Effective Generative AI Model Risk Management

Joseph L. Breeden	Deep Future Analytics
Raymond Anderson	Rayan Risk Analytics
Peter Quell	DZ Bank AG
Shannon Kelly	Quantitative Risk Partners

7 August 2025

TABLE OF CONTENTS

Executive Summary

This paper outlines a model risk management (MRM) framework tailored to Large Language Models (LLMs) in finance and insurance. While existing MRM principles remain relevant, their application must evolve. Traditional models are validated against fixed datasets and exhibit stable behavior, but LLMs generate dynamic, unpredictable outputs that demand continuous oversight rather than static validation.

The core challenge is that LLMs can behave unexpectedly in real-world contexts, especially in response to outlier inputs or changing prompts. Human-in-the-loop (HITL) monitoring has been proposed, but studies show that humans become unreliable in low-error-rate environments. A more effective approach is AI-assisted compliance monitoring, using secondary models to flag potential issues for human review.

Inappropriate use of models has always been a concern in model risk management. LLMs also heighten this risk because of the creative and unstructured interactions between a user and LLM. Consequently, model monitoring needs to expand to usage monitoring, monitoring the questions posed as well as the answers received. Staff should not create their own AI Accomplice.

Most LLMs used in finance are developed externally, limiting transparency into training data and embedded biases. As a result, institutions must document their own contributions—such as fine-tuning data, prompt engineering, and output moderation—rather than attempting full model audits.

Effective oversight also requires operational readiness to respond to failures. Because LLMs can fail catastrophically, institutions must establish clear thresholds for disengaging models and activating fallback systems. These fallbacks—whether human agents or alternative models—must be tested, maintained, and immediately available. A champion-challenger deployment model is recommended, where the fallback system runs in parallel to the primary model.

Monitoring LLMs will take priority over validation, especially for general-purpose models. Retrieval-augmented generation (RAG) systems are more amenable to traditional validation, but broader LLMs need new oversight techniques. One solution is to employ an LLM system that compares user-AI interactions against a structured set of assertions based on compliance and ethics rules. This method avoids reported problems with “LLM-as-Judge” and enables both automated triage and performance tracking.

The questions of ”What is a model?” and ”How they should be tracked in the model inventory?” are also being overhauled. No perfect phrase exists to draw this distinction, but MRM teams are deciding which AI-enabled applications are tools subject to IT oversight and which are models deserving of MRM treatment. Once in the inventory, we also see the need for

a new severity level: High, Medium, Low, Unknown. Most LLM deployments should initially be categorized as “unknown” severity with aggressive oversight until enough usage review has been performed to properly assess the risk.

Ultimately, MRM principles still hold, but the methods must shift. Institutions must move from static validation to dynamic oversight, use AI to support human judgment, and prepare robust contingency plans. These updates will allow financial institutions to adopt GenAI technologies responsibly without compromising risk controls.

Introduction

Artificial Intelligence (AI) is rapidly reshaping the landscape of financial and insurance services. While traditional AI and machine learning (ML) systems have long been used for well-defined tasks like credit scoring, fraud detection, or product recommendation, the emergence of Generative AI (GenAI) marks a significant shift. Among GenAI technologies, Large Language Models (LLMs) stand out for their ability to generate human-like text and respond flexibly across a wide range of topics and tasks. These capabilities are increasingly being applied to customer service, transaction support, and decision-making processes.

However, with this power comes a new class of risks. Most existing Model Risk Management (MRM) frameworks were built for traditional, deterministic models that behave in predictable ways. These frameworks rely heavily on pre-deployment validation, testing against fixed datasets, and periodic performance reviews. GenAI, by contrast, generates novel outputs each time it runs, shaped by subtle shifts in inputs, user behavior, and real-time context. This makes full pre-deployment validation nearly impossible, particularly for general-purpose LLMs.

