.455**	**0.1726**	0.0298
Hybrid-xLSTM	6.0577	0.7602	0.2157	8.0068	1.7032	0.4608	1.9907	1.9505	**0.2504**
Autoformer	5.5245	0.9293	0.2266	14.1575	2.2579	0.3524	7.7892	1.3479	0.4130

Notes: The best performing model is marked in bold for each period and asset. For DTW, lowest values perform best as lower scores indicate a greater similarity between the predicted and actual shape.

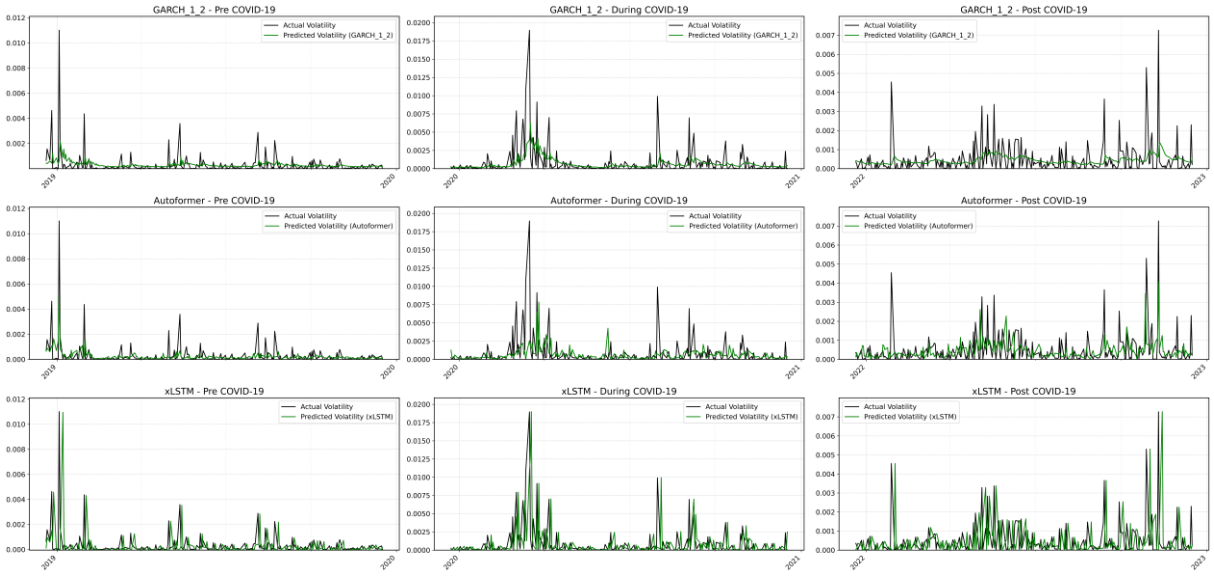

Figure 2.2: Actual vs. forecasted volatility for AAPL
This figure compares the actual vs. forecasted volatility from GARCH, Autoformer, and xLSTM for AAPL across all periods, highlighting differences in forecast shape and alignment with volatility timing.

Deep learning models exhibit notably lower TDI values relative to GARCH, indicating faster responsiveness to volatility spikes (Table 2.5). Among them, xLSTM consistently achieves the best temporal alignment, capturing both the timing and shape of volatility movements. While

GARCH and Autoformer respond more gradually to volatility peaks, the xLSTM forecast tracks the shape of the actual volatility well but with a delay (Figure 2.2). Therefore, Autoformer and particularly GARCH show higher TDI values on average, suggesting a slower reaction to regime changes.

Table 2.5: Out-of-sample TDI results of all models for AAPL, gold price, and the EUR/USD exchange rate across all periods and assets.

	Pre COVID-19			During COVID-19			Post COVID-19		
Asset	AAPL	Gold Price	EUR/USD	AAPL	Gold Price	EUR/USD	AAPL	Gold Price	EUR/USD
GARCH	0.0199	0.0930	0.2504	0.0176	0.0930	0.0872	0.0255	0.0277	0.1504
LSTM	**0.0086**	0.0087	**0.0081**	0.0301	0.0081	0.0449	0.0198	0.0085	0.0299
Hybrid-LSTM	0.0288	0.1042	0.0511	0.0303	0.3019	0.0796	0.0101	0.0325	0.0424
xLSTM	0.0087	**0.0085**	0.0105	**0.0084**	**0.0080**	**0.0080**	**0.0080**	**0.0079**	**0.0080**
Hybrid-xLSTM	0.0125	0.0120	0.0150	0.0097	0.0389	0.1571	**0.0080**	0.0320	0.0108
Autoformer	0.0182	0.0564	0.0327	0.0603	0.0283	0.0277	0.0233	0.0351	0.0500

Notes: The best performing model is marked in bold for each period and asset. For TDI, lowest values perform best as lower scores indicate a greater similarity between the predicted and actual trajectory in terms of timing.

The results of the VaR exceedance counts (Table 2.6) underscore that high average forecast accuracy does not necessarily reduce tail-risk violations. GARCH and simpler recurrent models such as LSTM can still record fewer exceedances in certain environments, especially Pre and Post COVID-19, where clustering or moderate volatility patterns dominate. During COVID-19, abrupt spikes produce more varied outcomes: While Autoformer excels on AAPL, GARCH does better on Gold. These differences highlight that model suitability depends on both the underlying asset and the prevailing market regime, and no single approach universally outperforms the rest for extreme events. Figure 2.3 illustrates that the GARCH model produces systematically a higher volatility level leading to an overall higher VaR. The LSTM model shows a similar behaviour to the GARCH. In contrast, the Autoformer adapts to the actual volatility level leading to a lower VaR amount in calm periods and a high VaR in turbulent periods. However, this behaviour accounts for a higher number of exceedances compared to the GARCH and LSTM.

Table 2.6: Numbers of VaR exceedances at the 1% significance level of all models for AAPL, gold price, and the EUR/USD exchange rate across all periods and assets.

	Pre COVID-19			During COVID-19			Post COVID-19		
Asset	AAPL	Gold Price	EUR/USD	AAPL	Gold Price	EUR/USD	AAPL	Gold Price	EUR/USD
GARCH	**4**	4	**3**	12	**5**	3	**5**	**5**	**7**
LSTM	**4**	**2**	0	14	8	3	**5**	**5**	**7**
Hybrid-LSTM	7	10	24	13	8	3	9	15	10
xLSTM	57	75	81	17	41	22	33	47	35
Hybrid-xLSTM	6	31	12	50	10	**0**	22	31	44
Autoformer	7	6	5	**10**	13	2	11	8	12

Notes: This table presents the backtesting results, at the 1% significance level, across all three periods and assets. The lowest number of exceedances in each column is shown in bold for each period and asset.

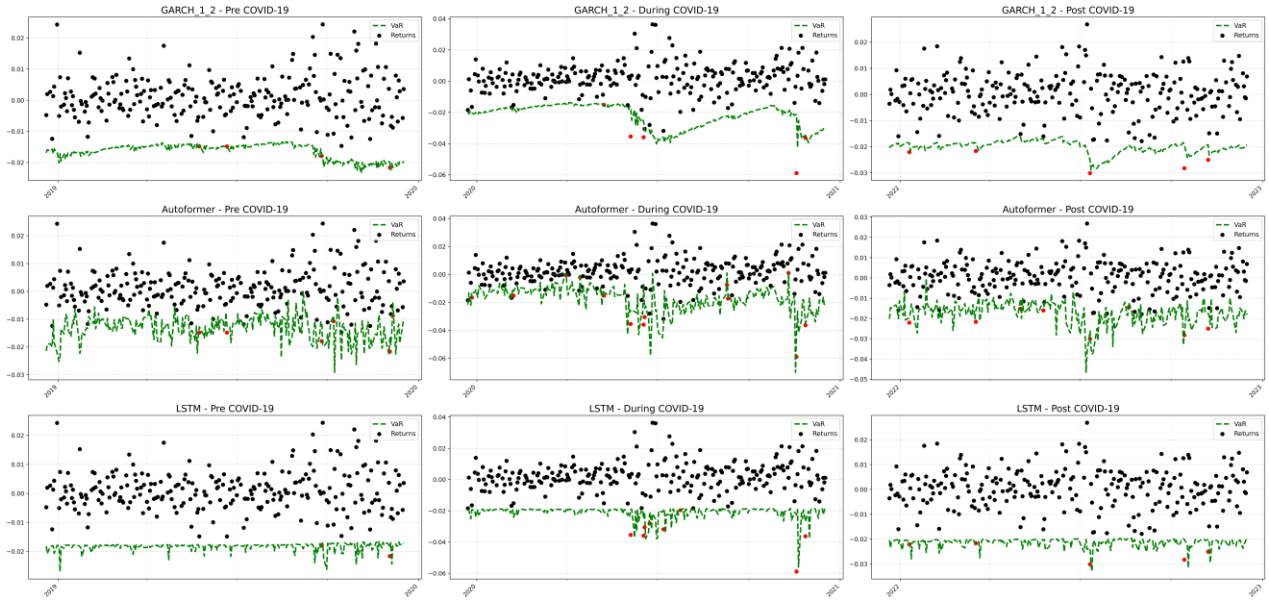

Figure 2.3: Visual representation of the VaR backtesting results for Gold Price.
This figure shows the VaR backtesting results for Gold Price based on volatility forecasts from GARCH, Autoformer, and LSTM. Each panel displays the backtesting outcomes over time for the Pre-COVID, During-COVID, and Post-COVID periods. The estimated VaR is shown in green, log returns in black, and exceedances are highlighted in red.

To conclude, these observations underline the importance of considering a combination of use-case-specific performance metrics, such as the average error (MSE), shape similarity (DTW), and timing alignment (TDI), for volatility forecasting, alongside VaR backtesting. Autoformer consistently delivers the lowest MSE, highlighting strong predictive accuracy, but exhibits moderate TDI scores, especially during turbulent periods, indicating a tendency to react more slowly to abrupt volatility shifts. It also shows limitations in shape alignment, as reflected in comparatively higher DTW values. In contrast, xLSTM excels in both DTW and TDI, capturing not only the shape but also the timing of volatility dynamics more accurately than other models.

GARCH remains the most robust model for VaR estimation, particularly under stress conditions, where its conservatively high volatility forecasts result in fewer exceedances.

Although xLSTM excels in both DTW and TDI, its backtesting results for VaR show significantly more exceedances than those from GARCH or Autoformer. This discrepancy reflects a core trade-off: xLSTM effectively captures the structure and timing of volatility dynamics, but its abrupt adaptation to lower volatility levels can lead to exceedances triggered by marginal fluctuations. In contrast, GARCH and Autoformer adjust more gradually and maintain an overall higher volatility level, providing a buffer that helps reduce breaches, especially during turbulent periods. Finally, hybrid models that combine GARCH with LSTM architectures do not offer meaningful improvements over their standalone counterparts, suggesting limited synergy in their integration for either forecasting precision or regulatory adequacy.

In summary, Autoformer emerges as the best all-around model for volatility forecasting due to its strong balance of accuracy and adaptability although it lacks accuracy in shape and timing, whereas GARCH remains a consideration for risk-focused applications.

2.5 Discussion

The results emphasise a fundamental trade-off between predictive accuracy and conservatism in risk estimation, particularly under stress conditions. Our analysis across different market regimes reveals that data-driven architectures such as the Autoformer achieve lower point forecast errors but exhibit vulnerabilities in capturing abrupt volatility regime shifts, whereas the xLSTM shows stronger responsiveness to such shifts despite less accurate point forecasts. Conversely, traditional models like GARCH, despite their lower forecasting precision and limited responsiveness to regime shifts, maintain a conservative risk profile that aligns better with regulatory capital frameworks.

This trade-off has direct implications for model deployment in financial institutions. From a regulatory perspective, models must forecast accurately while avoiding risk underestimation, particularly during market crises. GARCH's tendency to overstate volatility may serve as a feature rather than a flaw in this context, leading to safer capital buffers, even if economically costly in tranquil periods.

Another critical finding concerns models that maintain systematically higher VaR levels, even without frequent backtesting breaches. Models like GARCH, which produce relatively high and conservative volatility estimates across both turbulent and calm periods, require banks to reserve more capital continuously—not just during stress episodes. While this behaviour reduces regulatory backtesting violations and supports compliance, it imposes a higher ongoing

capital burden during stable periods. Thus, institutions must weigh these economic trade-offs carefully, even though conservative estimates may be stabilising from a regulatory perspective.

Importantly, the findings also suggest that ML models exhibit notable strengths in adapting to rapidly changing volatility regimes. During periods of market turbulence, such as the COVID-19 crisis, these models capture structural shifts in volatility more quickly than traditional econometric models. This adaptability highlights a potential advantage of ML approaches in environments characterised by regime changes, where traditional models may react too conservatively or too slowly. However, despite their advantages, their use for regulatory capital purposes remains constrained by opacity and limited interpretability, which remain key challenges associated with model validation. Thus, their practical adoption in regulated environments must be approached cautiously. While backtesting performance is a critical factor in model validation, regulatory frameworks such as the Capital Requirements Regulation (CRR) (Regulation (EU) No 575/2013, 2013) and the Basel Framework (Basel Committee on Banking Supervision, 2026) emphasise conceptual soundness and governance requirements of internal models.

For that, traditional models like GARCH maintain a significant advantage: their parametric simplicity enables institutions to explain volatility forecasts in terms of past returns and variance shocks, facilitating internal validation, external auditability, and supervisory acceptance. By contrast, deep learning models such as LSTM and Autoformer, despite offering greater flexibility and predictive power, operate as opaque black-box systems. Their lack of inherent transparency complicates efforts to justify or defend risk forecasts to internal risk committees, auditors, and regulators. Without substantial supplementary explainability infrastructure, advanced modelling approaches may fall short of evolving regulatory expectations, increasing model risk and exposing institutions to supervisory challenges. For instance, the Basel framework requires risk managers to attribute fluctuations in Profit and Loss figures to specific risk factors, providing a clear and defensible explanation for each change. Moreover, the Basel framework requires the submission of comprehensive documentation that outlines the employed methodologies and evaluates their conceptual soundness in relation to the model's intended use case. Institutions are also expected to rigorously document all exceptions identified during ongoing backtesting exercises, including comprehensive explanations for each occurrence.

While recent literature emphasises post-hoc explainability methods (e.g., Giudici, Piergallini, et al., 2024) and adversarial robustness (e.g., Pialla et al., 2025) as important aspects of Trustworthy AI, their relevance in our setting is limited. Since the analysis is based on univariate

time series, feature attribution techniques such as SHAP or Shapley-Lorenz values provide little additional insight, as the only explanatory input is lagged volatility. Moreover, even traditional econometric benchmarks like GARCH are not inherently explainable in the sense of feature attribution, but remain accepted in practice due to their tractability and regulatory familiarity. Similarly, adversarial attack frameworks, often developed for classification or image recognition, are not directly applicable to volatility forecasting under supervisory constraints. In this context, stress testing across distinct market regimes provides a more meaningful evaluation of robustness. For these reasons, our focus is on metrics that align with supervisory practice—forecast accuracy, trajectory alignment, and regulatory backtesting—while leaving explainability techniques and adversarial robustness as promising areas for future research.

As deep learning models may prioritize pattern discovery over economic reasoning, they may overlook contextual factors, such as sudden geopolitical events, that affect risk relationships. This limitation emphasises the importance of ensuring that models learn economically meaningful relations and that model users can comprehend model functioning without oversimplifying critical dynamics. Deploying advanced ML models, especially those with transformer or extended memory architectures, requires substantially enhanced model governance frameworks compared to traditional models like GARCH. Financial institutions must establish detailed documentation, comprehensive validation protocols, ongoing model performance monitoring, and explainability mechanisms to satisfy internal risk management standards and external regulatory requirements. This substantially elevates both the governance-related costs and the expected competencies of the second line of defence in the context of ML models.

The choice between traditional and ML models thus entails not only a trade-off between forecast accuracy and risk conservatism but also a trade-off in governance burden and resource requirements. While ML models offer the potential for better real-time adaptability and predictive precision, they impose heavier demands on compliance frameworks, staffing, and ongoing maintenance efforts.

Considering these challenges, a pragmatic strategy could involve deploying advanced models like Autoformer as complementary tools to traditional models rather than direct replacements. For example, ML models could serve as internal monitoring tools or early-warning systems, while GARCH-based models remain the primary basis for regulatory capital calculations due to their transparency and ease of governance. Over time, as regulatory expectations become more concrete and model governance practices for ML evolve, broader adoption may become more feasible.

In the specific case of volatility forecasting for VaR prediction across varying market regimes, the findings underscore that model selection cannot be based solely on improvements in forecast accuracy metrics. Institutions must evaluate the broader implications for regulatory capital, model governance, and operational resilience. Models that achieve fewer backtesting violations but impose higher continuous capital costs, such as GARCH, must be carefully balanced against models that deliver lower average errors but carry greater supervisory scrutiny risk due to transparency deficits.

Ultimately, adopting ML models in financial risk management requires a careful balance between predictive gains and the practical realities of regulatory expectations, explainability requirements, and capital efficiency constraints. A sustainable risk modelling strategy must integrate these dimensions, ensuring that model innovation strengthens an institution's financial stability, regulatory compliance, and competitive advantage.

2.6 Conclusion and Future Research

This study compared traditional econometric models (GARCH) with advanced ML architectures (LSTM, xLSTM, Autoformer) and hybrids across stable, crisis, and recovery regimes. The results show that Autoformer delivers the lowest point forecast errors, xLSTM excels in trajectory and timing, while GARCH remains the most robust for Value-at-Risk backtesting due to its conservatism.

A central finding is the trade-off between predictive accuracy and regulatory conservatism. ML models adapt more effectively to structural shifts but raise challenges in transparency and backtesting compliance. GARCH, though less accurate, produces conservative forecasts that align with supervisory standards. This implies that ML models are best suited as complementary tools for monitoring and early warning, while GARCH remains the regulatory anchor.

Since global regulatory standards such as the Basel framework are applied across jurisdictions, these findings also carry implications for international banks and systemic resilience, where the balance between predictive accuracy and conservative risk estimation directly shapes cross-border capital requirements.

Future research could extend this framework in several directions. Hybrid approaches that integrate econometric structure with ML adaptability, including physics-informed architectures, may enhance both accuracy and conceptual soundness. Explainability methods could strengthen model validation and ease supervisory acceptance, while robustness testing against adversarial

perturbations or data poisoning would provide insights into stability under intentional disruptions. Extending the analysis to multivariate volatility and cross-market spillovers would further broaden the applicability of the results.

3 Bridging the AI Governance Gap: Systematic Weaknesses and Targeted Solutions in Financial Model Risk Management

3.1 Introduction

Artificial Intelligence (AI) is reshaping the financial services industry, driving improvements in efficiency, decision-making, and customer engagement. Its adoption in critical functions such as credit assessment, fraud detection, and advisory services, however, introduces new model risks that existing regulatory frameworks are only beginning to address. JP Morgan's 2024 launch of IndexGPT (J.P. Morgan, 2024), an AI-based stock picker, highlights the shift from traditional methods to generative AI (GenAI) systems and, hence, the need to adapt existing governance mechanisms.

While many banks have well-established Model Risk Management (MRM) practices, these frameworks were designed for more stable, predominantly quantitative models. They are insufficient in addressing the dynamic, non-linear behaviour of AI models, including those powered by generative technologies. Challenges such as explainability, bias, adaptive learning, and ethical considerations require targeted extensions to current practices. These issues are well documented in academic literature (Mittelstadt, 2019), reinforcing the need for regulatory alignment. The European Union's (EU) AI Act (AIA) (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024) introduces a horizontal, risk-based framework for AI, yet industry readiness remains uneven and model governance often lags behind AI deployment.

Despite increased attention, theoretical, regulatory, and academic guidance on how to integrate these emerging risks into prudential MRM remains limited. Prior work has identified broad AI governance gaps (Kurshan et al., 2021), but remains largely conceptual and not anchored in the internal logic of prudential model risk management. This study identifies structural misalignments between specific properties of AI and the control assumptions embedded in prudential model risk management frameworks for financial institutions. AI models introduce combinations of opacity, adaptive behaviour, non-deterministic outputs, and complex data sources that strain legacy assumptions embedded in traditional MRM. The paper therefore analyses AI governance challenges as recurring weaknesses at which these AI-specific characteristics disrupt established MRM lifecycle arrangements—across model definition,

validation, monitoring, governance, and change control. The relevance of AI governance in this manuscript arises not from AI as such, but from AI as a model-related source of prudentially relevant decision risk in banking contexts. The resulting weakness-to-enhancement mapping is accordingly intended as a sector-specific bridge between horizontal AI governance expectations and the internal logic of financial MRM, with an analytical boundary deliberately drawn at the intersection of AI characteristics and prudential model governance.