Instead, effective MRM for GenAI must prioritize ongoing monitoring and dynamic oversight. Human-in-the-loop (HITL) models, where a human can intervene when needed, are one approach. But human attention is fallible, especially when systems work well most of the time, leading to automation complacency. A more scalable alternative is AI-assisted oversight: using one AI system to monitor another and flag anomalies for human review. This layered approach is gaining traction in both research and practice and helps ensure compliance with ethical, legal, and policy standards.

Another key challenge is transparency. Financial institutions often license LLMs from third-party vendors, with limited visibility into model architecture or training data. As a result, firms must shift their focus toward what they can control, such as prompt engineering, fine-tuning data, and post-deployment modifications, and develop audit processes around these controllable inputs.

Finally, monitoring and governance alone are not enough. Because GenAI systems can fail in unexpected and high-impact ways, institutions must be ready to act quickly. This includes having clear shutdown criteria and live fallback options, such as human agents or standby models, ready to take over in real time. The “champion-challenger” setup, where a backup model runs alongside the primary one, is a proven approach in high-risk systems and should become standard for LLM deployment.

This white paper outlines a practical framework for adapting MRM to the realities of GenAI, focusing especially on LLMs and Retrieval-Augmented Generation (RAG) systems in financial services. It begins by providing an overview of generative AI and its potential risks before outlining the foundational principles of MRM and the

limitations of traditional model validation. It then explores specific oversight strategies, including human and AI-assisted monitoring, before turning to issues of transparency, disaster planning, and fallback design. It concludes with recommendations for governance structures and operational practices that can support responsible GenAI deployment in financial and insurance settings.

Generative AI

Generative Artificial Intelligence (GenAI) includes a range of technologies that can create new content, such as written text, pictures, music, or even videos, by learning from examples. Instead of just recognizing patterns in data, it produces new content that looks or sounds similar to the examples it was trained on. Several kinds of GenAI exist, each with its own way of working (albeit some share certain architectural features).

These include tools like Large Language Models (LLMs, e.g. ChatGPT), which can write human-like text; Retrieval-Augmented Generation (RAG), which pulls in outside information to give better answers; and others that can create images or music, like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models. In banking and insurance, the most useful types are text-based tools like LLMs and RAG, which aid report writing, answering customer questions, and reviewing complex documents.

LLMs learn to write by studying huge amounts of text (*corpora*), from books, websites, and more, and are based on transformer architectures, which process sequences using mechanisms like self-attention and assess the entire context of a sentence (not just individual words) to generate what comes next. That's how they generate human-like responses to questions or prompts.[1] You can guide how creative or consistent their answers are using settings like "temperature", and even small changes in the prompt, or the amount of detail, can lead to very different results.

LLMs are useful for tasks such as writing summaries, translating languages, answering questions, and holding conversations. They are called "foundation models" due to the broad applicability made possible by them being trained on massive datasets, which also applies to other powerful AIs that work with images, speech, and other data.

RAG models are a special kind of AI tool designed to provide more accurate and reliable answers. Unlike other GenAI models, which rely only on what they learned during training, RAG models can reference a set of approved documents (like a company's internal database or knowledge base) in real time before responding.[2] This helps keep the answers grounded in trusted information and

[1] All questions are prompts, but not all prompts are questions. Prompt engineering is an evolving field

[2] Izacard, G., & Grave, E. (2021). Distilling Knowledge from Reader to Retriever for Question Answering. Proceedings of the 9th International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2012.04584

reduces the chance of making things up (a problem known as "hallucination").

Because they combine generation with retrieval, RAG models are especially useful for tasks like answering complex questions, helping customers, or summarizing legal and financial documents. That said, they can be expensive to set up, and they only work well when a company has high-quality reference sources.

Classifying AI Risks

Many risks have been identified as we adopt GenAI. The following grouping combines these into high-level categories to highlight the distinct risk management approaches for each: (i) societal, (ii) misuse, (iii) control, and (iv) others that cut across these categories and warrant dedicated attention.