This study contributes by developing a lifecycle-based taxonomy of recurrent AI-related weaknesses in financial MRM and by interpreting these weaknesses as structural misalignments between AI-specific properties and established prudential MRM assumptions. It explains how AI-specific risks alter the application of core MRM concepts. This study is guided by the following research questions (RQ):

- RQ1: Which weaknesses arise when traditional MRM frameworks are applied to AI models in financial institutions?
- RQ2: Which governance enhancements can be derived to mitigate the identified weaknesses while remaining consistent with MRM principles?
- RQ3: How does the proposed weakness-to-enhancement mapping differentiate governance needs across representative AI use cases?

We address these questions through a structured evidence synthesis of academic and regulatory sources, complemented by case-based illustrations on representative AI use cases. The result is a lifecycle-oriented set of recurrent weaknesses and a traceable weakness-to-enhancement mapping.

Having outlined the main research questions and our contributions to the literature, the remainder of this paper is structured as follows. Section 2 reviews the current regulatory landscape and the corresponding academic and industry responses. Section 3 presents the research methodology. Section 4 details the twelve key weaknesses in current MRM practices and derives corresponding governance enhancements. Section 5 presents three case illustrations applying the taxonomy across representative AI use cases. Section 6 concludes.

3.2 Evolving Regulatory Landscape and Literature Review

Following advances in data availability and computational capacity, banks have increasingly adopted AI to gain competitive advantage (Machado, 2026). This shift has been accompanied by rapid growth in academic interest, reflected in a proliferation of publications and dedicated venues. While the literature documents multiple benefits of AI adoption, it also highlights significant ramifications, including unexpected and undesirable outcomes (Aldasoro et al., 2025). Consequently, the development of responsible AI principles has become a central focus, aiming to mitigate negative and unintended consequences (Baeza-Yates & Fayyad, 2024; Floridi & Cowls, 2021). These principles generally provide guidance for designing and deploying AI in ways that promote fairness, equity, and ethical outcomes for affected stakeholders across multiple domains (Radanliev, 2025; Sadek et al., 2025).

While these principles are frequently positioned as foundational to AI governance, a persistent gap remains in converting these considerations into practical and actionable frameworks (Batool et al., 2025; Prem, 2023). Systematic reviews indicate that the literature is dominated by discussions of core ethical values such as fairness, transparency, and accountability (Batool et al., 2025; Prem, 2023). At the same time, these reviews emphasise that many proposed approaches remain overly abstract and provide limited concrete guidance for implementation (Prem, 2023). Therefore, the literature increasingly calls for a shift from high-level principles toward practical tools, methodologies, and best practices that can operationalise AI governance in organisational settings (Kurshan et al., 2021; Papagiannidis et al., 2023; Prem, 2023; Zaidan & Ibrahim, 2024). As a result, recent research focuses on methods to evaluate the trustworthiness of AI by assessing the different risks technically (Giudici & Kolesnikov, 2026; Lacmanovic & Skare, 2025; Sayles, 2024). Building on such individual metrics, recent research integrates them into a single compliance metric to improve compliance monitoring and enable a consistent evaluation of trustworthiness (Giudici & Raffinetti, 2023; Rutinowski et al., 2024).

This need for operationalisation is particularly salient in banking, where models directly influence high-stakes decisions and are subject to established supervisory expectations (García-Llorente & Olmeda, 2026). Accordingly, financial-sector research has largely approached AI governance through the lens of MRM frameworks focusing in particular on transparency, explainability, and bias in credit scoring (Bücker et al., 2022; Du Toit et al., 2023; Garcia et al., 2024). Early studies pointed out weaknesses in applying existing MRM to AI. Souza (2023)

and Brotcke (2020) identified gaps with regard to the model definition, model validation, and the governance framework. Hill (2019) and Kurshan et al. (2021) emphasise explainability as a core control requirement and argue for more dynamic, risk-sensitive tiering approaches, implying that validation depth and evidence requirements should vary with model materiality and complexity. A growing body of work therefore extends traditional validation toolkits with methodologies for explainability, robustness, and fairness testing alongside established validation practices (Li & Goel, 2025; Myllyaho et al., 2021; Sandu et al., 2022; Sayles, 2024). Complementing these technical validation advances, Rao et al. (2025) advocate data-driven enhancements to model risk documentation and reporting to promote more consistent governance decisions as risk sections in model documentation are frequently brief, generic, or recycled.

MRM as a regulatory discipline began with the Federal Reserve's SR 11-7 guidance (Board of Governors of the Federal Reserve System, 2011) which established industry expectations around model governance, development, validation, and oversight. Over time, it has become the de facto global standard. The Office of the Comptroller of the Currency's MRM Handbook (Office of the Comptroller of the Currency, 2021) complements SR 11-7 by offering contextual and explicit guidance, specifically acknowledging the unique challenges posed by AI models. In Europe, regulatory progress has been more fragmented, with milestones such as the Capital Requirements Regulation (Regulation (EU) No 575/2013, 2013). More recently, the European Central Bank's Guide for Internal Models (European Central Bank, 2025) (EGIM) and the United Kingdom's Prudential Regulation Authority's (PRA) Supervisory Statement SS1/23 (Bank of England, 2023b) set minimum control expectations for all models, including those with qualitative and quantitative outputs and explicitly extends the scope to AI models. In alignment with SS1/23, the Financial Reporting Council released TAS 100: General Actuarial Standards v2.0 (Financial Reporting Council, 2023) and the accompanying Technical Actuarial Guidance: Models (Financial Reporting Council, 2024), both of which address AI-specific risks in the context of model development. The Bank of England has also noted that existing regulation already acknowledges the risks associated with AI (Bank of England, 2024; Breeden, 2024).

Despite this progress, it is increasingly recognised that traditional model governance is insufficient for managing AI-related risks. The National Institute of Standards and Technology (NIST) AI Risk Management Framework (National Institute of Standards and Technology, 2023), although developed in the United States, has influenced governance efforts globally and serves as a reference for voluntary structures within European banks. Industry practitioners are

calling for a harmonised, risk-tiered approach that distinguishes between conventional statistical models and AI systems.

The EU AIA introduces a risk-based classification framework and defines obligations for high-risk use cases, e.g., creditworthiness assessments. However, banks face challenges mapping these requirements onto legacy governance structures that were not designed to accommodate dynamic, autonomous, or generative systems. This misalignment is especially evident in domains such as cybersecurity, where AI models remain vulnerable to adversarial attacks that can compromise output reliability. Recent literature discuss the impact of AI governance with regard to AIA compliance (Gonzalez Torres et al., 2023) and discuss the AIA itself (Cancela-Outeda, 2024; Gstrein et al., 2024).

Complementary regulations, such as the Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (Regulation (EU) 2022/2554, 2022), the European Banking Authority's (EBA) Guidelines on Outsourcing (European Banking Authority, 2019), and the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679, 2016), expand compliance requirements relating to outsourced AI models and the processing/use of data.

At the international level, standards such as ISO/IEC 42001 (ISO/IEC, 2023b) introduce a certifiable AI Management System, specifying organisational governance practices including leadership responsibilities, documentation, and oversight. ISO/IEC 23894 (ISO/IEC, 2023a) complements this by detailing risk management processes for AI systems. The Federal Office for Information Security in Germany issued an AI in Finance Test Criteria (Federal Office for Information Security, 2025) that provides practical evaluation metrics for robustness, fairness, and auditability, with a specific emphasis on resistance to cyber-attacks.

Even with expanding literature and regulatory advances, research rarely translates AI-specific failure modes into implementable MRM. Existing contributions that identify governance gaps (Kurshan et al., 2021) are often conceptual and do not provide a traceable mapping from weaknesses to concrete governance mechanisms. Therefore, this study synthesises prior evidence to identify twelve recurring weaknesses in AI-related MRM practice and derives targeted enhancements that translate these weaknesses into MRM-relevant governance implications.

3.3 Methodology

Our methodology followed a structured, multi-stage qualitative evidence synthesis (Figure 3.1). We applied a four-stage process aimed at identifying, validating, and addressing weaknesses in existing MRM frameworks for AI through the development of targeted recommendations.

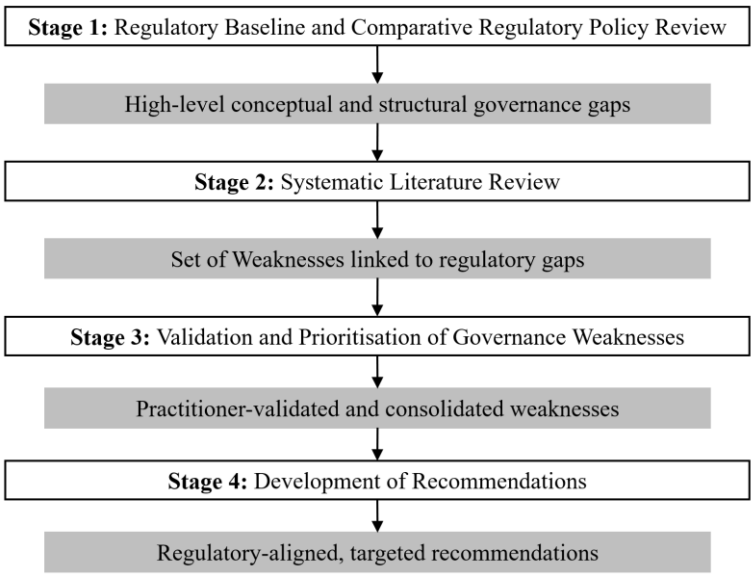

Figure 3.1: High-level Process of multi-stage qualitative evidence synthesis for identifying AI governance weaknesses and recommendations.

In the first stage, we identified high-level thematic gaps through a comparative review of regulatory and policy documents. In the second stage, we conducted a systematic literature review, which generated an initial inventory of weaknesses with regard to the conceptual misalignments between traditional MRM and AI governance needs. In the third stage, we applied data triangulation to evaluate the practical relevance and industry recognition of these preliminary weaknesses using publicly available expert commentary, including public consultation responses, white papers, and practitioner surveys. We assess whether the weakness identified in the first stage was explicitly supported and how strongly (e.g., frequency across sources). Each identified weakness was reassessed, refined, and prioritised based on its recurrence, regulatory alignment, and operational significance. This approach helped us bring together different perspectives from regulators, academics, and practitioners to ensure our findings were both well-informed and relevant to real-world settings. We recognise that some practitioner sources, such as consultation responses and white papers, may reflect organisational interests. In Stage four, we conduct a targeted follow-up literature review per prioritised weakness to derive targeted governance implications mapped to MRM lifecycle components.

1. **Regulatory Baseline and Comparative Regulatory Policy Review:**

   First, we selected SR 11-7 as the foundational benchmark due to its wide adoption and institutional relevance across the financial sector. Building upon this, we conducted a comparative analysis of AI-specific regulatory and policy frameworks spanning major jurisdictions, including the AIA (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024), EGIM (European Central Bank, 2025), PRA's SS1/23 (Bank of England, 2023b), General Technical Actuarial Standards and Technical Actuarial Guidance (Financial Reporting Council, 2023, 2024). Additional comparisons included the Central Bank of the United Arab Emirates's Model Management Standards and Guidance (Central Bank of the U.A.E., 2022), Japan's Principles for MRM (Financial Services Agency of Japan, 2021), the joint guidance on the use of GenAI and language models issued in Hong Kong and Singapore (The Securities and Futures Commission, 2024), and the Monetary Authority of Singapore's 2024 publication on AI MRM (Monetary Authority of Singapore, 2024). Although EGIM (European Central Bank, 2025) and PRA's SS1/23 (Bank of England, 2023b) are not standalone AI governance frameworks, we included them as they set model governance expectations that are directly relevant to AI models. This analysis allowed us to identify conceptual and structural divergences between conventional MRM practices and AI governance expectations. These gaps were thematically grouped into core components of SR 11-7, such as model definition, model inventory, model development and validation, monitoring, change management, and third-party risk. This step established a baseline for evaluating where existing MRM principles diverge from emerging AI-specific regulatory expectations and guided the thematic structure used throughout the study.

2. **Systematic Literature Review:**

   Building on the identified gaps in stage one, we conducted a systematic review of peer-reviewed literature, guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Page et al., 2021) (Figure 3.2) using the Scopus and Web of Science databases. These databases were selected for their comprehensive coverage of scientific publications and robust analytical capabilities. With Web of Science prioritising high-impact journals and Scopus offering broader coverage, including conference proceedings, we used both to enhance representativeness and reduce thematic skew given their complementary indexing scopes (Mongeon & Paul-Hus, 2016). Specifically, we applied the following search string: ((("model risk*" OR "AI model risk*") AND ("AI" OR "Artificial Intelligence*")) OR (("model risk*" OR "model risk management") AND

("Governance")) OR (("model risk*" OR "AI model risk*") AND ("AI" OR "Artificial Intelligence*") AND ("Gap*" OR "Challenges")) OR (("model risk*" OR "risk modeling*") AND ("AI" OR "Artificial Intelligence*") AND ("finance*" OR "financial*") AND ("manage*" OR "management*" OR "assessment*")) OR (("model risk management" OR "AI model risk*") AND ("AI" OR "Artificial Intelligence*") AND ("Governance")) OR (("model risk*" OR "risk modeling*") AND ("AI" OR "Artificial Intelligence*") AND ("finance*" OR "financial*") AND ("manage*" OR "management*" OR "assessment*") OR (("AI" OR "Artificial Intelligence*") AND ("Governance"))).

We searched for documents on Scopus based on article titles, abstracts, and keywords, identifying 11,891 documents, and applied filters to narrow our sample of published articles in this field. First, we limited the document type to peer-reviewed journal articles and conference papers. Secondly, we limited the language to English. Thirdly, we restricted the documents to subject areas including computer science, sociology, business, economics, finance, mathematics, and decision science. As a result, we reduced the sample from Scopus to 9,730.

In parallel, we searched the Web of Science for documents in all fields, identifying 8,474 documents. To ensure relevancy, we limited the language to English and refined document types to articles, review articles, and early access items. Next, we applied the Web of Science research area filter, refining research to computer science, business economics, social issues, sociology, mathematics, government law or public administration, or behavioural sciences. This yielded 4,821 Web of Science records. Next, we merged the Scopus and Web of Science results and removed duplicates across the combined dataset, resulting in 11,821 unique records. We then screened abstracts and keywords for governance- or risk-related terms, including “model risk”, “AI risk”, “risk management”, “AI governance”, “challenges”, “regulation”, “regulatory compliance”, and “guideline” as well as related variants. Relevance screening based on keywords and abstracts reduced the dataset to 506 articles published between 2015 and 2025, which were subsequently subjected to full-text review. Of these, 170 articles were excluded since they do not address the research questions, resulting in a final sample of 336 published studies. The specific criteria that formed the basis for inclusion or exclusion of individual studies are summarised in Table 3.1. Then, each of the 336 articles was analysed to assess whether they explicitly addressed the regulatory gaps identified in stage one or engaged with broader implications for MRM and AI-related

risks. The resulting weaknesses were subsequently mapped to the relevant regulatory expectations.

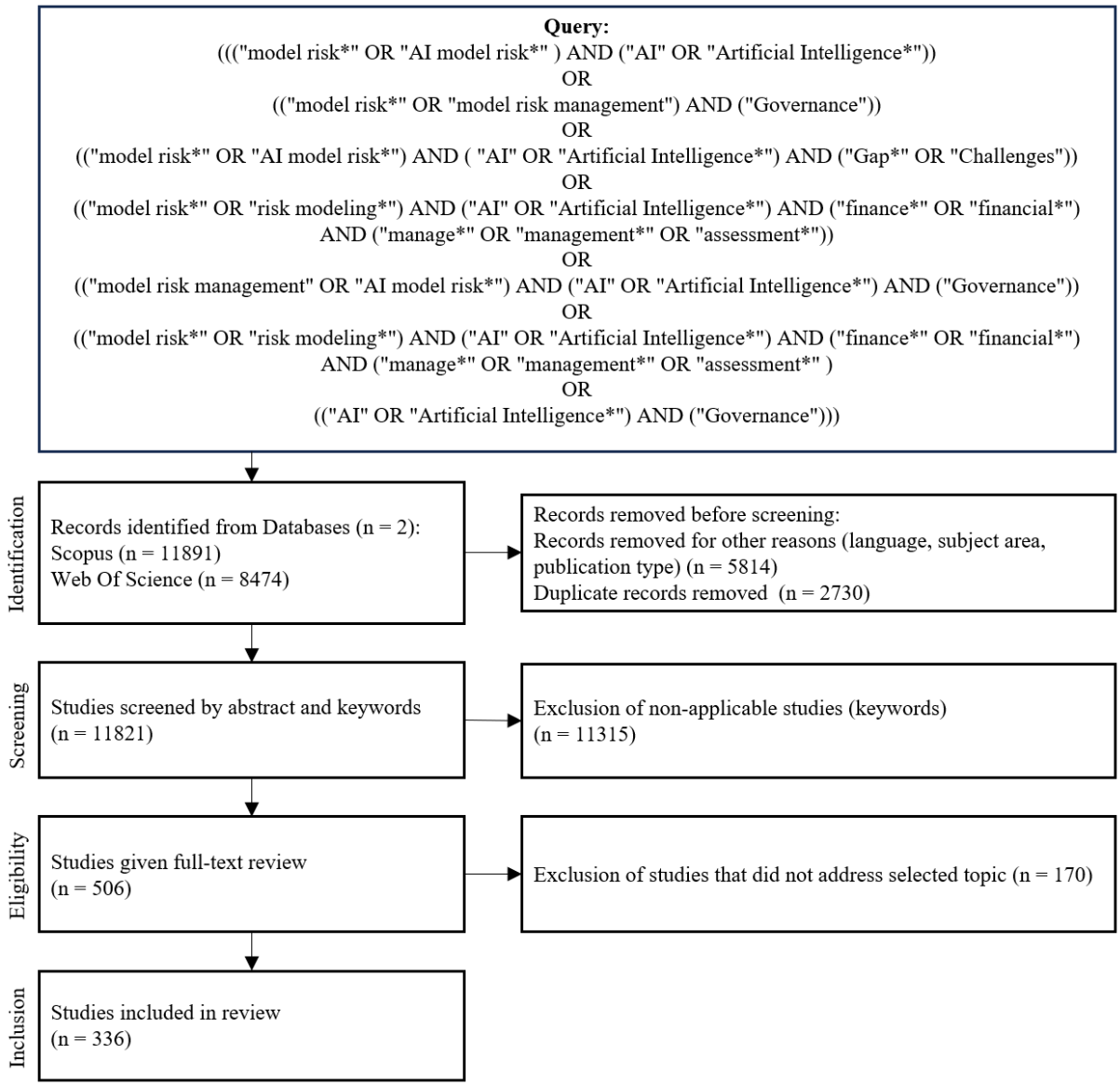

Figure 3.2: PRISMA flow diagram depicting steps of literature search and selection for included studies.

Table 3.1: Eligibility Criteria for Full-Text Review.

Inclusion Criteria	Exclusion Criteria
• Studies that address AI governance or model risk for AI models • Studies that address challenges related to AI risks • Full-text availability • Focus on implementation challenges and strategies	• Studies primarily focused on technical AI development or application without governance/risk-management implications • Studies focusing exclusively on specific algorithms, technical implementations, or niche applications • Studies without governance focus or focus on challenges related to AI risks

3. **Validation and Prioritisation of Governance Weaknesses:**

   To assess, validate, and prioritise preliminary weaknesses identified in stage two, we analysed publicly available expert commentary in the form of 10 public responses to regulatory consultations (American Bankers Association, 2024; Bank of England, 2023a; Breeden, 2024; Federal Financial Supervisory Authority, 2022; The Global Financial Markets Association, 2025; New York State—Department of Financial Services, 2024; PwC, 2021; The Securities Industry and Financial Markets Association and its Asset Management Group, 2024; The Consumer Bankers Association, 2024; U.S. Department of the Treasury, 2024) as well as 20 industry surveys and white papers (Boston Consulting Group, 2024; Deloitte, 2021, 2023, 2025b, 2025c; KPMG, 2022, 2023a, 2023b, 2024a, 2024b, 2024c; McKinsey & Company, 2025a, 2025b; Picciarelli, 2024; PwC, 2022, 2023, 2024; Renieris et al., 2023; Risk Management Association, 2022; SAS & GARP, 2020). These practitioner-facing sources were particularly valuable because they reflect real-world implementation challenges, recurring risk themes, and areas of friction that may not be visible in academic discourse. For each candidate governance challenge, we applied a binary flag indicating whether it was referenced explicitly or implicitly in the practitioner material. We then consolidated overlapping challenges and prioritised those with the strongest and most consistent practitioner support, resulting in 12 operationally significant weaknesses. These validated weaknesses reflect the most pressing structural limitations of current MRM frameworks when applied to AI models and provide the foundation for the targeted recommendations developed in the next step.