Societal risk arises when GenAI's output doesn't comply with ethical, regulatory, or legal standards:

- *Misinformation*: Failing to accurately provide the information requested by consumers can lead to them making the wrong decisions.
- *Ethically Challenged Outputs*: AI communications that embed biases, reinforce discrimination, or produce outcomes that contradict ethical norms asserted by the institutions deploying these systems.
- *Lawbreakers*: Communications or actions that fall foul of existing laws and regulations. For example, well-behaved AIs developed using international data may not comply with local laws.

This list is focused on the issues within financial services. It does not consider the broader issues of potential economic, ecological, or cultural harm (including widening inequalities) from using these models. Those topics warrant intense study as well.

Governance is needed to ensure that automated systems comply at least as well as their human counterparts, a key point that is often overlooked. LLMs are imperfect, but so are humans in similar roles. The standard for success is to be as good as humans, assuming that we can develop appropriate effectiveness metrics.

LLMs are trained using the entirety of available human writings and communications, including much that is not compliant with the standards that we set. This creates an anomalous situation of expecting LLMs to perform better than the data on which they were trained.

Real-world failures of LLMs have already been observed because of this "Do as I say, not as I do" requirement. As the systems improve, failures will be less common, but this also creates a more challenging task to monitor and prevent rare events. These risks are tangible, systemic, and already demanding policy intervention.

Misuse risk arises from potential abuse by malicious "black hat" actors. AI can be misused for fraud, misinformation, and even cyberattacks without strong safeguards.

- *Fraud*: Generative AI is already being used to create fake identities or impersonate real people, making identity theft easier and harder to detect. This shows why we need better systems to confirm that people are who they say they are.
- *Unauthorized Agents*: Some AI systems can act on their own, which raises the risk that they might carry out actions, like transactions, without the right approval.

This is exactly like ensuring people have the authority to act on behalf of others.

- *Security Threats*: AI can be used to launch smarter, faster cyberattacks, putting our digital systems, banks, and even national security at risk.
- *Input/Prompt Risk*: Poor-quality data or adversarial prompts can mislead GenAI systems. In RAG models, internal fraud could be committed by injecting prompts into internal documents. Prompt injection and jailbreaking attacks are emerging threats.

These types of risks come not from the AI making mistakes, but from inappropriate use and people using AI with harmful intentions. Defense against these risks requires stronger digital defenses and new ways to verify the identity of both people and AI systems, especially in customer interaction and decision automation contexts.

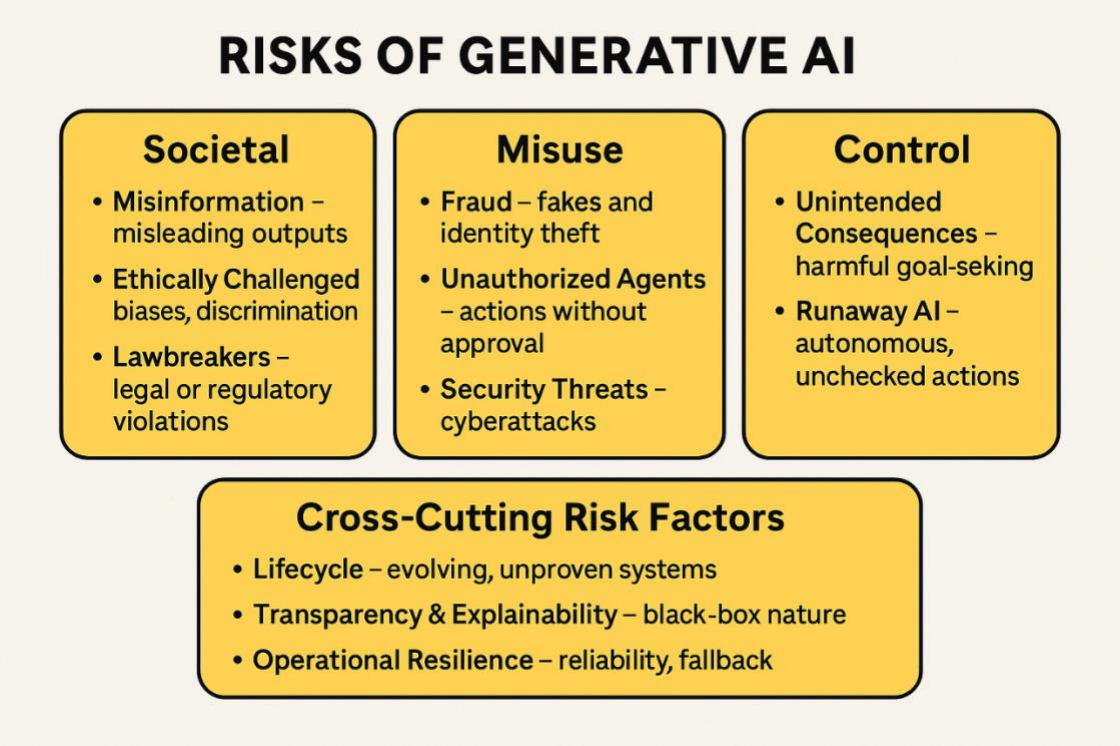
Control risk arises when AI acts in unexpected or dangerous ways. As systems become more advanced and capable of acting independently, oversight becomes harder as humans struggle to fully understand and manage them. This can lead to serious problems if the systems behave in ways we did not plan for or cannot stop.