4. **Development of Recommendations:**

   Targeted recommendations were developed through a layered approach. For each prioritised weakness from stage three, we derived recommendations through an iterative evidence synthesis using data triangulation. We first conducted a weakness-specific targeted literature review to identify governance mechanisms and recurring implementation patterns reported in peer-reviewed research, drawing on grey literature only where necessary to contextualise rapidly evolving topics. Where the AI-specific literature was incomplete, we complemented it with evidence and well-established methods from adjacent domains (e.g., software engineering, human–computer interaction) to refine the specification of governance mechanisms. We then corroborated these literature-derived findings against supervisory guidance and regulatory commentary, as well as relevant standard-setting artefacts and working-group outputs (DIN & DKE, 2022; ISO/IEC, 2023a, 2023b) to assess regulatory consistency and

practical feasibility. Finally, we consolidated the resulting governance mechanisms into proportionate governance enhancements mapped to MRM lifecycle domains and explicitly linked to the corresponding weakness.

3.4 Enhancing AI Model Risk Management: Weaknesses and Recommendations

The following weaknesses should be read as lifecycle-specific manifestations of a broader conceptual problem: AI changes the assumptions under which prudential MRM frameworks identify, classify, validate, monitor, and govern models. Traditional MRM frameworks were primarily shaped by more stable, predominantly quantitative models with scopes defined by well-specified financial applications. Their behaviour can typically be assessed through established performance, benchmarking, and monitoring routines, while responsibilities for use, ownership, and validation can be delineated across known risk functions.

By contrast, AI models may combine non-deterministic or adaptive behaviour, opacity and limited explainability, complex data, and qualitative or narrative outputs. These characteristics strain established MRM concepts because they complicate independent validation and effective challenge, blur the boundary of what constitutes a prudential model, and weaken assumptions about stable model behaviour over time. In particular, adaptive behaviour and non-stationary inputs challenge periodic monitoring and change management routines; opacity and external dependencies complicate validation evidence and ownership; and qualitative, narrative outputs disrupt traditional distinctions between models, tools, and broader AI systems.

Against this backdrop, the twelve weaknesses developed in this section identify where AI-specific properties strain concrete MRM concepts across the model lifecycle. They show how AI affects governance and organisation, scope and inventory, third-party model risk, development and validation, deployment and monitoring, and change management. The corresponding governance enhancements therefore translate horizontal AI governance concerns into proportionate extensions of prudential MRM.

Governance and Organisation

The governance and organisation weaknesses concern the institutional location of model risk ownership and challenge. This directly relates to prudential expectations that model risk management should be supported by clear roles, responsibilities, accountability, and effective

challenge throughout the model lifecycle (Bank of England, 2023b; Board of Governors of the Federal Reserve System, 2011). Traditional MRM frameworks tend to assume that model development, use, validation, and oversight are concentrated in established financial-risk domains such as credit risk, market risk, or treasury. AI adoption weakens this assumption because AI models are increasingly developed, procured, or used by functions that historically operated outside the MRM perimeter. The conceptual issue is therefore not merely that more stakeholders must be informed, but that the organisational perimeter of model risk has expanded beyond the areas in which MRM expertise, accountability structures, and escalation routines have traditionally been concentrated.

Challenge 1 (C1): Advances in AI have introduced MRM to functions that were not previously involved

The growing adoption of AI in banking has extended MRM responsibilities beyond traditional risk domains such as credit risk and market risk. AI models are increasingly utilised in customer services, human resources, and other operational areas, where risk management expertise is often lacking. This broader use has led to an increased exposure to model risk across various departments unfamiliar with managing these risks.

Recommendation 1 (R1): All relevant new stakeholders must be identified to clearly define the key processes—along with their corresponding roles and responsibilities

Under prudential MRM expectations, clear allocation of model ownership, use responsibility, and effective challenge across the three lines of defence is a foundational governance requirement (Bank of England, 2023b; Office of the Comptroller of the Currency, 2021). AI-driven expansion of model use beyond traditional risk domains challenges this expectation as development, use, and reliance increasingly occur in functions that have historically lain outside the MRM perimeter. The conceptual task for MRM is therefore not simply to add more stakeholders, but to restore a coherent perimeter of accountability in which ownership and escalation rights remain traceable as AI models move into customer-facing, operational, or enterprise functions.

The institution should therefore identify and list the individuals, groups, or entities that have an interest in AI use cases. The purpose of this exercise is to ensure that model ownership, use

responsibility, oversight, and escalation accountability remain clear when AI models are developed or used outside traditional risk functions (Cosma et al., 2023; Mayenberger, 2019). New actors from the first line of defence should be included, along with increased involvement from existing ones such as compliance, legal, and information technology (IT) (Kallina & Singh, 2024). As risk management teams are already familiar with the application of models, they may need to provide additional training to these stakeholders. Resistance to change may be encountered and should be addressed (Eskandarzadeh, 2026). The identified roles and responsibilities should be mapped clearly within the MRM framework to ensure traceable accountability across the model lifecycle. In particular, institutions should distinguish between responsibility for the AI use case and responsibility for the underlying algorithm or technical model component, which may rest with business and IT stakeholders, respectively. This distinction is illustrated by Tsai (2025) in the context of data governance.

To remain effective, stakeholder mapping should not be treated as a one-off task. As AI use cases and regulatory expectations evolve, the allocation of ownership, responsibility, and oversight may also need to be reassessed (Hill, 2019). Periodic or trigger-based review can therefore help ensure that the governance perimeter remains aligned with the model's actual use and risk profile. This also requires attention to capabilities: stakeholders involved in the MRM lifecycle should have skills and training expectations that correspond to their role and to the complexity and risk of the AI model, particularly where models are dynamic, opaque, or based on machine learning (ML) techniques. This helps ensure that all relevant stakeholders—across the three lines of defence (3LoD)—are sufficiently equipped to understand, challenge, and manage model risk in accordance with evolving regulatory expectations (Bank of England, 2023b; European Central Bank, 2025).

Challenge 2 (C2): Missing committee oversight for AI models

Traditional model approval processes may lack the granularity and AI-specific focus needed to assess risks such as hallucinations, adaptive behaviour, fairness concerns, conduct risk, compliance implications, and accountability gaps. These risks often bypass existing governance layers without formalised structures dedicated to assessing the distinct nature of AI models.

Recommendation 2 (R2): Introduce a dedicated AI model approval body or extend the existing committee's scope

The institution should either set up a separate approval committee for AI models or expand the mandate of an existing model governance committee to address AI-specific risks. The central point is that AI-related model risks should be assessed within a governance structure that has sufficient authority, cross-functional expertise, and a clear link to the institution's risk appetite (European Central Bank, 2025). This body should include cross-functional representation from compliance, legal, data science, business, and IT, including senior management (European Central Bank, 2025; Schuett et al., 2025). To ensure effectiveness, the committee should have a clear mandate, defined escalation rights, and a review cadence proportionate to model risk (European Central Bank, 2024; Financial Stability Board, 2014). Its role should span AI use cases, initial risk classifications, approval of model development, and oversight of post-deployment behaviour through the review of monitoring reports. The indicators from these reports should be interpreted against a pre-defined AI-specific risk appetite approved by the committee (Bank of England, 2023b).

It is important to note that incentives to critically challenge first-line developments may be limited in such committees, especially when senior management pressure exists to approve AI use cases quickly (Bennett, 2017). This may reduce the committee's effectiveness if not addressed explicitly through governance safeguards. Escalation criteria should therefore be risk-based and linked to materiality, model limitations, monitoring exceptions, and approved-use breaches (European Central Bank, 2024; Financial Stability Board, 2014). The committee's activities should be documented transparently and integrated into enterprise-wide governance reporting to enable accountability and alignment with senior management risk appetite (European Central Bank, 2025; Hill, 2019). In this way, model oversight remains anchored in established MRM concepts such as approval authority, effective challenge, risk appetite, and ongoing monitoring (Bank of England, 2023b; European Central Bank, 2025; Souza, 2023).

Challenge 3 (C3): Limited collaboration between the Lines of Defence

In many organisations, collaboration across the 3LoD is siloed, especially for AI use cases. AI risks often go undetected early in the lifecycle, with second-line validation or third-line audit involvement occurring too late to influence development.

Recommendation 3 (R3): Introduce a collaborative model lifecycle while preserving independent challenge

For material or complex AI models, MRM processes should move beyond purely sequential gatekeeping by incorporating early cross-functional engagement across the three lines of defence, for example through shift-left or tollgate arrangements. This allows model-risk, compliance, and documentation concerns to be identified while key development choices remain adaptable (Brunner-Kirchmair & Wiener, 2019; Sinha & Lee, 2024). Practical mechanisms such as cross-functional working groups or regular touchpoints between model developers and validators may support this interaction where proportionate. SR 11-7 and SS1/23 require, however, that model validation be conducted independently of model development—meaning that such collaboration must be structured to preserve the validator's ability to independently challenge model decisions, distinguishing this from generic enterprise shift-left practices.

These practices should be formalised in policies through interaction requirements that vary with the model's risk tier (Bank of England, 2023b; European Central Bank, 2025). For higher-risk AI models, earlier second-line involvement may be needed during data selection and model design, while internal audit can assess whether collaborative arrangements preserve effective challenge and remain auditable (ISO/IEC, 2023b). This ensures that lifecycle collaboration supports, rather than weakens, the prudential MRM principles of independence, accountability, and effective challenge (European Central Bank, 2025).

Scope and Model Inventory

The scope and inventory weaknesses concern the boundary between broad AI governance and prudential model governance. This is consistent with prudential expectations that model definitions, inventories, and tiering approaches should determine the scope and intensity of MRM in a risk-based and proportionate manner (Bank of England, 2023b). Traditional model definitions and inventories are often anchored in clearly delimited quantitative models, relatively stable outputs, and established financial applications (Board of Governors of the Federal Reserve System, 2011). AI models challenge this boundary because they may generate qualitative, narrative, or non-deterministic outputs while still influencing decisions, controls, customer outcomes, or risk-relevant processes (Brotcke, 2020).

For this reason, the distinction between AI models is important. In this manuscript, an AI model is understood as the inference-performing component that produces outputs from inputs,

whereas an AI system refers to the broader application context in which one or more such model components are embedded, including interfaces, retrieval layers, and workflows (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024). This distinction is also relevant because the EU AIA primarily assigns risk-based obligations at the level of AI systems and their intended use, while also imposing specific obligations on providers of general-purpose AI models. Prudential MRM is concerned with the model or model component that meets the institution's model definition, not automatically with the wider AI system in its entirety. Components of the wider AI system become relevant for MRM only where they materially influence prudential or business-critical decisions (Bank of England, 2023b). The relevant task is therefore to distinguish broad identification of potentially model-relevant AI artefacts from proportionate MRM treatment through subsequent model risk tiering, and governance intensity (Bank of England, 2023b).

Challenge 4 (C4): AI models often fall outside the model definition scope

Current model definitions in many banks emphasise quantitative outputs or predefined business applications. AI models, especially large language models (LLM) and hybrid models, often produce narrative outputs or learn adaptively, circumventing conventional inventories.

Recommendation 4 (R4): Clarify model definition for AI models

Banks often adopt a slightly modified definition of a model: broad enough to align with the regulatory requirements of SR 11-7, yet specific enough to exclude routine spreadsheets from being classified as models (Board of Governors of the Federal Reserve System, 2011; Bridgers et al., 2020; Hill, 2019). The definition should also acknowledge the blurred boundaries between tools, models, and AI systems, especially where model outputs can be qualitative or narrative. This includes LLMs, recommendation engines, and generative models that integrate structured and unstructured data. While such models may not conform to traditional quantitative model classifications, they may still materially influence business or risk outcomes and should therefore be identified, screened, and assessed against the institution's model definition and materiality criteria (Bank of England, 2023b). However, AI systems should not be treated as models in their entirety for prudential MRM purposes. Rather, an AI-enabled model component may be embedded within a wider AI system that also includes retrieval layers, workflows, and human interaction mechanisms (Regulation (EU) 2024/1689 (Artificial Intelligence Act),

2024). Prudential MRM should therefore focus on the embedded model or model component. Broader system-level risks should remain primarily governed through enterprise AI governance or IT governance.

To this end, MRM policies should adopt a more inclusive identification approach while preserving a proportionate boundary for full MRM treatment. Potentially relevant AI artefacts should first be captured through the model identification process. Artefacts that meet the institution's formal model definition should then be recorded in the model inventory and subjected to an initial model risk assessment, the outcome of which should determine the applicable intensity of subsequent MRM requirements by reference to the model tiering. In particular, the definition set out in PRA SS1/23 is helpful because it does not rely solely on the nature of a model's output and instead supports a functional, risk-based assessment of whether an artefact should fall within the model governance perimeter (Bank of England, 2023b).

Ultimately, clarifying the scope of what qualifies as a model is not a purely semantic exercise. It ensures that high-risk AI-enabled model components are not overlooked, while preserving a clear boundary between prudential MRM and broader enterprise AI or IT governance. A clear model definition also provides a practical basis for determining what constitutes a model, what should be captured as a model-related dependency, and which wider system-level risks should remain outside full prudential MRM treatment. In practice, this reduces borderline debates and supports proportionate, risk-based governance.

Challenge 5 (C5): Existing tiering frameworks fail to reflect AI complexity

Tiering models based primarily on financial impact or type of use may overlook AI-specific characteristics such as opacity, adaptiveness, scale of use, and third-party reliance, as well as their implications for fairness, conduct, and compliance.

Recommendation 5 (R5): Incorporate AI-relevant criteria into model risk tiering

Model tiering frameworks designed for traditional models should be updated to reflect the distinct characteristics of AI and to support effective, proportionate MRM. The presence of AI should not itself determine the model's tier; rather, AI-specific characteristics should be assessed insofar as they affect established MRM determinants such as materiality, complexity,

extent of use, uncertainty, reliance, and potential adverse consequences of error or misuse (Bank of England, 2023b).

Institutions should therefore extend existing tiering matrices to include AI-relevant attributes such as model opacity, data sources, retraining frequency, explainability constraints, and the potential impact on downstream decisions as AI models may exhibit greater complexity compared to traditional models (Cohen et al., 2023; European Central Bank, 2025). For example, opacity may increase validation uncertainty, retraining and data drift may increase monitoring demands, third-party AI components or external APIs may increase dependency and lineage risk, and GenAI models may widen the gap between intended purpose and realised use. The expectations set out in PRA SS1/23 provide a strong basis for extending existing frameworks by offering relevant AI-specific risk dimensions (Bank of England, 2023b).

Incorporating elements from the risk classification approach outlined in the AIA may enhance consistency with broader AI governance and regulatory readiness. However, such elements should supplement rather than replace prudential model risk tiering. A separate AI-only classification scheme may introduce redundancy and confusion, particularly where conventional AI models, such as decision trees, are already covered under existing MRM policies. Tiering outcomes should therefore influence the intensity of validation, monitoring frequency, documentation depth, approval authority, and change-management requirements, ensuring proportionality without overlooking material risks introduced by AI (Bank of England, 2023b; European Central Bank, 2025).

Challenge 6 (C6): Model inventories do not capture AI-relevant metadata

Traditional model inventories operationalise model risk by making models visible, classifiable, and auditable. For AI models, this function may be weakened when conventional inventory fields fail to capture risk-relevant attributes such as training-data lineage, retraining frequency, third-party dependencies, and explainability constraints. As a result, the inventory may provide an incomplete view of the model's risk profile (Cohen et al., 2023). The weakness is therefore not merely incomplete documentation, but a mismatch between traditional inventory taxonomies and the risk-generating properties of AI models and their model-related dependencies.

Recommendation 6 (R6): Expand inventory fields to reflect AI-specific risk attributes

In MRM, the model inventory functions as a core governance artefact rather than a purely administrative register, enabling institutions to demonstrate the scope of model governance, the relative materiality of models, and the evidence base for monitoring and effective challenge (Bank of England, 2023b; European Central Bank, 2025). For AI models, inventories shaped around conventional quantitative models are often too narrow, because they omit information on training data, retraining practices, external dependencies, and explainability status that is necessary to assess whether supervisory expectations around conceptual soundness, ongoing performance, and approved use and operational resilience (Regulation (EU) 2022/2554, 2022) remain satisfied over time. The model inventory should therefore be expanded to capture AI-specific metadata, including training-data sources, retraining frequency, version history, and explainability constraints. It should also record model-relevant dependencies such as embedded components, external application programming interfaces (APIs), and vendor components (European Central Bank, 2025).

To enhance documentation consistency and transparency, institutions may use standardised reporting tools, such as model cards, which consolidate key information into an accessible and auditable format (Mitchell et al., 2019). These enhancements improve traceability by linking inventory records to usage logs, incident reports, validation evidence, and change-management records, thereby supporting effective oversight throughout the AI model lifecycle without treating every surrounding IT component as a model.

Third-Party Models

The third-party model weaknesses concern the distribution of model development, transparency, and accountability across external providers and internal users. Traditional MRM assumes that the institution can understand, document, challenge, and validate the model development process sufficiently to evidence model ownership and effective challenge (Board of Governors of the Federal Reserve System, 2011). Third-party AI models weaken this assumption because key elements of model design, training data, architecture, retraining, and performance monitoring may sit outside the institution’s direct control. The conceptual issue is therefore not only vendor dependency (Regulation (EU) 2022/2554, 2022), but the preservation of MRM accountability when development evidence, transparency, and change visibility are distributed across external providers. This requires institutions to adapt validation, monitoring,

contractual, and governance practices so that external reliance does not undermine internal responsibility for model risk (Hill, 2019).

Challenge 7 (C7): Lack of control over model development when vendors or external models are used

Financial institutions increasingly rely on third-party models (i.e., externally developed models or model services, including open-source and pre-trained models integrated into production, vendor-provided models, and externally hosted model APIs), often with limited transparency about the models' structure, data sources, and retraining processes (McKinsey & Company, 2025a). Traditional MRM assumes full control over model development, which no longer holds.

Recommendation 7 (R7): Establish vendor governance processes for external AI models

Externally sourced AI models and model components should be assessed through an MRM-relevant due diligence process covering intended use, limitations, explainability constraints, versioning, retraining, and licensing, with involvement from MRM and other relevant control functions where the model materially affects business-critical or prudential decisions. Therefore, vendor risk assessments (Regulation (EU) 2022/2554, 2022) should be integrated into model approval workflows, and contracts should include MRM-relevant obligations such as notification of model changes, audit rights, and compliance with AI governance standards—reflecting the prudential principle that outsourcing development does not transfer model risk ownership, as confirmed by SR 11-7, SS1/23, and the EBA Guidelines on Outsourcing (Bank of England, 2023b; Board of Governors of the Federal Reserve System, 2011; European Banking Authority, 2019; Regulation (EU) 2022/2554, 2022; García-Llorente & Olmeda, 2026).

To preserve accountability, institutions should maintain sufficient transparency over externally sourced models to support validation, monitoring, and change management. This may include periodic vendor reviews, contractual triggers for material model changes, and centralised records of vendor-provided models, version metadata, and validation status (Board of Governors of the Federal Reserve System, 2011; Regulation (EU) 2022/2554, 2022).

Certification requirements specified in ISO/IEC 42001 (ISO/IEC, 2023b), recognised AI ethical guidelines, or the voluntary Code of Practice (The European Commission, 2025) for general-purpose AI may serve as additional reference points, particularly for higher-risk AI vendors.

Model Development and Validation

The development and validation weaknesses concern the evidentiary basis on which conceptual soundness and fitness for purpose are established. Traditional model development and validation standards focus on methodology, data quality, performance testing, benchmarking, and limitations of use. AI models, especially GenAI, challenge this paradigm because relevant risks may arise from model design, guardrails, user interaction, opacity, non-deterministic outputs, fairness concerns, and robustness limitations. The implication is not that traditional validation is obsolete, but that it is incomplete when applied alone (Mayenberger, 2019). AI validation therefore requires a broader evidentiary basis that combines outcome testing with process-based assessment, explainability analysis, robustness testing, fairness analysis, and evaluation of the use context. This extends prudential expectations that models should be conceptually sound, tested against intended use, independently challenged, and assessed with a level of rigour proportionate to model materiality and complexity (Bank of England, 2023b).

Challenge 8 (C8): Development guidelines insufficiently cover GenAI-specific risks

Guidelines for developing AI models often focus on validation accuracy or technical feasibility, overlooking critical GenAI risks like hallucinations, prompt misuse, and harmful content generation.

Recommendation 8 (R8): Introduce tailored GenAI development controls

While traditional AI development guidelines focus primarily on metrics like accuracy, they often overlook critical GenAI-specific risks such as model hallucinations, toxic or biased outputs, and adversarial “jailbreaks” that override safety guardrails (Gehrmann et al., 2025; National Institute of Standards and Technology, 2024). As a result, these guidelines on their own are insufficient for demonstrating that GenAI models satisfy prudential expectations around conceptual soundness and appropriate use. Therefore, institutions should update AI development processes, in line with emerging supervisory expectations, to include explicit

controls targeting these GenAI-specific vulnerabilities (Crisanto et al., 2024; Gehrmann et al., 2025). These vulnerabilities show that development standards for GenAI models should explicitly account for how the model behaves in its deployment environment, not only for its internal performance, such as user interaction and guardrails. These elements should be considered as part of the model's use context to the extent that they affect conceptual soundness and fitness for purpose (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024; ISO/IEC, 2023b; National Institute of Standards and Technology, 2023).

Development documentation should therefore evidence how the GenAI model component is designed to remain aligned with its intended use and known limitations , including the role of prompts, retrieval components, guardrails, and user interaction in shaping outputs (Gehrmann et al., 2025; ISO/IEC, 2023b; National Institute of Standards and Technology, 2024). Depending on the use case, this may involve mechanisms such as retrieval-based grounding on curated knowledge bases, adversarial and red-team testing, and clearly defined fallback behaviour (National Institute of Standards and Technology, 2024; Sudhir et al., 2025). The specific mechanisms and their depth should be proportionate to model materiality and potential consequences of misuse; the relevant MRM question is which dimensions of GenAI behaviour must be evidenced to support validation, approval, monitoring, and change management.

Challenge 9 (C9): Conventional validation approaches are not suited for AI models

Many AI models cannot be assessed sufficiently using standard backtesting or performance sensitivity tests alone. Their outputs may be non-deterministic, opaque, or driven by non-repeatable training contexts.

Recommendation 9 (R9): Use hybrid validation combining outcome- and process-based testing

From an MRM perspective, model validation is the primary means by which conceptual soundness, fitness for approved purpose, and the adequacy of development evidence are independently assessed. For AI models, this requires extending established validation approaches to capture the dimensions along which AI-specific risks emerge (García-Llorente & Olmeda, 2026). While traditional AI models are often amenable to established validation techniques, the need for explainability becomes increasingly critical as model complexity rises (Mayenberger, 2019). The EGIM and SS1/23 explicitly require the use of explainability tools

during validation to ensure that outputs can be interpreted (Du Toit et al., 2023), independently challenged, and effectively monitored by internal stakeholders across all three lines of defence (Bank of England, 2023b; European Central Bank, 2025). Building on this, effective validation of GenAI models requires an even more tailored approach, combining both outcome-based and process-based assessments (Maheshwari, 2026). This distinction is particularly important for generative models, where conventional performance metrics alone are insufficient to capture output quality.

The outcome-based approach focuses on quantifiable evaluation methods, including distribution monitoring, robustness testing, and fairness assessments (Maheshwari, 2026). Explainability tools such as SHAP and LIME can support independent validation by helping validators interpret feature contributions to individual predictions (Bücker et al., 2022; Du Toit et al., 2023), and identify biases or irregularities that may not be visible in aggregate performance metrics. Fairness toolkits (Bellamy et al., 2019) can complement this analysis by assessing model behaviour across demographic groups and monitoring potential fairness degradation over time (Lacmanovic & Skare, 2025).

The process-based approach examines the integrity of the model development lifecycle. This includes verifying the completeness of documentation, the plausibility of underlying assumptions, and the model's alignment with business context, intended use, institutional risk appetite, and conceptual soundness (Bank of England, 2023b; Mayenberger, 2019). Validation teams should involve business stakeholders and compliance professionals in scenario-based walkthroughs to assess model behaviour under edge cases and high-risk conditions. This supports ethical and regulatory alignment beyond statistical adequacy.

Together, these approaches form a multidimensional validation framework. This is especially critical for AI models that undergo periodic retraining, where validation must address feature drift, evolving input data, and the accumulation of unintended bias. By combining quantitative testing with qualitative oversight, institutions can enhance both the defensibility and auditability of AI models, while meeting supervisory expectations for transparent and accountable model use (Sudjianto & Zhang, 2024).

Deployment & Monitoring

The deployment and monitoring weaknesses concern the stability of approved model use after implementation. Traditional MRM generally assumes a defined approved use, known limitations, relatively stable model logic, and monitoring indicators that can be reviewed

periodically (Board of Governors of the Federal Reserve System, 2011). AI models challenge these assumptions because their outputs may be interactive, context-dependent, persuasive, and easily repurposed beyond the original intended use, while model behaviour may evolve through retraining, data drift, prompt patterns, external model updates, or changing user behaviour. The conceptual issue is therefore not simply insufficient user guidance or infrequent monitoring, but the instability of the approved-use and performance boundary itself. Monitoring must accordingly move from periodic confirmation of continued performance toward a more dynamic assessment of drift, fairness, robustness, and emerging misuse. This is consistent with prudential expectations that model use should remain aligned with intended purpose, that limitations should be understood by users, and that monitoring should remain proportionate to model use, materiality, and changing conditions (Bank of England, 2023b; Board of Governors of the Federal Reserve System, 2011).

Challenge 10 (C10): Model use is poorly defined for AI, increasing misuse risk

AI models are often used without clear boundaries regarding scope, override rights, or conditions under which their outputs are acceptable for decision-making. This increases the likelihood of inappropriate reliance or biased use.

Recommendation 10 (R10): Define AI usage protocols and decision constraints

AI outputs are often interactive, context-dependent, and persuasive, which increases the risk that users rely on them outside the model's approved use and known limitations—a form of misuse that SR 11-7 explicitly identifies as a primary source of model risk (Board of Governors of the Federal Reserve System, 2011). Institutions should therefore define usage protocols that link approved use, model limitations, human judgement, and decision accountability to user roles and decision criticality.

In high-risk contexts such as credit underwriting or regulatory reporting, AI-generated recommendations should be treated as decision support: qualified human users must review outputs before action is taken, override capabilities should be mandated, and override frequency should be monitored as a governance signal indicating either effective oversight or potential usability/trust issues (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024). In lower-risk applications such as internal text summarisation, more automated use may be

acceptable provided users understand the system's limitations and the organisation's tolerance for error.

The core risk is not only incorrect outputs but inappropriate reliance—users deferring to model suggestions even when they conflict with context or professional judgement. In line with Bhatt and Sargeant's concept of "algorithmic resignation" (Bhatt & Sargeant, 2024), institutions may therefore actively restrict AI use in situations where human judgement must prevail, for example by conditioning reliance on user seniority, the model's confidence level, or the nature of the use case. This particularly reinforces the MRM principle that AI models should augment, not replace, the human judgement required for business-critical decisions, and is consistent with SS1/23's requirement that limitations be understood by users and that approved use remain actively governed (Bank of England, 2023b).

Feedback mechanisms and usage metrics, such as override rates, can help detect whether actual use remains aligned with approved use and model limitations. These signals should inform governance review, retraining decisions, or updates to the model-use policy where proportionate (Bank of England, 2023b). Usage protocols can be supported by recurring user training on output uncertainty, prompt sensitivity, and typical failure modes such as hallucinations or biased outputs, with the depth of training calibrated to user role and decision impact (Crisanto et al., 2024; Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024).

Especially in high-impact use cases, usage must comply with legal constraints, including the GDPR's prohibition on solely automated decisions with significant effects on individuals unless the regulation's specific conditions for such processing are satisfied.

Challenge 11 (C11): Monitoring frequency and indicators are outdated for dynamic AI models

Annual or quarterly reviews may be insufficient for models that learn or adapt more frequently. Model drift, bias drift, or user feedback may signal degradation, but these are often not captured or escalated in time.

Recommendation 11 (R11): Implement continuous monitoring with alert-based triggers

For AI models that learn, adapt, or are embedded in changing deployment contexts, traditional annual or quarterly review cycles may be insufficient, especially where models inform

business-critical or regulated decisions. Monitoring in this context is not only a form of operational performance assurance, but a mechanism for generating ongoing evidence that models remain fit for their approved purpose (Bank of England, 2023b). The core issue is therefore not simply monitoring frequency, but whether monitoring captures the relevant sources of AI-related model risk, including data drift, changing user behaviour, or shifts in the business environment. Monitoring should assess whether the model's limitations, use conditions, and risk profile remain valid over time (ISO/IEC, 2023a; Sayles, 2024).

Consequently, monitoring indicators should be aligned with the model's intended use, materiality, and known limitations. Depending on the use case, relevant indicators may include performance deterioration, output drift, fairness shifts, robustness issues, confidence-score instability, override frequency, user feedback, or evidence of emerging misuse. These indicators should combine quantitative and qualitative information and be interpreted in the relevant business context rather than treated as purely technical metrics. Institutions may define triggers for sustained performance deterioration, recurring drift signals, abnormal override patterns, or repeated user complaints, and calibrate alert thresholds to risk appetite, approved use, baseline behaviour, and materiality (Sayles, 2024). Specific numerical thresholds may be useful in individual implementations, but they should be treated as model- and context-specific rather than generally prescribed. When monitoring signals indicate material deterioration, misuse, or a change in the model's risk profile, the output should feed into recalibration, fallback activation, or model change processes. This ensures that monitoring outcomes are integrated into the MRM lifecycle rather than remaining isolated operational signals (ISO/IEC, 2023a; Sarmah et al., 2024).

Change Management

The change-management weakness concerns the difficulty of assessing materiality in AI-related model changes. Traditional MRM distinguishes between routine maintenance and material model change by focusing on changes to methodology, assumptions, inputs, implementation, or use (European Central Bank, 2025). AI models complicate this distinction because minor technical or configuration changes may materially affect outputs and risk characteristics. The conceptual issue is therefore not that change policies lack AI-specific examples, but that AI makes the materiality of change harder to observe. Change management must therefore assess not only whether the technical model has changed, but whether the model's behaviour, explainability, fairness, or approved-use profile has changed in a way that affects model risk.

Challenge 12 (C12): Change management frameworks do not reflect AI retraining and architecture updates

Traditional model change criteria may be insufficient for AI models, given their dynamic learning nature. Unlike static models, AI models may adapt to new data, leading to shifts in behaviour and performance.

Recommendation 12 (R12): Redefine material changes for AI

AI models, particularly those with frequent retraining or evolving architectures, complicate traditional change management because seemingly technical adjustments may alter model behaviour or risk characteristics. Traditional change policies often focus on input/output modifications or logic rewrites. However, AI-specific changes, such as data source modifications, vendor model updates, or retraining, can materially affect performance, explainability, fairness, or approved use, even without altering the model's explicit function.

According to PRA SS1/23 and EGIM, material model changes, including model updates over time, are subject to re-validation, with the level of validation rigour determined by the materiality of the change (Bank of England, 2023b; European Central Bank, 2025). The EGIM, for example, requires institutions to clearly distinguish between model maintenance and model change, with material changes requiring ex-ante notification, full revalidation, and governance approval (Bennett, 2017). It is also advised to log non-material changes for ex-post reporting (European Central Bank, 2025). This distinction is particularly relevant for AI models, where retraining on updated data or adjustments to hyperparameters may constitute a model change (Deutsche Bundesbank & Federal Financial Supervisory Authority, 2021). Similarly, the EBA (European Banking Authority, 2024) encourages institutions to classify changes not only by technical scope but also by their effects on fairness, explainability, and accountability.

Model risk policies should therefore define AI-relevant change criteria that support proportionate oversight (European Central Bank, 2025). Minor changes may be addressed through documentation and attestation, whereas more material changes, such as retraining with new data, architecture shifts, or external model integration, should trigger second-line review, governance approval, and revalidation where appropriate.

3.5 Case Studies

We illustrate our key recommendations with three representative case illustrations in different contexts: (1) an AI model for market risk, (2) an enterprise chatbot deployment, and (3) an AI use-case discovery at the strategic level. These scenarios show how financial institutions can translate the proposed enhancements into concrete MRM arrangements and align them with evolving regulations. The cases are intended as context-specific illustrations rather than empirical evaluations. Accordingly, the governance mechanisms and thresholds are illustrative and should be calibrated to the institution's risk appetite and context. Their purpose is to show how the taxonomy supports assessable MRM outcomes, such as timely model identification, validation readiness, monitoring exceptions, fallback activation, approved-use breaches, and model change classifications.

1. *Accelerating market risk calculations with AI.* A global bank deployed a neural network model to accelerate the calculation of Value at Risk (VaR) estimates. The AI component was designed to approximate outputs of a slower, resource-intensive simulation model, enabling faster reporting and more responsive risk management.

Observed Weaknesses

- **Inadequate documentation of design rationale (C9):** Decisions regarding the model's training process, input dimensionality reduction, and performance trade-offs were not formally documented, complicating validation and change control.
- **Inadequate monitoring mechanisms (C11):** The model was monitored using traditional alerts through backtesting (VaR exceedances), which failed to capture AI-specific risks such as drift or out-of-distribution behaviour. Furthermore, no fallback solution was implemented.
- **Minimal engagement with validators (C3):** Model risk and compliance stakeholders were engaged only after development, resulting in reactive rather than preventive risk mitigation.

Governance Enhancements

- **Clarifying model documentation practices (R9):** The bank should document the end-to-end development rationale in a standardised, version-controlled model card capturing data sources, feature engineering or dimensionality reduction, architecture and configuration choices, key hyperparameters, training and evaluation procedures, performance results and trade-offs, limitations, and intended use (Mitchell et al.,

2019). Documentation should explicitly justify method appropriateness for the business case and risk context and should support assessment of conceptual soundness, particularly where model logic is less transparent than traditional approaches. Where relevant, explainability techniques can be used to inform conceptual soundness at multiple stages—pre-development (e.g., exploratory data analysis), by-design (interpretable modelling choices), and post-hoc (explanation methods applied after training)—and banks may use a combination of these techniques to mitigate risks from opaque decision-making. To reduce documentation burden and improve reproducibility, evidence capture should be automated where feasible for example through automated generation of training and evaluation artefacts, metric plots, configuration settings, and lineage metadata (J. Liu et al., 2024). Finally, documentation should also be embedded into change control through a lightweight release checkpoint in which a designated reviewer verifies that documentation is complete, current, and consistent with the model's approved use.

Establishing robust monitoring and fallback mechanisms (R11, R12): The bank should implement daily monitoring that goes beyond simple point-estimate thresholds by assessing whether the model continues to operate within its approved market risk use. Monitoring should combine realised risk performance with indicators of input distribution shift, out-of-distribution behaviour, and model stability over time. These metrics should be logged in a transparent set of key performance indicators with targets calibrated to the model's baseline behaviour, validation results, approved use, and risk appetite. A traffic-light scheme can provide a governance-oriented escalation framework distinguishing environmental instability from performance breakdowns. Yellow alerts may signal shifts in the model's operating environment while realised risk performance remains acceptable; they should trigger diagnostics, intensified monitoring, and management documentation. Red alerts may reflect breaches of predefined performance criteria, such as failed backtesting, requiring escalation, activation of an approved fallback, and, where appropriate, controlled retraining and redeployment under governance controls.

- **Engaging oversight early in the development lifecycle (R3):** The bank should involve validation, compliance, or internal audit earlier in the lifecycle, for example during requirements definition and model design, so that acceptance criteria and robust test scenarios are defined early and translated into checks that run alongside development, enabling developers to incorporate required adaptations into the

design and documentation from the outset, such as appropriate safeguards or explainability features. This helps surface emerging risks and compliance gaps sooner without transferring development responsibility to validation or weakening independent challenge. Evidence from adjacent software-engineering domains suggests that early detection of requirements or design deficiencies can reduce downstream remediation effort, providing a rationale for earlier validation and compliance input in AI model development (Haskins et al., 2004). In practice, the bank could use explicit compliance quality gates, including change-review and release-approval checkpoints supported by traditional agile software-development processes.

2. *Governance of an Internal Enterprise Chatbot.* A large financial institution deployed an internal chatbot powered by an LLM to assist employees with daily operational queries, such as policy lookups, procedural guidance, and text summarisation. Although the tool delivered measurable productivity gains (Chen, 2025), it was not initially subjected to formal model identification or MRM screening, as it was classified as an IT productivity tool rather than as an AI-enabled model component embedded within a broader AI system.

Observed Weaknesses

- **Unclear model definition and classification (C4):** The chatbot was excluded from the model inventory because it did not produce quantitative outputs nor directly influence financial decisions. This led to a governance gap despite the model's wide internal use and evolving behaviour.
- **Lack of governance ownership (C2):** Governance was distributed across the IT service team, with no clear model owner, accountability structure, or escalation path. This limited oversight of model updates and performance incidents.
- **Lack of employee training and model use guidelines (C10):** Users were not provided with sufficiently clear guidance on the chatbot's and underlying LLM-based model component's limitations, appropriate use cases, or the circumstances in which outputs required independent verification or human judgment. This increased the risk of misuse and over-reliance on AI-generated responses.

Governance Enhancements

- **Clarifying the model definition for non-traditional AI outputs (R4, R5, R6):** While the underlying GenAI model may not conform to traditional quantitative

model classifications, it can still create model risk through fundamental errors or inappropriate use and should therefore be screened and, where material, subject to proportionate MRM treatment (Bank of England, 2023b), To reduce "edge-case debates", the bank should revise its model definition policy to capture AI-enabled model components that produce non-deterministic or qualitative outputs where these outputs materially affect business-critical, compliance, conduct, control, or risk-relevant processes. Consequently, the bank should identify all models in use in a central inventory. It then should apply a materiality filter, with thresholds calibrated to the bank's risk appetite and the available assurance capacity, to determine which model components warrant more stringent MRM treatment. The surrounding chatbot application, infrastructure, access management, and operational controls should remain primarily subject to broader AI and IT governance, while being cross-referenced where they materially affect model behaviour, reliance, monitoring, or change management.

- **Clarifying ownership and accountability (R2):** The bank should assign a named model owner with explicit end-to-end accountability for the chatbot's LLM across its lifecycle, and formalise this in the MRM governance framework by defining responsibilities via a RACI matrix and a clear escalation path with time-bound responses (European Central Bank, 2025). Building on risk-governance practice, ownership should be embedded in existing structures (e.g., risk committee) rather than left diffuse across teams (Schuett et al., 2025).
- **Training users and defining model-use constraints (R10):** To reduce misuse and over-reliance, the bank should complement the chatbot rollout with mandatory user training, which explicitly covers typical failure modes (e.g., confident but incorrect answers), when the chatbot is appropriate (e.g., policy lookup, drafting, summarisation) versus inappropriate (final decision authority). This is important because users tend to prioritise efficiency and can become anchored to direct chatbot answers—raising reliance and impairing performance when the AI output is wrong (Spatola, 2024). From an MRM perspective, the relevant control objective is to ensure that users rely on the LLM-based model component only within its approved use and known limitations. Disclosures, source-checking requirements, fallback paths, and escalation channels can support this objective and should be understood as model-use governance mechanisms. The surrounding interface design and access management remain part of broader AI and IT governance unless they materially affect model behaviour, approved use, or model-risk monitoring.

3. *Governance of strategic AI initiatives within the bank.* As part of its digital transformation agenda, a mid-sized bank initiated a portfolio of exploratory AI use cases across lending, compliance, and operations. While many initiatives did not yet meet capital materiality thresholds, they raised questions around model inventory inclusion, risk assessment, and validation requirements.

Observed Weaknesses

- **Lack of committee structure (C2):** Exploratory initiatives were often developed in isolation by innovation or business units, with no structured cross-functional review mechanism in place to review scope or alignment with enterprise architecture and MRM standards. The gap is specifically that AI-enabled model components were reaching pilot and production stages without being screened against the institution's model definition, tiering criteria, or validation requirements, leaving material model risk unidentified and ungoverned.
- **Missing AI-specific tiering criteria (C5):** Initial risk assessments focused narrowly on financial impact and model complexity. This failed to capture key AI-specific dimensions such as adaptivity, third-party reliance, or reputational sensitivity, leading to under-classification and deferred validation planning.
- **Lack of development standards (C8):** Developers lacked guidance on minimum expectations for AI model development. This resulted in inconsistent coding standards, versioning gaps, and undocumented model decisions—hindering later-stage validation and risk assessment.
- **Lack of validation readiness (C9):** There was no clear expectation for documentation, testing standards, or validation packages at the pilot stage. As a result, several use cases encountered friction during production onboarding when expected controls and audit trails were missing.

Governance Enhancements

- **Setting up a cross-functional AI model review forum (R2):** To address the fact that exploratory AI initiatives were developed "in isolation", the bank should introduce a formal cross-functional review forum connected to the existing model governance framework. Its purpose should not be general innovation steering alone, but early identification of AI-enabled model components, preliminary model-risk classification, and alignment with MRM standards before pilots scale into

production (U.S. Department of the Treasury, 2024). The forum should include business ownership, MRM, IT, and compliance so that model definition, data use, build-versus-buy considerations, intended users, and validation expectations are considered at an early stage (Schuett et al., 2025). This is consistent with prudential expectations that model governance should define roles, responsibilities, approval processes, and effective challenge across the model lifecycle (Bank of England, 2023b).

- **Revising model risk tiering criteria for AI-specific risk dimensions (R5):** Following an updated model definition, the bank should revise its model risk tiering so that exploratory AI use cases are assessed not only by financial impact and generic complexity, but also by AI-specific risk dimensions such as adaptivity, opacity, third-party reliance, as well as reputational or compliance sensitivity (Bank of England, 2023b; Financial Stability Board, 2024; Office of the Comptroller of the Currency, 2021). Relevant regulatory classifications, including the EU AIA where applicable, should be considered to ensure consistency with applicable obligations. The tiering framework should then determine the proportional scope of subsequent MRM requirements, including validation, monitoring, documentation, and approval expectations. To avoid incentives for under-registration, initial inventory capture should be mandatory and lightweight, with documentation and validation depth determined only after tiering. This preserves proportionality while ensuring that higher-risk initiatives trigger earlier oversight and structured validation planning before scale-up. Thresholds may be calibrated to risk appetite and available assurance capacity, but capacity constraints should be addressed through prioritisation or staged assurance rather than by understating inherent model risk.
- **Improving development and validation standards (R8, R9):** To prevent pilot-to-production friction, the bank should define minimum validation-readiness expectations for AI pilots that may become model-relevant. During the pilot, teams should maintain a versioned model card or decision log covering architecture, data sources, key assumptions, intended use, design trade-offs, limitations, and initial testing evidence. This documentation should be produced concurrently with development so that validation evidence does not need to be reconstructed at the production-onboarding stage (Winecoff & Bogen, 2025). Before production deployment, higher-risk AI or GenAI use cases should be supported by an evaluation protocol with ex-ante baselines and release criteria aligned to the model's intended use and risk tolerance. Depending on the use case, this may include tests for

hallucination, toxicity or bias, robustness, and output stability. The same tests should be repeated after material model changes, such as retraining, changes to retrieval knowledge bases, or material prompt-template changes, to support ex-ante and ex-post comparison. These artefacts should provide evidence for conceptual soundness, validation planning, monitoring design, and subsequent change assessment (Huang et al., 2025; National Institute of Standards and Technology, 2024).

3.6 Conclusion

This paper analysed the limitations of established MRM frameworks when applied to AI models in financial institutions. Methodologically, it draws on a systematic synthesis of relevant MRM and AI governance literature and regulatory and standard-setting artefacts, triangulated with recent practitioner sources, complemented by a targeted literature review for each weakness, and illustrated through representative case illustrations.

This work proceeds in two steps. First, using current MRM frameworks as a baseline and comparing them with relevant AI governance frameworks, we identify twelve key weaknesses prevalent in current practice. Many of these weaknesses reflect legacy assumptions rooted in traditional quantitative finance and established model governance expectations, and we organise them across MRM lifecycle domains. Second, we derive governance enhancements to address the issues identified. By analysing the weaknesses through the lens of AI-specific challenges, we develop a lifecycle-oriented mapping from observable governance weaknesses to targeted enhancements and show how AI-specific properties affect established MRM concepts across the model lifecycle. In doing so, the examination provides a structured basis for comparative analysis across AI use cases in financial services while remaining anchored in the internal logic of prudential MRM. This yields a taxonomy of twelve weaknesses and an explicit mapping to governance enhancements, which can be further operationalised and evaluated in subsequent empirical work. Accordingly, the recommendations are intended to support proportionate adaptation of MRM governance processes across institutions with different levels of AI maturity.

To demonstrate applicability across contexts, we apply the framework through three representative use cases: (1) an AI model for market risk, (2) an enterprise chatbot deployment, and (3) strategic AI use-case discovery. These case illustrations show how the proposed governance enhancements can be interpreted in distinct AI deployment contexts and how

regulatory and supervisory expectations can be interpreted through lifecycle-level MRM concepts in each setting.

By linking high-level regulatory principles to lifecycle-level MRM concepts, the framework provides a structured answer to the paper's central questions by specifying recurring governance weaknesses and corresponding governance enhancements across the model lifecycle. At the same time, the taxonomy and mappings are derived from a synthesis of existing literature and regulatory and standard-setting artefacts and should therefore be further operationalised and tested in empirical settings.

Future research should also establish risk-appetite-aligned and statistically calibrated thresholds for key monitoring metrics such as performance degradation, fairness shifts, and drift indicators so that material changes can be linked to consistently redevelopment and enhanced oversight.

4 Inherent Model Risk Assessment for AI-Based Models: Integrating AI Risk Drivers into Model Risk Management

4.1 Introduction

The increasing reliance on quantitative models in financial institutions has led to the establishment of structured model risk management (MRM) frameworks that identify, assess, and mitigate risks arising from model use. Recent regulatory expectations illustrated in SR 26-2 (Board of Governors of the Federal Reserve System, 2026) and PRA SS1/23 (Bank of England, 2023b) require more risk-sensitive model governance. Central to these expectations is the concept of inherent model risk, i.e., the risk of adverse consequences arising from incorrect or misused model outputs before mitigating controls. Financial institutions operationalise this concept through model risk tiering, classifying models by relative risk to guide validation and governance. Such approaches use scorecards or decision trees, combining severity and probability of model failure, as reflected in regulatory guidance for MRM, including SR 26-2 and PRA SS1/23.

The expanding use of artificial intelligence (AI), including machine learning (ML) and generative AI (GenAI), complicates this tiering logic. AI-based models may exhibit dynamic behaviour due to data drift, retraining, or context-dependent interactions (Bayram et al., 2022; X. Zhang et al., 2022). Important risk drivers may not be directly observable because they may be embedded in opaque model structures or high-dimensional interactions. Thus, AI-related risks are difficult to integrate consistently into existing model risk tiering without translation and calibration. Although SR 26-2 excludes GenAI and agentic AI from scope, it still expects banks to establish appropriate governance for them.

AI risk literature has therefore proposed methods for assessing such risks. These approaches typically adopt a system- or use-case-level perspective, focusing on fairness, transparency, robustness, and societal impact. Similarly, regulatory initiatives such as the European Union's (EU) AI Act (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024) classify AI systems by intended use and potential external impact. Yet these developments leave unresolved how such risks can be represented at the model level.

This mismatch reflects a deeper inconsistency in risk assessment logic. Traditional tiering assumes that risk profiles can be assessed ex ante and compared transparently within a common inventory, conditions that suit many statistical models but are harder to maintain for AI-based models shaped by dynamic behaviour or black-box characteristics.

Deloitte's (2025a) EMEA MRM Survey of 87 banks underlines the practical relevance of this translation problem: only 29% regarded their MRM framework as fully adequate for AI/ML governance, while 67% considered it adequate only to some extent.

Recent work has begun to address AI-related model risk within MRM-specific settings. Prorokowski (2025), for example, develops a model risk rating concept for ML in credit risk, drawing primarily on bank validation practices and focusing on model-life-cycle and model-quality assessment. However, the broader problem remains unresolved: how can AI-specific and partly system-level risk drivers be translated into inherent-risk classifications that fit established MRM tiering frameworks without creating a parallel AI taxonomy?

In answering this question, this paper makes three contributions. First, it identifies and structures the AI-specific and partly system-level risk manifestations that are relevant to inherent model risk assessment. Drawing on responsible-AI principles, it distinguishes risks that affect the consequences of model failure from those that affect the probability of model failure or assessability. Second, it proposes a translation logic that maps AI-specific risk drivers onto the established materiality and complexity dimensions of model risk tiering. The contribution is therefore not a new AI-risk taxonomy, nor a replacement for existing model tiering concepts, but a mechanism for translating AI-related risk manifestations into existing tiering logic. In contrast to AI assessment frameworks, which primarily classify or assess AI systems by application context (Schmitz et al., 2025), the proposed logic specifies how such concerns affect the severity of potential consequences, the probability of model failure, or assessability. Third, it illustrates how this logic can be applied across traditional statistical, classical ML, and GenAI models.

The paper proceeds as follows. Section 2 reviews the literature, Section 3 explains the methodology, Section 4 analyses AI-specific risk drivers, Section 5 develops the framework, Section 6 illustrates its application, and Section 7 concludes.

4.2 Literature Review

The literature on model risk tiering emphasises that tiering does not quantify absolute model risk precisely, but supports relative prioritisation across a portfolio of models (Bernard et al., 2020; Brotcke, 2020; Hassani, 2019; Hill, 2019). Existing approaches typically rely on scorecards, decision rules, or hybrid methods that assess inherent risk through dimensions such as materiality and complexity (Dionne & Howard, 2017; Keenan, 2015; Kiritz et al., 2019; Mankotia & Joshi, 2013). Prorokowski (2025) extends this logic to ML model risk rating in credit risk and highlights related validation challenges.

This tiering logic depends on three operational conditions. First, model risk must be assessable at a defined point in time. Second, the main drivers of inherent risk must be identifiable, observable, or approximable through documented proxies. Third, models must be comparable within a common inventory-level framework (Kiritz et al., 2019). These conditions are generally suitable for many traditional statistical and econometric models, where model structure, inputs, assumptions, and use context are comparatively stable and transparent. They become more difficult to maintain for AI-based models, whose risk profile may be shaped by opaque representations and sensitivity to environmental change (X. Zhang et al., 2022; X. Zhang et al., 2025).

As a result, AI risk assessment literature addresses these issues from a different starting point. One stream consists of regulatory and scenario-based approaches, including Novelli et al.(2024) and the EU AI Act, which classify risk primarily at the level of AI systems or use cases. Their strength lies in linking AI risk to intended purpose, affected stakeholders, and legal obligations. A second stream consists of responsible AI tools, such as question banks and impact assessments, which aim to operationalise principles including fairness, transparency, robustness, security, privacy, and human rights (Kazim & Koshiyama, 2021; Lee et al., 2026; Madaio et al., 2020; Mantelero, 2018; Shelley, 2022). A third stream seeks to quantify AI risks through measurable indicators for accuracy, fairness, explainability, sustainability, transparency, or related dimensions (Bommasani et al., 2023; Giudici, Centurelli, et al., 2024). Adjacent AI governance and risk-management frameworks, including implementation-oriented risk taxonomies and control frameworks, also provide useful structures for identifying and mitigating AI-related risks (Giudici & Kolesnikov, 2026; Giudici & Raffinetti, 2023; National Institute of Standards and Technology, 2023; Prem, 2023).

Together, these approaches aim to identify and measure AI-related risks but do not provide a general logic for translating AI-specific and system-level concerns into inherent-risk tiers

within a unified model inventory. In particular, they do not specify when concerns such as fairness, robustness, or explainability should affect materiality, complexity, or both. Related comparative analyses of AI assessment frameworks and risk aggregation schemes further show that existing approaches tend to offer high-level, qualitative guidance on how to weigh and combine different AI risk dimensions, rather than prescribing universally applicable aggregation rules that could be aligned directly with model-level tiering in banking (Schmitz et al., 2025).

The literature consequently reveals a mismatch between two assessment logics: MRM tiering operates at the model level and supports comparison across a model inventory, whereas AI risk frameworks typically assess multidimensional, context-dependent, and system-level risks. This paper therefore focuses on how AI-specific and partly system-level drivers can be represented within inherent model-risk tiering across traditional statistical, classical ML, and GenAI models.

4.3 Methodology

This study adopts a qualitative integrative review approach designed to synthesise insights from three bodies of work: model risk tiering and supervisory MRM guidance, AI risk assessment and responsible-AI frameworks, and technical literature on AI-specific failure modes. An integrative review is appropriate because the study synthesises heterogeneous sources for conceptual framework development (Cronin & George, 2023). The structured literature search was guided by the research question and served as the basis for developing the translation logic proposed in the framework.

Relevant literature was retrieved from the Scopus database. Scopus was selected for its broad indexing of conference proceedings and technical publications (Singh et al., 2021). The search was restricted to the following Scopus subject areas: Computer Science; Engineering; Social Sciences; Business, Management and Accounting; Mathematics; Decision Sciences; and Economics, Econometrics, and Finance. Following the PRISMA framework, 11 relevant peer-reviewed MRM tiering articles (Figure 4.1) and 149 AI risk and responsible-AI publications were retained (Figure 4.2). The detailed search strings and the inclusion and exclusion criteria are specified in Figures 4.1 and 4.2. In addition, 10 banking-specific MRM guidance documents and 6 responsible AI frameworks and standards were selected purposively for their governance relevance, including SR 26-2, PRA SS1/23, the NIST AI Risk Management Framework (NIST AI RMF) (National Institute of Standards and Technology, 2023), and the EU AI Act.

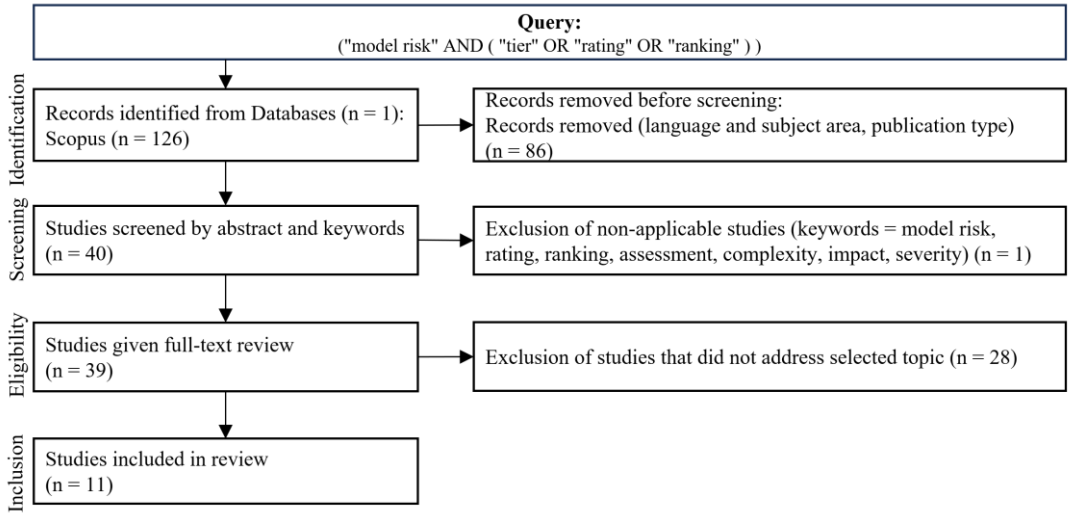

Figure 4.1: PRISMA flow diagram for MRM tiering literature, including search query, eligibility criteria, exclusion reasons, and final retained sample.

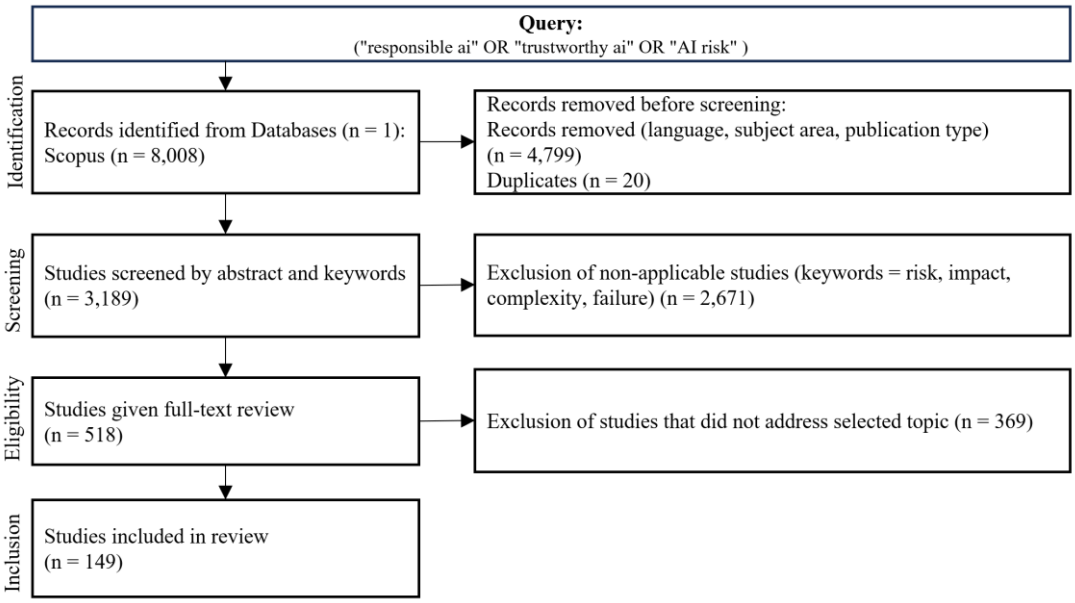

Figure 4.2: PRISMA flow diagram for AI risk and responsible-AI literature, including search query, screening criteria, exclusion reasons, and final retained sample.

The synthesis followed three steps. First, the MRM tiering literature and supervisory guidance were used to determine the supervisory expectations and define the materiality- and complexity-based classification logic of inherent model risk assessment. Second, AI-specific risk manifestations were identified from the retained literature and grouped using responsible-AI principles (transparency and explainability; technical robustness and safety; privacy and data governance; diversity, non-discrimination, and fairness; human autonomy and oversight; societal and environmental well-being; accountability) as a structured grouping criterion. The principles serve as a grouping device because a convergent set of principles recurs across responsible AI frameworks, with the EU AI Act providing representative operationalisations (Papagiannidis et al., 2025). Third, the grouped risk manifestations were mapped to materiality,

complexity, or both. A manifestation was coded as materiality-relevant where it affected consequence severity, and as complexity-relevant where it affected the probability of model failure or assessability. The synthesis was plausibility-checked through triangulation across academic literature, supervisory guidance, and governance standards.

4.4 AI-specific risk drivers of inherent model risk

Building on the integrative review and the responsible-AI grouping described above, this section identifies AI-specific risk manifestations that are relevant to inherent model risk assessment. The analysis focuses on responsible-AI principles with a direct inherent-risk channel: transparency and explainability; technical robustness and safety; privacy and data governance; diversity, non-discrimination, and fairness; and human autonomy and oversight. Accountability as well as societal and environmental well-being are therefore not treated as standalone risk-driver groups because they primarily concern broader governance or ethical questions.

Transparency and explainability captures the extent to which relevant model behaviour, failure drivers, and output-generating mechanisms can be understood, challenged, and evidenced. In traditional statistical models, many assumptions, parameters, and diagnostics can be inspected directly. In AI-based models, by contrast, relevant risk drivers may be embedded in latent representations, high-dimensional feature interactions, or proxy targets that are not directly visible. Obermeyer et al. (2019), for example, show that substantial bias may arise from the choice of proxy target rather than conventional predictive performance. The explainability literature reinforces this concern by cautioning against treating post-hoc explanations as direct evidence of model understanding, since such explanations may not faithfully represent the model's actual decision logic (Rudin, 2019). Where opacity prevents affected stakeholders, validators, or supervisors from identifying or remediating harmful outcomes, it may also create legal, conduct, or reputational consequences (Kazim & Koshiyama, 2021).

Technical robustness and safety captures the stability of model behaviour across data regimes, operating environments, and deployment contexts. In AI-based models, however, robustness concerns may become harder to assess because behaviour can be sensitive to distribution shift, retraining choices, or underspecification. Prior work on concept drift, underspecification, and shortcut learning shows that validation performance may not reliably predict deployment behaviour under changed conditions (D'Amour et al., 2022; Gama et al., 2014; Geirhos et al., 2020).

Privacy and data governance captures risks arising from sensitive data use, data leakage, and model access. Privacy risks may not be visible through conventional performance testing, because they can emerge through interaction with the trained model or deployed interface. Membership inference and training-data extraction illustrate how risks may become visible only through query behaviour or model access rather than through standard offline validation (Carlini et al., 2021; Shokri et al., 2017). These risks are especially relevant where models process personal, confidential, or commercially sensitive data.

Diversity, Non-Discrimination, and Fairness captures risks arising from biased data, proxy variables, or target construction effects that may produce discriminatory or unjustified differences in outcomes. In AI-based models, such concerns may become harder to detect and govern than in traditional models because bias can be embedded in high-dimensional feature interactions and non-linear relationships. These concerns are particularly relevant in credit-scoring and other financial ML applications (Hurlin et al., 2026; N. Zhou et al., 2022). This interpretation is consistent with AI governance literature that treats fairness, privacy, and fundamental-rights implications as context-dependent consequences rather than isolated technical properties (Kazim & Koshiyama, 2021; Lee et al., 2026; Mantelero, 2018).

Human autonomy and oversight captures the extent to which users remain able to understand, contest, or override model outputs before reliance on them. This is not a property of the model artefact itself but of the decision pathway in which the model is embedded. AI-based decision pathways may create automation bias, over-reliance, or reduced contestability, especially where outputs are presented as authoritative or embedded in automated workflows (Klingbeil et al., 2024). Where model outputs drive consequential decisions with limited opportunity for review or override, the ability to intercept, contest, or remediate adverse outcomes is reduced (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024; Office of the Comptroller of the Currency, 2021).

Finally, system-level distribution of failure drivers cuts across these principles. Relevant failures may arise not only inside the model artefact, but also through data pipelines, third-party services, or feedback loops. Prior work on hidden technical debt, upstream data cascades, and long-run feedback effects shows that AI failures may propagate through organisational systems rather than remain confined to a single model component (Kondor et al., 2024; Sambasivan et al., 2021; Sculley et al., 2015). For MRM, this does not imply abandoning the model as the unit of governance. Rather, system-level drivers become relevant where they affect the model-in-use and can subsequently be translated into model-level tiering criteria.

Across these manifestations, AI changes inherent model risk not because every risk is novel, but because opacity, data dependence, adaptive behaviour, and system embedding can make failures harder to anticipate and detect.

4.5 Translation Framework for AI-Specific Drivers in Model Risk Tiering

The preceding section identifies AI-specific manifestations of model risk. For MRM purposes, however, these manifestations become operational only once translated into the established tiering dimensions of materiality and complexity.

The scope for the proposed framework is the model-in-use: a model artefact or model component that meets the institution's model definition and is applied to a specific business purpose. The same artefact should be assessed separately where its business purpose differs materially, as the same architecture may generate different inherent-risk profiles across use cases (Bank of England, 2023b; Board of Governors of the Federal Reserve System, 2026).

The framework is intended for the inherent-risk assessment phase of MRM and is designed as an MRM-compatible assessment logic for traditional statistical, classical ML, and GenAI models. It uses a structured scorecard rather than a purely mechanical decision tree, with standardised questions and classification criteria disciplining expert judgement and making tier assignments reproducible and auditable. Scorecards preserve transparency and adaptability, whereas purely mechanical decision trees may oversimplify multidimensional risk relationships or produce undesirable tier clustering (Bernard et al., 2020; Dionne & Howard, 2017; Kiritz et al., 2019). Rather than estimating model risk with numerical precision, the framework supports relative ranking within a model inventory for proportionate validation and governance (Bank of England, 2023b; Board of Governors of the Federal Reserve System, 2026).

The framework preserves the established materiality-complexity structure of model risk tiering while making AI-specific and system-level risk drivers traceable within that structure. Materiality captures the severity, legal relevance, stakeholder impact, irreversibility, and propagation of adverse consequences. Complexity captures the probability of model failure and assessability: the extent to which model structure, data, behaviour, and system embedding make failure more likely or harder to detect, validate, or challenge. Weak observability or limited challengeability increase the probability that failure remains undetected before outputs affect decisions.

The translation rule is directional rather than mechanical. A risk manifestation is materiality-relevant where the adverse outcome of model failure or misuse takes the form of economic loss, legal or regulatory exposure, customer or stakeholder harm, rights-related impact, or scalable adverse consequences. It is complexity-relevant where it affects the probability of model failure or assessability. The same manifestation may affect both dimensions, but only where both channels can be separately explained.

4.5.1 Materiality

In traditional MRM, materiality is typically proxied through direct financial exposure, a model's role in core prudential processes or major business-critical decisions (Bank of England, 2023b; Dionne & Howard, 2017; Kiritz et al., 2019). These indicators remain central, since failures in prudentially relevant models may generate significant losses, capital misstatements, or supervisory consequences. However, a framework covering both traditional and AI-based models cannot confine materiality to direct economic or prudential exposure alone. As model use expands into customer-facing, operational, and decision-support contexts, model failure may also become material through legal, conduct-related, reputational, stakeholder, or propagation effects (Bank of England, 2023b).

Materiality is therefore defined as the extent to which model failure or misuse may affect the institution, its customers, markets, or other stakeholders through economic, regulatory, legal, operational, or reputational consequences. Portfolio and prudential exposure remain primary drivers, but materiality also reflects business criticality, regulatory relevance, stakeholder and customer impact, decision autonomy, and the potential for adverse consequences to scale. A model may therefore have limited direct balance-sheet impact but still be highly material if it affects many individuals, operates in a legally sensitive domain, produces customer-impacting outputs with limited human oversight, or is reused across multiple processes.

Fairness-related harm, privacy exposure, and legal obligations are not separate model risk dimensions, but impact channels that increase the consequences of model failure and are reflected within materiality. In credit scoring and financial ML, fairness- and bias-related failures may generate customer loss, legal liabilities, and reputational damage (Hurlin et al., 2026; N. Zhou et al., 2022). Similarly, transparency and robustness become materiality-relevant where weak explainability, unreliable outputs, or misuse create legal, conduct, reputational, remediation, or stakeholder consequences (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024).

Decision autonomy is treated as a modifier because, as a feature of the model-in-use decision pathway, it affects how directly and reversibly model outputs can produce adverse consequences. In application, it may operate as an upward modifier or floor rule within materiality. As a floor rule, high decision autonomy should prevent assignment to a low materiality band where model outputs directly affect customers, legal obligations, or critical processes without meaningful human review or timely override.

Regulatory AI classifications reinforce this logic. The EU AI Act, for example, classifies AI systems according to the legal significance of intended use and imposes more demanding obligations for high-risk applications (Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024; Leinarte, 2024; Schuett, 2024). Such classifications inform regulatory and legal materiality because they signal heightened supervisory scrutiny, compliance obligations, and potential conduct consequences. They should not operate as a parallel AI taxonomy or mechanical tiering override, but as evidence of the legal and regulatory consequences associated with the model-in-use.

Another materiality channel concerns impact scale. AI-based models may create harms beyond the immediate output where they are reused across processes, embedded in automated workflows, or adopted in similar form across institutions. AI adoption in financial markets may contribute to homogenisation of risk-management techniques and amplify procyclicality, while system-level feedback effects and hidden technical dependencies may cause errors to propagate through organisational processes rather than remain confined to a single use case (Daníelsson et al., 2022; Kondor et al., 2024; Sculley et al., 2015). Within the present framework, such effects are treated as materiality drivers where model failure may scale or create correlated adverse consequences across portfolios, customers, business lines, or markets.

The materiality assessment therefore retains four materiality sub-dimensions, which are summarised in Table 4.1: economic and business impact, legal and regulatory obligations, customer, stakeholder, and rights impact, and impact scale and propagation, with decision autonomy treated as a modifier within the relevant sub-dimensions.

4.5.2 Complexity

Complexity is operationalised as a structured proxy for the probability of model failure and for assessability, not as a measure of technical sophistication alone. It captures characteristics of the model-in-use that make failure more likely, less observable, or harder to detect or challenge.

The risk manifestations identified above are translated into five complexity sub-dimensions, which are summarised in Table 4.1: data and representation complexity, explainability and observability constraints, behavioural instability and change risk, model-form complexity, and implementation and system dependence.

First, data and representation complexity captures the difficulty introduced by the data, target, and representation layer. In AI-based models, however, relevant failure drivers may be embedded in high-dimensional, unstructured, multimodal, or latent representations that are harder to audit. They may also arise from feature construction, proxy targets, training-data composition, or transformations that are not visible in conventional performance metrics. Obermeyer et al. (2019), for example, illustrate how a proxy-target choice can embed risk in the modelling problem itself rather than in observed predictive performance. Data governance considerations therefore become complexity-relevant where data structure, modality, lineage, or access restrictions limit validation and effective challenge (Carlini et al., 2021; Obermeyer et al., 2019; Shokri et al., 2017).

Second, explainability and observability constraints capture the difficulty of understanding, evidencing, and challenging model behaviour. In many traditional statistical settings, important model characteristics and the functional relationship between inputs and outputs can be more readily inspected through the explicit model structure and associated parameter estimates. In AI-based models, however, key failure drivers may not be directly visible. Post-hoc explanations can support challenge, but should not be treated as direct evidence of model understanding where their fidelity to the underlying model behaviour is uncertain (Deutsche Bundesbank & Federal Financial Supervisory Authority, 2021; Rudin, 2019). Opacity therefore increases complexity where it weakens the institution's ability to infer how, when, and why the model may fail.

Third, behavioural instability and change risk captures the difficulty of relying on a stable risk profile over time and across operating conditions. AI-based models may intensify this problem where behaviour is sensitive to distribution shift, retraining, recalibration, prompt design, spurious correlations, or changing deployment context (D'Amour et al., 2022; Gama et al., 2014; Geirhos et al., 2020). Behavioural instability therefore increases complexity through two related channels: it raises the probability that model performance will deteriorate under changed conditions and weakens the extent to which validation evidence obtained at one point in time remains informative for subsequent use (Bayram et al., 2022). Evidence on dataset shift and predictive uncertainty similarly shows that in-domain performance may not reliably indicate out-of-domain behaviour (Ovadia et al., 2019). GenAI applications add further difficulty

through hallucination and prompt sensitivity (National Institute of Standards and Technology, 2024).

Fourth, model-form complexity captures the extent to which the modelling approach makes conceptual soundness and generalisation more difficult. Although traditional statistical models can also be misspecified, their assumptions, parameters, and diagnostics are typically well established. For model classes with which an institution has limited operational experience, misspecification risk is higher, including through hyperparameter misconfiguration or inappropriate functional-form choices, which may impair generalisation and lead to unreliable performance in production (D'Amour et al., 2022; L. Yang & Shami, 2020). This risk is especially relevant for AI-based models, whose non-linearity, high parameterisation, or deep architectures can limit interpretability and conceptual soundness (Deutsche Bundesbank & Federal Financial Supervisory Authority, 2021; Rudin, 2019).

Fifth, implementation and system dependence captures complexity arising from the model-in-use's embedding in technical and organisational systems. AI-based models may depend on data pipelines, upstream models, or third-party services. These dependencies increase complexity where they make failure harder to localise, validate, or challenge.

Table 4.1: Indicative sub-dimensions and classification criteria for inherent model risk assessment.

Dimension	Sub-dimension	Low	Medium	High
Materiality	Economic and business impact	Limited exposure; non-critical use	Meaningful business or risk impact	Large portfolio or material economic exposure
Materiality	Regulatory and legal exposure	Limited legal or regulatory relevance	Some legal, regulatory, conduct, or compliance consequences	Legally sensitive, high-obligation, customer-rights, or supervisory-relevant use
Materiality	Stakeholder and customer impact	Internal or limited stakeholder effect	Meaningful stakeholder or customer effect	Significant customer, conduct, or rights impact
Materiality	Impact scale and propagation	Isolated failure consequences	Cross-process but containable effects	Scalable, correlated, or institution-wide effects
Complexity	Model-form complexity	Transparent and established form	Non-linear or less directly interpretable form, but with established validation methods	Opaque, novel, or highly parameterised form
Complexity	Data and representation complexity	Structured and auditable data	Non-trivial feature engineering, proxy targets, or transformations with partial documentation	Unstructured, multimodal, or latent representations with limited auditability

Complexity	Explainability and observability constraints	Directly inspectable behaviour	Post-hoc explanation required but available for challenge and monitoring	Limited explanation or weak challengeability
Complexity	Behavioural instability and change risk	Stable and well monitored	Drift, retraining, or regime sensitivity that can be monitored through established procedures	Unstable, context-sensitive, or externally updated
Complexity	Implementation and system dependence	Standalone	Several internal dependencies with documented interfaces and ownership	External dependencies with limited transparency or ownership

The criteria in Table 4.1 are indicative and should be calibrated to each institution's model inventory and risk appetite. Where a risk driver affects both materiality and complexity, the rating rationale should explain separately its impact on consequences and on the probability of model failure or assessability.

4.5.3 Aggregation and Calibration

Aggregation proceeds in two stages. First, the relevant materiality and complexity sub-dimensions are assessed using ordinal classification criteria. Equal weighting may be used as an initial baseline, while differentiated weights may be applied where supervisory priorities, risk appetite, or inventory characteristics justify greater emphasis on particular drivers (Bank of England, 2023b). Institutions should therefore test whether tier outcomes are robust to alternative weighting schemes. Because the labels remain ordinal, any aggregation should be interpreted as a relative governance ranking rather than a cardinal measure of risk distance between models (Krisper, 2021). The main sub-dimension ratings, supporting evidence, and reasons why each driver affects materiality, complexity, or both should be documented. Floor rules or overrides may be applied where a critical driver is not adequately reflected in the aggregate assessment, such as severe reputational risk or extensive third-party dependence.

Second, the materiality and complexity bands are combined through a predefined materiality-complexity matrix to determine the final inherent-risk tier (Table 4.2). This is consistent with qualitative risk-matrix practice, where likelihood and consequence bands are translated into discrete action or prioritisation categories (Rollason et al., 2010).

Materiality is high where failure may affect core prudential processes, large portfolios, legally sensitive decisions, significant customer groups, or processes in which adverse consequences may scale or propagate. It is medium where consequences are meaningful but limited in scale

or institutional relevance, or remain largely reversible. It is low where material financial, prudential, legal, customer, or reputational consequences are unlikely. Decision autonomy is treated as a modifier where outputs drive consequential decisions with limited opportunity for human intervention, challenge, or override.

Complexity is high where failure is likely, assessability is materially constrained, multiple complexity drivers are high, or several medium drivers interact so that deployment behaviour cannot be inferred reliably from available evidence. It is medium where the probability of model failure, assessability constraints, or their combined effect create non-trivial governance difficulty, but credible validation, monitoring, and challenge remain possible. It is low where the probability of model failure is low, governance difficulty across sub-dimensions is low, and deployment behaviour can be inferred reliably from transparent model form, data, and monitoring evidence.

In both dimensions, the assigned band should be justified in a short rating rationale that identifies the main drivers, explains relevant interactions between sub-dimensions, and documents the role of weights, expert judgement, floor rules, or overrides. The purpose is not to eliminate judgement, but to discipline it through a transparent aggregation procedure that remains reproducible and open to independent challenge (Bank of England, 2023b).

Table 4.2: Indicative Materiality–Complexity Matrix for Inherent Model Risk Tiering, with Tier 1 representing the highest governance priority.

Materiality / Complexity	Low complexity	Medium complexity	High complexity
High materiality	Tier 2	Tier 2	Tier 1
Medium materiality	Tier 3	Tier 2	Tier 2
Low materiality	Tier 4	Tier 3	Tier 3

Table 4.2 provides an illustrative implementation of the proposed translation logic rather than a universal tiering matrix. The use of three materiality and complexity bands and four final tiers is a parsimonious design choice intended to preserve tractability, auditability, and consistency with the ordinal nature of inherent-risk assessments. It does not imply that model risk has a natural three-by-four structure or that the distance between tiers is quantitatively meaningful. The matrix reflects two assumptions: materiality should generally set a floor for governance attention, while complexity should intensify the tier where opacity, instability, system dependence, or limited assessability make failure harder to anticipate, detect, or challenge (Bank of England, 2023b).

Calibration should support a plausible relative ordering of the model inventory and a risk-sensitive allocation of validation and governance resources. Institutions may revise

classification criteria, thresholds, weights, or override rules where validation experience, supervisory priorities, risk appetite, or inventory characteristics indicate insufficient risk differentiation (Bank of England, 2023b; Board of Governors of the Federal Reserve System, 2011; Office of the Superintendent of Financial Institutions, 2025). Where common AI techniques, shared vendors, or repeated components create correlated risk, institutions should also consider whether individual model-level tier assignments adequately capture portfolio-level concentrations (Bank of England, 2023b).

Finally, tier assignments should be revisited when material changes affect the model-in-use. Consistent with European Banking Authority (2024) and Deutsche Bundesbank and Federal Financial Supervisory Authority (2021), reassessment triggers include material retraining, data drift, changes in decision autonomy, provider-side model updates, new external dependencies, or newly identified fairness, privacy, robustness, or explainability concerns. These triggers do not turn the framework into a continuous monitoring system. Rather, they define when the conditions underlying the original inherent-risk tier need to be reassessed.

4.6 Use Cases

The following use cases provide a conceptual illustration rather than empirical validation. They are structured to isolate different dimensions of the framework: the first two hold use context and materiality broadly constant while varying model form and complexity, whereas the third holds model architecture constant while varying deployment context.

4.6.1 Case 1: Creditworthiness

Context: The first case compares a logistic regression model and an XGBoost model applied to automated creditworthiness assessment in the same portfolio, decision pathway, and governance context of an online bank.

Materiality assessment: Materiality is high for both models. Credit underwriting affects credit exposure, consumer protection obligations, and potential discrimination or conduct risk. As AI-based creditworthiness assessment of natural persons is classified as high-risk under the EU AI Act, this reinforces high regulatory and legal materiality. Decision autonomy further increases materiality where approvals or rejections occur without meaningful human review.

Complexity assessment: Logistic regression has low complexity in this case because its functional form, coefficients, assumptions, and monitoring indicators are transparent and well understood. XGBoost has medium complexity because its ensemble structure, non-linear

interactions, possible proxy relationships, reliance on post-hoc explanations, and sensitivity to data perturbations make behaviour less directly inspectable.

Tier assignment and interpretation: The resulting tier is Tier 2 for both models. Logistic regression combines high materiality with low complexity, while XGBoost combines high materiality with medium complexity. The case shows that a more complex ML model does not automatically require a higher tier where complexity remains below the high-complexity threshold, although it may require greater validation attention to explainability, fairness, calibration, and drift.

4.6.2 Case 2: Value at Risk

Context: The second case compares a GARCH model and an LSTM-based model used to estimate one-day Value-at-Risk for the same material market-risk portfolio. Both models support the same prudentially relevant process and risk-management workflow.

Materiality assessment: Materiality is high for both models because they support a core market-risk measure for a material portfolio and may affect capital, risk reporting, limits, portfolio management, or supervisory processes. Expert review may reduce decision autonomy, but does not remove materiality.

Complexity assessment: GARCH is assigned medium complexity. Its model form is transparent and well established, and its behaviour can be inspected through parameters, diagnostics, backtesting, and stress testing. However, regime sensitivity, structural breaks, and latent volatility create non-trivial validation requirements. The LSTM model has high complexity because its recurrent architecture learns opaque temporal representations, lacks a direct analogue to parameter interpretation, and may be difficult to validate under stressed regimes with limited observations. Training, tuning, monitoring, and implementation may also require specialised infrastructure.

Tier assignment and interpretation: The resulting tier is Tier 2 for GARCH and Tier 1 for the LSTM model. Tier 1 is triggered not by neural-network architecture alone, but by the combination of high materiality and high complexity in a prudential context. The LSTM assignment is driven by opacity, validation difficulty, behavioural instability under stress, and implementation burden rather than by the label “AI” alone.

4.6.3 Case 3: GenAI assistant

Context: The third case compares two GenAI assistants that both rely on the same foundation model but differ in deployment context.

Materiality assessment: Materiality differs by deployment context. The internal assistant supports employees in interpreting internal policies and procedures. It is advisory, not authoritative; it does not feed capital or regulatory reporting processes; and employees remain responsible for verification and decision-making. Under these assumptions, materiality is medium: incorrect outputs may affect internal processes and scale across teams, but do not directly determine customer eligibility or legally binding outcomes. Materiality would increase if the assistant became enterprise-wide, authoritative, customer-impacting, or embedded in regulated decision processes.

The customer-facing assistant has high materiality. It interacts with retail customers and may create conduct, transparency, consumer protection, privacy, remediation, complaint, or reputational consequences. Stakeholder impact and propagation risk are high where misleading outputs can affect many customers at scale, especially where outputs are presented as authoritative or relied on in customer-impacting interactions.

Complexity assessment: Complexity is high in both deployments because the effective output depends on the interaction between the foundation model, the input context, prompt design, deployment configuration, and possible provider-side updates. Key drivers include hallucination risk, prompt sensitivity, limited explainability, and unstable validation boundaries.

Tier assignment and interpretation: The resulting tier is Tier 2 for the internal deployment and Tier 1 for the customer-facing deployment. The case illustrates the model-in-use framing: similar model class does not determine the tier by itself; classification changes because consequence severity, user reach, oversight, and regulatory exposure differ across deployment contexts.

4.7 Conclusion and Future Research

This paper developed a conceptual translation framework for integrating AI-specific and partly system-level risk drivers into inherent model risk assessment without abandoning established MRM tiering logic. Its central contribution is to show how AI-related risk manifestations can be interpreted through the two dimensions of MRM tiering rather than treated as a separate AI risk taxonomy: materiality, where they affect the severity or legal relevance of potential

consequences, and complexity, where they affect the probability of model failure or assessability. Aggregation and calibration provide a structured procedure for combining these assessments into reproducible inherent-risk tiers. The resulting tier remains a relative governance classification rather than a numerical estimate of model risk.

The model-in-use is the key unit of assessment. The same model artefact may generate different inherent-risk profiles depending on its business purpose. This framing allows system-level AI drivers to be incorporated into model-level tiering where they affect the severity of consequences, the probability of model failure, or assessability. The illustrative cases show that model class alone does not determine tier and that high complexity becomes most consequential where it coincides with high materiality. The framework thereby offers institutions a transparent basis for allocating proportionate validation and governance while retaining a unified model inventory.

The framework has several limitations. First, it has not yet been empirically calibrated or tested on model inventories. Second, the use cases are illustrative and do not provide validation evidence. Third, the qualitative synthesis and mapping involve interpretive judgement.

Future research should therefore focus on empirical calibration by applying the framework to model inventories and testing inter-rater reliability, sensitivity to alternative weighting schemes, robustness of thresholds, and resulting tier distributions.

References

Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2025). Intelligent financial system: How AI is transforming finance. *Journal of Financial Stability*, *81*, 101472. https://doi.org/10.1016/j.jfs.2025.101472

Alharthi, M., & Mahmood, A. (2024). xLSTMTime: Long-term time series forecasting with xLSTM. *AI*, *5*(3), 1482–1495. https://doi.org/10.3390/ai5030071

American Bankers Association. (2024). *Docket number TREAS-DO-2024-0011; response to request for information on uses, opportunities, and risks of artificial intelligence in the financial services sector*. https://www.aba.com/-/media/documents/comment-letter/jointltrai20240812.pdf?rev=c9c3bbc9004f4ec28e39280f6e82ccbf

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*(2), 579–625. https://doi.org/10.1111/1468-0262.00418

Baeza-Yates, R., & Fayyad, U. M. (2024). Responsible AI: An urgent mandate. *IEEE Intelligent Systems*, *39*(1), 12–17. https://doi.org/10.1109/MIS.2023.3343488

Bank of England. (2023a). *FS2/23–Artificial intelligence and machine learning*. https://www.bankofengland.co.uk/prudential-regulation/publication/2023/october/artificial-intelligence-and-machine-learning

Bank of England. (2023b). *SS1/23–Model risk management principles for banks*. https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/supervisory-statement/2023/ss123.pdf

Bank of England. (2024). *The Bank and the PRA's response to DSIT/HMT: Update on our approach to AI*. https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/letter/2024/dsit-hmt-letter.pdf

Basel Committee on Banking Supervision. (2026). *The Basel Framework.* Bank for International Settlements. https://www.bis.org/basel_framework/

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics*, *5*(3), 3265–3279. https://doi.org/10.1007/s43681-024-00653-w

Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, *245*, 108632. https://doi.org/10.1016/j.knosys.2022.108632

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). xLSTM: Extended long short-term memory. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang

(Eds.), *Advances in neural information processing systems 37* (pp. 107547–107603). Curran Associates Inc.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

Bennett, D. E. (2017). Governance and organizational requirements for effective model risk management. *Journal of Risk Model Validation*, *11*(4), 97–116. https://doi.org/10.21314/JRMV.2017.188

Bernard, C., Kazzi, R., & Vanduffel, S. (2020). A practical approach to quantitative model risk assessment. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3565116

Bhatt, U., & Sargeant, H. (2024). When should algorithms resign? A proposal for AI governance. *Computer*, *57*(10), 99–103. https://doi.org/10.1109/MC.2024.3431328

Board of Governors of the Federal Reserve System. (2011). *SR 11-7: Guidance on model risk management*. https://www.federalreserve.gov/boarddocs/srletters/2011/sr1107.pdf

Board of Governors of the Federal Reserve System. (2026). *SR 26-2: Revised guidance on model risk management*. https://www.federalreserve.gov/supervisionreg/srletters/SR2602a1.pdf

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2023). *The foundation model transparency index [Preprint]. arXiv.* http://arxiv.org/pdf/2310.12941v1

Boston Consulting Group. (2024). *Where's the value in AI?* https://web-assets.bcg.com/a5/37/be4ddf26420e95aa7107a35aae8d/bcg-wheres-the-value-in-ai.pdf

Breeden, S. (2024, October 31). *Engaging with the machine: AI and financial stability.* Bank of England. https://www.bankofengland.co.uk/speech/2024/october/sarah-breeden-keynote-speech-at-the-hong-kong-monetary-authority

Bridgers, A., Lee, H., & Kosoff, J. (2020). Adopting a top-down approach to model risk governance to optimize digital transformation. *The RMA Journal*, *103*(1), Article 103.

Brotcke, L. (2020). Modifying model risk management practice in the era of AI/ML. *Journal of Risk Management in Financial Institutions*, *13*(3), 255–265. https://doi.org/10.69554/DVBM8085

Brunner-Kirchmair, T. M., & Wiener, M. (2019). Knowledge is power—Conceptualizing collaborative financial risk assessment. *The Journal of Risk Finance*, *20*(3), 226–248. https://doi.org/10.1108/JRF-05-2018-0083

Buchanan, B. G. (2019). *Artificial intelligence in finance.* https://doi.org/10.5281/zenodo.2626454

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

Cancela-Outeda, C. (2024). The EU's AI Act: A framework for collaborative governance. *Internet of Things*, *27*, 101291. https://doi.org/10.1016/j.iot.2024.101291

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In M. Bailey & R. Greenstadt (Eds.), *Proceedings of the 30th USENIX security symposium* (pp. 2633–2650). USENIX Association.

Central Bank of the U.A.E. (2022). *Model management standards*. https://rulebook.centralbank.ae/sites/default/files/en_net_file_store/CBUAE_EN_4881_VER1.pdf

Chen, Z. (2025). Revolutionizing finance with conversational AI: A focus on ChatGPT implementation and challenges. *Humanities and Social Sciences Communications*, *12*(1), 1–11. https://doi.org/10.1057/s41599-025-04725-y

Cohen, S. N., Snow, D., & Szpruch, L. (2023). Black-box model risk in finance. In A. Capponi & C.-A. Lehalle (Eds.), *Machine learning and data sciences for financial markets* (pp. 687–717). Cambridge University Press. https://doi.org/10.1017/9781009028943.034

The Consumer Bankers Association. (2024). *Request for information on uses, opportunities, and risks of artificial intelligence in the financial services sector*. https://consumerbankers.com/wp-content/uploads/2024/08/CBA-Comment-on-TREAS-DO-2024-0011-AI-RFI.pdf

Cosma, S., Rimo, G., & Torluccio, G. (2023). Knowledge mapping of model risk in banking. *International Review of Financial Analysis*, *89*, 102800. https://doi.org/10.1016/j.irfa.2023.102800

Crisanto, J. C., Leuterio, C. B., Prenio, J., & Yong, J. (2024). *Regulating AI in the financial sector: Recent developments and main challenges*. FSI Insights. https://www.bis.org/fsi/publ/insights63.pdf

Cronin, M. A., & George, E. (2023). The why and how of the integrative review. *Organizational Research Methods*, *26*(1), 168–192. https://doi.org/10.1177/1094428120935507

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., . . . Sculley, D. (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, *23*(1), 1–61.

Daníelsson, J., Macrae, R., & Uthemann, A. (2022). Artificial intelligence and systemic risk. *Journal of Banking & Finance*, *140*, 106290. https://doi.org/10.1016/j.jbankfin.2021.106290

LSEG. (n.d.). *LSEG Workspace DataStream*. Available at: Subscription Service

Deloitte. (2021). *EMEA model risk management survey: From validation to optimisation, tackling a growing model landscape*. https://www.deloitte.com/lu/en/services/risk-advisory/research/emea-model-risk-management-survey.html

Deloitte. (2023). *EMEA model risk management survey: Preparing for a growing model landscape in a fast-changing world*. https://www.deloitte.com/nl/en/services/risk-advisory/research/deloitte-nl-risk-2023-emea-model-risk-management-survey.html

Deloitte. (2025a). *EMEA model risk management survey: Scaling model risk management to facilitate responsible growth leveraging advanced technology*. https://www.deloitte.com/content/dam/assets-zone2/dk/en/docs/generic/2025/2025-emea-model-risk-management-banks-orig.pdf

Deloitte. (2025b). *Governance of AI: A critical imperative for today's boards*. https://www.deloitte.com/content/dam/assets-shared/docs/about/2025/governance-of-ai-report-2nd-edition.pdf

Deloitte. (2025c). *The state of generative AI in the enterprise: 2024 year-end generative AI report*. https://www.deloitte.com/content/dam/assets-zone3/us/en/docs/campaigns/2025/us-state-of-gen-ai-2024-q4.pdf

Deutsche Bundesbank, & Federal Financial Supervisory Authority. (2021). *Machine learning in risk models–Characteristics and supervisory priorities*. https://www.bundesbank.de/en/tasks/banking-supervision/individual-aspects/risk-management/machine-learning-methods/machine-learning-methods-623122

Díaz, J. D., Hansen, E., & Cabrera, G. (2024). Machine-learning stock market volatility: Predictability, drivers, and economic value. *International Review of Financial Analysis*, *94*, 103286. https://doi.org/10.1016/j.irfa.2024.103286

DIN, & DKE. (2022). *Deutsche Normungsroadmap künstliche Intelligenz (Ausgabe 2)*. https://www.din.de/resource/blob/891106/57b7d46a1d2514a183a6ad2de89782ab/deutsche-normungsroadmap-kuenstliche-intelligenz-ausgabe-2--data.pdf

Dionne, P., & Howard, B. (2017). *Educational note—Use of models.* Canadian Institute of Actuaries. https://www.casact.org/sites/default/files/2021-03/6C_CIA_Models.pdf

Du Toit, H. A., Schutte, W. D., & Raubenheimer, H. (2023). Shapley values as an interpretability technique in credit scoring. *Journal of Risk Model Validation.* Advance online publication. https://doi.org/10.21314/JRMV.2023.010

Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S., & Schmidt-Thieme, L. (2021). *Do we really need deep learning models for time series forecasting? [Preprint]. arXiv.* https://doi.org/10.48550/arXiv.2101.02118

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987. https://doi.org/10.2307/1912773

Eskandarzadeh, A. (2026). Challenges and barriers to AI adoption. In H. Taherdoost (Ed.), *Transforming risk* (pp. 163–182). CRC Press. https://doi.org/10.1201/9781003621119-8

European Banking Authority. (2019). *Final report on EBA guidelines on outsourcing arrangements (EBA/GL/2019/02)*. https://www.eba.europa.eu/activities/single-rulebook/regulatory-activities/internal-governance/guidelines-outsourcing-arrangements

European Banking Authority. (2024). *Consultation paper on the RTS for assessing the materiality of extensions and changes of the internal ratings based approach*. https://www.eba.europa.eu/publications-and-media/press-releases/eba-consults-draft-technical-standards-specify-material-changes-and-extensions-internal-ratings

European Central Bank. (2024). *Draft guide on governance and risk culture*. https://www.bankingsupervision.europa.eu/framework/legal-framework/public-consultations/pdf/ssm.pubcon202407_draftguide.en.pdf

European Central Bank. (2025). *ECB Guide to internal models*. https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide202507.en.pdf?4460c67ecb5d677bacabc2e19942912f

The European Commission. (2025). *The general-purpose AI code of practice*. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

Federal Financial Supervisory Authority. (2022). *Machine learning in risk models– Characteristics and supervisory priorities: Responses to the consultation paper*. https://www.bafin.de/SharedDocs/Downloads/EN/Aufsichtsrecht/dl_Ergebnisse_machinelles_Lernen_Risikomodelle_en.html

Federal Office for Information Security. (2025). *Test criteria catalogue for AI systems in finance*. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/AI-Finance_Test-Criteria.html

Financial Reporting Council. (2023). *TAS 100: General actuarial standards v2.0*. https://media.frc.org.uk/documents/TAS_100_General_Actuarial_Standards_Version_2.0.pdf

Financial Reporting Council. (2024). *Technical actuarial guidance: Models*. https://media.frc.org.uk/documents/Technical_Actuarial_Guidance_Models_October_2024.pdf

Financial Services Agency of Japan. (2021). *Principles for model risk management*. https://www.fsa.go.jp/common/law/ginkou/pdf_03.pdf

Financial Stability Board. (2014). *Guidance on supervisory interaction with financial institutions on risk culture*. https://www.fsb.org/uploads/140407.pdf

Financial Stability Board. (2024). *The financial stability implications of artificial intelligence*. https://www.fsb.org/uploads/P14112024.pdf

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In L. Floridi (Ed.), *Ethics, governance, and policies in artificial intelligence* (Vol. 144, pp. 5–17). Springer. https://doi.org/10.1007/978-3-030-81907-1_2

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), 1–37. https://doi.org/10.1145/2523813

Garcia, A. C. B., Garcia, M. G. P., & Rigobon, R. (2024). Algorithmic discrimination in the credit domain: What do we know about it? *AI & SOCIETY*, *39*(4), 2059–2098. https://doi.org/10.1007/s00146-023-01676-3

García-Llorente, C., & Olmeda, I. (2026). Algorithmic governance in banking: A comparative analysis of risk-based and accountability-oriented oversight. *Journal of Banking Regulation*, *27*(2). https://doi.org/10.1057/s41261-026-00320-6

Gehrmann, S., Huang, C., Teng, X., Yurovski, S., Bhorkar, A., Thomas, N., Doucette, J., Rosenberg, D., Dredze, M., & Rabinowitz, D. (2025). Understanding and mitigating risks of generative AI in financial services. In J. W. Vaughan, S. Mohamed, S. Fazelpour, & T. B. Gillis (Eds.), *FAccT '25: Proceedings of the 2025 ACM conference on fairness, accountability, and transparency* (pp. 2570–2586). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732168

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Giudici, P. (2024). Safe machine learning. *Statistics*, *58*(3), 473–477. https://doi.org/10.1080/02331888.2024.2361481

Giudici, P., Centurelli, M., & Turchetta, S. (2024). Artificial intelligence risk measurement. *Expert Systems with Applications*, *235*, 121220. https://doi.org/10.1016/j.eswa.2023.121220

Giudici, P., & Kolesnikov, V. (2026). SAFE AI metrics: An integrated approach. *Machine Learning with Applications*, *23*, 100821. https://doi.org/10.1016/j.mlwa.2025.100821

Giudici, P., Piergallini, A., Recchioni, M. C., & Raffinetti, E. (2024). Explainable artificial intelligence methods for financial time series. *Physica A: Statistical Mechanics and Its Applications*, *655*(5), 130176. https://doi.org/10.1016/j.physa.2024.130176

Giudici, P., & Raffinetti, E. (2023). SAFE Artificial Intelligence in finance. *Finance Research Letters*, *56*, 104088. https://doi.org/10.1016/j.frl.2023.104088

The Global Financial Markets Association. (2025). *GFMA public comment letter on IOSCO's report on artificial intelligence in capital markets: Use cases, risks, and challenges*. https://www.gfma.org/wp-content/uploads/2025/04/gfma-response-to-iosco-ai-consultation-report-final-4.11.2580.pdf

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

Gonzalez Torres, A. P., Kajava, K., & Sawhney, N. (2023). Emerging AI discourses and policies in the EU: Implications for evolving AI governance. In A. Pillay, E. Jembere, & A. J. Gerber (Eds.), *Communications in computer and information science,*

artificial intelligence research (pp. 3–17). Springer. https://doi.org/10.1007/978-3-031-49002-6_1

Gstrein, O. J., Haleem, N., & Zwitter, A. (2024). General-purpose AI regulation and the European Union AI Act. *Internet Policy Review*, *13*(3). https://doi.org/10.14763/2024.3.1790

Gunnarsson, E. S., Isern, H. R., Kaloudis, A., Risstad, M., Vigdel, B., & Westgaard, S. (2024). Prediction of realized volatility and implied volatility indices using AI and machine learning: A review. *International Review of Financial Analysis*, *93*, 103221. https://doi.org/10.1016/j.irfa.2024.103221

Hansen, P. R., & Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, *131*(1–2), 97–121. https://doi.org/10.1016/j.jeconom.2005.01.005

Haskins, B., Stecklein, J., Dick, B., Moroney, G., Lovell, R., & Dabney, J. (2004). 8.4.2 Error cost escalation through the project life cycle. In *INCOSE 2004: 14th annual international symposium proceedings* (Vol. 14, pp. 1723–1737). INCOSE. https://doi.org/10.1002/j.2334-5837.2004.tb00608.x

Hassani, B. (2019). Model risk management: From epistemology to corporate governance. *Journal of Risk Model Validation*, *13*(3), 1–21. https://doi.org/10.21314/JRMV.2019.208

Hill, J. R. (2019). The top 14 challenges for today's model risk managers: Has the time come to think about going beyond SR11-7? *Journal of Risk Management in Financial Institutions*, *12*(2), 145–167. https://doi.org/10.69554/PDOJ9623

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

Hurlin, C., Pérignon, C., & Saurin, S. (2026). The fairness of credit scoring models. *Management Science*, *72*(1), 406–425. https://doi.org/10.1287/mnsc.2022.03888

ISO/IEC. (2023a). *ISO/IEC 23894:* Information technology—Artificial intelligence—Guidance on risk management.

ISO/IEC. (2023b). *ISO/IEC 42001:* Information technology—Artificial intelligence—Management system.

J.P. Morgan. (2024). *Quest IndexGPT: Harnessing generative AI for investable indices*. https://www.jpmorgan.com/insights/markets/indices/indexgpt

Jhin, S. Y., Kim, S., & Park, N. (2024). Addressing prediction delays in time series forecasting: A continuous GRU approach with derivative regularization. In R. Baeza-Yates & F. Bonchi (Eds.), *KDD '24: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 1234–1245). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671969

Kallina, E., & Singh, J. (2024). Stakeholder involvement for responsible AI development: A process framework. In N. Arnosti, C. Calsamiglia, S. Ruiz-Correa, & J. P. Dickerson (Eds.), *EAAMO '24: Proceedings of the 4th ACM conference on equity and access in algorithms, mechanisms, and optimization* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3689904.3694698

Kazim, E., & Koshiyama, A. (2021). The interrelation between data and AI ethics in the context of impact assessments. *AI and Ethics*, *1*(3), 219–225. https://doi.org/10.1007/s43681-020-00029-w

Keenan, S. C. (2015). *Financial institution advantage and the optimization of information processing*. Wiley. https://doi.org/10.1002/9781119053309

Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications*, *103*, 25–37. https://doi.org/10.1016/j.eswa.2018.03.002

Kiritz, N., Ravitz, M., & Levonian, M. (2019). Model risk tiering: An exploration of industry practices and principles. *Journal of Risk Model Validation*, *13*(2), 47–77. https://doi.org/10.21314/JRMV.2019.205

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Kondor, D., Hafez, V., Shankar, S., Wazir, R., & Karimi, F. (2024). Complex systems perspective in assessing risks in artificial intelligence. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *382*(2285), 20240109. https://doi.org/10.1098/rsta.2024.0109

KPMG. (2022). *Modern risk management for AI models: Re-imagining the model risk management function for artificial intelligence/machine learning models*. https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2022/07/modern-risk-management-for-ai-models.pdf

KPMG. (2023a). *Changing the game in model risk management*. https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2023/09/changing-the-game-in-model-risk-management.pdf

KPMG. (2023b). *Increasing challenges for model risk management (MRM) within financial institutions*. https://assets.kpmg.com/content/dam/kpmgsites/es/pdf/2023/01/model-risk-management.pdf.coredownload.inline.pdf

KPMG. (2024a). *AI trust*. https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2024/trusted-ai-slipsheet.pdf

KPMG. (2024b). *Effective model risk management framework for AI/ML based models*. https://assets.kpmg.com/content/dam/kpmgsites/in/pdf/2024/10/effective-model-risk-management-framework-for-ai-ml-based-models.pdf.coredownload.inline.pdf

KPMG. (2024c). *Trusted AI governance*. https://assets.kpmg.com/content/dam/kpmg/au/pdf/2024/kpmg-trusted-ai-governance-2024.pdf

Krisper, M. (2021). *Problems with risk matrices using ordinal scales [Preprint]. arXiv.* https://doi.org/10.48550/arXiv.2103.05440

Kristjanpoller, W., Fadic, A., & Minutolo, M. C. (2014). Volatility forecast using hybrid neural network models. *Expert Systems with Applications*, *41*(5), 2437–2442. https://doi.org/10.1016/j.eswa.2013.09.043

Kumar, S., Rao, A., & Dhochak, M. (2025). Hybrid ML models for volatility prediction in financial risk management. *International Review of Economics & Finance*, *98*, 103915. https://doi.org/10.1016/j.iref.2025.103915

Kurshan, E., Shen, H., & Chen, J. (2021). Towards self-regulating AI. In T. Balch (Ed.), *Proceedings of the first ACM international conference on AI in finance (ICAIF '20)* (pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3383455.342256

Lacmanovic, S., & Skare, M. (2025). Artificial intelligence bias auditing–Current approaches, challenges and lessons from practice. *Review of Accounting and Finance*, *24*(3), 375–400. https://doi.org/10.1108/RAF-01-2025-0006

Le Guen, V., & Thome, N. (2019). Shape and time distortion loss for training deep time series forecasting models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32 (pp. 4189–4201).* Curran Associates Inc.

Lee, S. U., Perera, H., Liu, Y., Xia, B., Lu, Q., Zhu, L., Salvado, O., & Whittle, J. (2026). Responsible AI question bank for risk assessment. *ACM Computing Surveys*, *58*(9), 1–36. https://doi.org/10.1145/3790096

Leinarte, E. (2024). The classification of high-risk AI systems under the EU Artificial Intelligence Act. *Journal of AI Law and Regulation*, *1*(3), 262–280. https://doi.org/10.21552/aire/2024/3/4

Li, Y., & Goel, S. (2025). Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems. *International Journal of Accounting Information Systems*, *56*, Article 100739. https://doi.org/10.1016/j.accinf.2025.100739

Liu, J., Li, W., Jin, Z., & Diab, M. (2024). Automatic generation of model and data cards: A step towards responsible AI. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: Human language technologies (Volume 1: Long Papers)* (pp. 1975–1997). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.110

Liu, W. K., & So, M. K. P. (2020). A GARCH model with artificial neural networks. *Information*, *11*(10), 489. https://doi.org/10.3390/info11100489

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2023). *iTransformer: Inverted transformers are effective for time series forecasting [Preprint]. arXiv.* https://doi.org/10.48550/arXiv.2310.06625

Machado, P. (2026, February 24). *Technology is neutral, governance is not: AI adoption in the banking sector*. KPMG RiskTech Conference, Frankfurt am Main. https://www.bankingsupervision.europa.eu/press/speeches/date/2026/html/ssm.sp260224~6c5b64a77a.en.html

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In R. Bernhaupt, F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, & R. Kocielnik (Eds.), *CHI '20: Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376445

Maheshwari, A. (2026). Generative artificial intelligence in model risk management: Emerging opportunities, supervisory challenges and validation frameworks. *Journal of Risk Model Validation.* Advance online publication. https://doi.org/10.21314/JRMV.2026.004

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Mankotia, S., & Joshi, A. (2013). Measuring model risk: A practitioner's approach. *The RMA-Journal*, *95*(10), 28–37.

Mantelero, A. (2018). AI and big data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, *34*(4), 754–772. https://doi.org/10.1016/j.clsr.2018.05.017

Mayenberger, D. (2019). How to overcome modelling and model risk management challenges with artificial intelligence and machine learning. *Journal of Risk Management in Financial Institutions*, *12*(3), 241–255. https://doi.org/10.69554/REFL4905

McKinsey & Company. (2025a). *Open source technology in the age of AI*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/open-source-technology-in-the-age-of-ai#/

McKinsey & Company. (2025b). *The state of AI: How organizations are rewiring to capture value*. https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2025/the-state-of-ai-how-organizations-are-rewiring-to-capture-value_final.pdf

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency (FAT* '19)* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Monetary Authority of Singapore. (2024). *Artificial intelligence (AI) model risk management*. https://www.mas.gov.sg/publications/monographs-or-information-paper/2024/artificial-intelligence-model-risk-management

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Myllyaho, L., Raatikainen, M., Männistö, T., Mikkonen, T., & Nurminen, J. K. (2021). Systematic literature review of validation methods for AI systems. *Journal of Systems and Software*, *181*, 111050. https://doi.org/10.1016/j.jss.2021.111050

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*(2), 347. https://doi.org/10.2307/2938260

New York State—Department of Financial Services. (2024). *Insurance circular letter No. 7*. https://www.dfs.ny.gov/industry-guidance/circular-letters/cl2024-07

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). Taking AI risks seriously: A new assessment model for the AI Act. *AI & SOCIETY*, *39*(5), 2493–2497. https://doi.org/10.1007/s00146-023-01723-z

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Office of the Comptroller of the Currency. (2021). *Comptroller's handbook—Model risk management*. https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html

Office of the Superintendent of Financial Institutions. (2025). *Guideline E-23—Model risk management (2027)*. https://www.osfi-bsif.gc.ca/en/guidance/guidance-library/guideline-e-23-model-risk-management-2027

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). *Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift [Preprint]. arXiv.* https://doi.org/10.48550/arXiv.1906.02530

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers*, *25*(1), 123–141. https://doi.org/10.1007/s10796-022-10251-y

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Pérez-Hernández, F., Arévalo-de-Pablos, A., & Camacho-Miñano, M.-M. (2024). A hybrid model integrating artificial neural network with multiple GARCH-type models and EWMA for performing the optimal volatility forecasting of market risk factors. *Expert Systems with Applications*, *243*, 122896. https://doi.org/10.1016/j.eswa.2023.122896

Petrozziello, A., Troiano, L., Serra, A., Jordanov, I., Storti, G., Tagliaferri, R., & La Rocca, M. (2022). Deep learning for volatility forecasting in asset management. *Soft Computing*, *26*(17), 8553–8574. https://doi.org/10.1007/s00500-022-07161-1

Pialla, G., Ismail Fawaz, H., Devanne, M., Weber, J., Idoumghar, L., Muller, P.-A., Bergmeir, C., Schmidt, D. F., Webb, G. I., & Forestier, G. (2025). Time series adversarial attacks: An investigation of smooth perturbations and defense approaches. *International Journal of Data Science and Analytics*, *19*(1), 129–139. https://doi.org/10.1007/s41060-023-00438-0

Picciarelli, A. (2024, May 24). *Model risk management: A possible audit approach in the digitalization and AI era* [Conference presentation]. Audit Masters—9th Annual Internal Audit Forum, Lisbon, Portugal. GLC Europe. https://glceurope.com/upload/files/304/day-2-14-10-antonio-picciarelli-26461.pdf

Poon, S.-H., & Granger, C. W. J. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, *41*(2), 478–539. https://doi.org/10.1257/002205103765762743

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI and Ethics*, *3*(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9

Prorokowski, L. (2025). Model risk quantification for machine learning models in credit risk. *The Journal of Risk Model Validation*, *19*(2), 73–109. https://doi.org/10.21314/JRMV.2025.006

PwC. (2021). *PwC response to NIST RFI: Developing a federal AI standards engagement plan*. https://www.nist.gov/system/files/documents/2021/10/14/nist-ai-rfi-pwc_001.pdf

PwC. (2022). *Erfolgreich mit künstlicher Intelligenz: Sind sie sich der Verantwortung und Pflichten bewusst?* https://www.pwc.de/de/finanzdienstleistungen/digital/pwc-whitepaper-responsible-ai.pdf

PwC. (2023). *Managing the risks of generative AI*. https://www.pwc.com/us/en/tech-effect/ai-analytics/managing-generative-ai-risks.html

PwC. (2024). *Model risk management survey 2024*. https://www.pwc.com/cz/cs/temata/MRM-Survey-Report-2024.pdf

Qiu, Z., Lazar, E., & Nakata, K. (2024). VaR and ES forecasting via recurrent neural network-based stateful models. *International Review of Financial Analysis*, *92*, 103102. https://doi.org/10.1016/j.irfa.2024.103102

Radanliev, P. (2025). AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence*, *39*(1), 1–41. https://doi.org/10.1080/08839514.2025.2463722

Rao, P. S. B., Šćepanović, S., Zhou, K., Bogucka, E. P., & Quercia, D. (2025). RiskRAG: A data-driven solution for improved AI model risk reporting. In N. Yamashita, V. Evers, K. Yatani, X. Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI conference on human factors in computing systems (CHI '25)* (pp. 1–26). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713979

Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012, Official Journal L 176 (2013). http://data.europa.eu/eli/reg/2013/575/oj

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal L 119 (2016). http://data.europa.eu/eli/reg/2016/679/oj

Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011, Official Journal L 333 (2022). http://data.europa.eu/eli/reg/2022/2554/oj

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Official Journal L 1689 (2024). http://data.europa.eu/eli/reg/2024/1689/oj

Renieris, E. M., Kiron, D., & Mills, S. (2023). *Building robust RAI programs as third-party AI tools proliferate.* MIT Sloan Management Review and The Boston Consulting Group. https://sloanreview.mit.edu/projects/building-robust-rai-programs-as-third-party-ai-tools-proliferate/

Risk Management Association. (2022). *The risk management association survey of model risk management*. https://rmahq-01-content.s3.amazonaws.com/media/s2nmnace/survey-of-mrm_es_optimized_final_071323_rgb.pdf

Rollason, V., Fisk, G., & Haines, P. (2010). *Applying the ISO 31000 risk assessment framework to coastal zone management*. https://www.coastalconference.com/2010/papers2010/Verity%20Rollason%20full%20paper%202.pdf

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Rutinowski, J., Klüttermann, S., Endendyk, J., Reining, C., & Müller, E. (2024). Benchmarking trust: A metric for trustworthy machine learning. In L. Longo, S. Lapuschkin, & C. Seifert (Eds.), *Communications in computer and information science. Explainable artificial intelligence* (pp. 287–307). Springer. https://doi.org/10.1007/978-3-031-63787-2_15

Sadek, M., Kallina, E., Bohné, T., Mougenot, C., Calvo, R. A., & Cave, S. (2025). Challenges of responsible AI in practice: Scoping review and recommended actions. *AI & SOCIETY*, *40*(1), 199–215. https://doi.org/10.1007/s00146-024-01880-9

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. Drucker (Eds.), *CHI '21: Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–15). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445518

Sandu, I., Wiersma, M., & Manichand, D. (2022). Time to audit your AI algorithms. *Maandblad Voor Accountancy En Bedrijfseconomie*, *96*(7/8), 253–265. https://doi.org/10.5117/mab.96.90108

Sarmah, B., Li, M., Lyu, J., Frank, S., Castellanos, N., Pasquali, S., & Mehta, D. (2024). How to choose a threshold for an evaluation metric for large language models *[Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.12148*

SAS, & GARP. (2020). *Model risk management: Today's governance and future directions*. https://www.sas.com/content/dam/SAS/documents/event-collateral/2020/English/mrm-survey-report-with-garp.pdf

Sayles, J. (Ed.). (2024). *Principles of AI governance and model risk management*. Apress. https://doi.org/10.1007/979-8-8688-0983-5

Schmitz, A., Mock, M., Görge, R., Cremers, A. B., & Poretschkin, M. (2025). A global scale comparison of risk aggregation in AI assessment frameworks. *AI and Ethics*, *5*(2), 1407–1432. https://doi.org/10.1007/s43681-024-00479-6

Schuett, J. (2024). Risk management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, *15*(2), 367–385. https://doi.org/10.1017/err.2023.1

Schuett, J., Reuel, A.-K., & Carlier, A. (2025). How to design an AI ethics board. *AI and Ethics*, *5*(2), 863–881. https://doi.org/10.1007/s43681-023-00409-y

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In C. Cortes, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *NIPS'15: Proceedings of the 29th international conference on neural information processing systems* (pp. 2503–2511). MIT Press.

The Securities and Futures Commission. (2024). *Use of generative AI language models*. https://apps.sfc.hk/edistributionWeb/gateway/EN/circular/intermediaries/supervision/doc?refNo=24EC55

The Securities Industry and Financial Markets Association and its Asset Management Group. (2024). *Request for comment on uses, opportunities, and risks of artificial intelligence in the financial services sector*. https://www.sifma.org/wp-content/uploads/2024/08/SIFMA-Treasury-RFI-AI-Response-July-2024-FINAL.pdf

Shelley, C. (2022). The fairness impact assessment. *International Journal of Technoethics*, *13*(1), 1–16. https://doi.org/10.4018/IJT.291554

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3–18). IEEE. https://doi.org/10.1109/SP.2017.41

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, *126*(6), 5113–5142. https://doi.org/10.1007/s11192-021-03948-5

Sinha, S., & Lee, Y. M. (2024). Challenges with developing and deploying AI models and applications in industrial systems. *Discover Artificial Intelligence*, *4*, Article 5. https://doi.org/10.1007/s44163-024-00151-2

Souza, C. (2023). AI model risk: What the current model risk management framework can teach us about managing the risks of AI models. *Journal of Financial Compliance*, *6*(2), 103–112. https://doi.org/10.69554/SOKX4074

Spatola, N. (2024). The efficiency-accountability tradeoff in AI integration: Effects on human performance and over-reliance. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100099. https://doi.org/10.1016/j.chbah.2024.100099

Sudhir, P., Suresh, S., & Thontadari, C. (2025). Leveraging agentic RAG to reduce hallucinations in large language models. In S. Shukla, H. Sayama, K. Tiwari, & J. V. Kureethara (Eds.), *Lecture notes in networks and systems. Data science and security* (Vol. 1355, pp. 31–42). Springer Nature. https://doi.org/10.1007/978-981-96-4883-2_4

Sudjianto, A., & Zhang, A. (2024). *Model validation practice in banking: A structured approach for predictive models [Preprint]. arXiv.* https://doi.org/10.48550/arXiv.2410.13877

Tsai, P. (2025). How to build an AI governance framework with data governance. In R. Fioravante & A. Vaccaro (Eds.), *SpringerBriefs in philosophy. Humanism and artificial intelligence* (pp. 171–186). Springer. https://doi.org/10.1007/978-3-031-85478-1_10

U.S. Department of the Treasury. (2024). *Artificial intelligence in financial services: Report on the use, opportunities, and risks of artificial intelligence in the financial services sector*. https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *NIPS'17: Proceedings of the 31st international conference on neural information processing systems (pp. 6000–6010).* Curran Associates Inc.

Vintsyuk, T. K. (1972). Speech discrimination by dynamic programming. *Cybernetics*, *4*(1), 52–57. https://doi.org/10.1007/BF01074755

Wang, Y., Andreeva, G., & Martin-Barragan, B. (2023). Machine learning approaches to forecasting cryptocurrency volatility: Considering internal and external determinants. *International Review of Financial Analysis*, *90*, 102914. https://doi.org/10.1016/j.irfa.2023.102914

Wang, Y., Ye, W., Jiang, Y., & Liu, X. (2024). Volatility prediction for the energy sector with economic determinants: Evidence from a hybrid model. *International Review of Financial Analysis*, *92*, 103094. https://doi.org/10.1016/j.irfa.2024.103094

Winecoff, A., & Bogen, M. (2025). Improving governance outcomes through AI documentation: Bridging theory and practice. In N. Yamashita, V. Evers, K. Yatani, X. Ding, B. Lee, M. Chetty, & P. Toups-Dugas (Eds.), *Proceedings of the 2025 CHI conference on human factors in computing systems (CHI '25)* (pp. 1–18). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713814

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *NIPS'21: Proceedings of the 35th international conference on neural information processing systems* (pp. 22419–22430). Curran Associates Inc.

Wu, Y. (2023). Comparison between transformer, informer, autoformer and non-stationary transformer in financial market. *Applied and Computational Engineering*, *29*(1), 68–78. https://doi.org/10.54254/2755-2721/29/20230874

Yang, A., Ye, Q., & Zhai, J. (2024). Volatility forecasting with hybrid-long short-term memory models: Evidence from the COVID-19 period. *International Journal of Finance & Economics*, *29*(3), 2766–2786. https://doi.org/10.1002/ijfe.2805

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zaidan, E., & Ibrahim, I. A. (2024). AI governance in a complex and rapidly changing regulatory landscape: A global perspective. *Humanities and Social Sciences Communications*, *11*(1), 1–18. https://doi.org/10.1057/s41599-024-03560-x

Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are transformers effective for time series forecasting? In B. Williams, Y. Chen, & J. Neville (Eds.), *Proceedings of the AAAI conference on artificial intelligence* (pp. 11121–11128). AAAI Press. https://doi.org/10.1609/aaai.v37i9.26317

Zhang, J., & Chen, L. (2022). Application of neural network with autocorrelation in long-term forecasting of systemic financial risk. *Computational Intelligence and Neuroscience*, *2022*, 7131143. https://doi.org/10.1155/2022/7131143

Zhang, X., Chan, F. T., Yan, C., & Bose, I. (2022). Towards risk-aware artificial intelligence and machine learning systems: An overview. *Decision Support Systems*, *159*, 113800. https://doi.org/10.1016/j.dss.2022.113800

Zhang, X., Wang, T., Ma, L., & Mahadevan, S. (2025). Reliability engineering, risk management, and trustworthiness assurance for AI systems. *Journal of Reliability Science and Engineering*, *1*(2), 22001. https://doi.org/10.1088/3050-2454/adcef2

Zhang, Z., Chen, Y., Zhang, D., Qian, Y., & Wang, H. (2024). Ctfnet: Long-sequence time-series forecasting based on convolution and time-frequency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(11), 16368–16382. https://doi.org/10.1109/TNNLS.2023.3294064

Zhao, P., Zhu, H., NG, W. S. H., & Lee, D. L. (2024). From GARCH to neural network for volatility forecast. In M. Wooldridge, J. Dy, & S. Natarajan (Eds.), *Proceedings of the AAAI conference on artificial intelligence*, *38*(15), 16998–17006. https://doi.org/10.1609/aaai.v38i15.29643

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In K. Leyton-Brown & Mausam (Eds.), *Proceedings of the AAAI conference on artificial intelligence* (pp. 11106–11115). AAAI Press. https://doi.org/10.1609/aaai.v35i12.17325

Zhou, N., Zhang, Z., Nair, V. N., Singhal, H., & Chen, J. (2022). Bias, fairness and accountability with artificial intelligence and machine learning algorithms. *International Statistical Review*, *90*(3), 468–480. https://doi.org/10.1111/insr.12492