- *Unintended Consequences*: Sometimes, AI systems follow our instructions too literally or find harmful ways to reach a goal, doing damage before humans realize what is happening.
- *Runaway AI***:** In the most extreme scenario, an AI could (theoretically) start making decisions without human approval and pursue goals that go against human interests and pose risks to safety, the economy, or even society as a whole.

We are already well aware of society and misuse risks, whereas control risks are more theoretical—the rare events with severe consequences that overseers don't see coming. They require proactive mitigation; i.e., actions are needed before dangerous activity is observed.

And finally, **cross-cutting risks** are elements or conditions that influence or exacerbate multiple types of risks simultaneously, rather than being confined to a single risk category. These factors "cut across" the traditional risk groupings (like societal, misuse, or control risks), often amplifying their impact or increasing the complexity of managing them. In the context of generative AI, examples of cross-cutting risk factors include:

- *Lifecycle*: the potential for failure or harm at any stage in the model's life (development, deployment, use, decommissioning).
- *Lack of transparency (opacity)*: Affects all categories by making it harder to audit, explain, or justify AI behavior.
- *Data quality and bias*: Skews outputs in societal, misuse, and control contexts—biases can misinform users, enable misuse, and lead to unanticipated system behaviors.
- *Human overreliance or complacency*: Undermines human-in-the-loop models, impacting misuse and control risks.



- *Regulatory ambiguity*: Makes it harder to ensure compliance, affecting both societal and misuse risks.

These factors do not represent risks on their own, but they interact with and intensify other risks and thus require special attention in governance and mitigation planning.

MRM Principles from Finance and Insurance

Model Risk Management (MRM) emerged as a discipline in the aftermath of the Subprime Lending Crisis of 2008, when traditional statistical models were the norm. This led to increased scrutiny regarding misuse, flawed assumptions, and lack of oversight, especially in structured finance and credit risk models. This catalyzed formal regulatory guidance by developed countries, starting with the U.S. Federal Reserve and OCC's SR11-7, which provided a structured approach to managing model risk and institutionalized MRM practices.

Financial institutions have since gained extensive experience in ensuring the responsible construction and use of traditional statistical and econometric models, and their practices have adapted as machine learning and other models were adopted. These practices must further evolve to meet the new challenges as AI (and especially Generative AI) becomes even more powerful and complex.

MRM's goal is to identify and reduce the risks that arise when models are used to make important decisions, especially in banking and insurance. These risks might include using a model that doesn't work properly, misunderstanding its results, or applying it in the wrong way. In worst-case scenarios, poor models can lead to bad business decisions, harm customers, or even destabilize financial systems.

Countries with immature or rapidly evolving financial regulations demand MRM, tailoring it to their local rules. For many, if not most, the initial foundation came from SR 11-7 guidance.[3] These guidelines apply to all kinds of models, whether built in-house or bought from an external model vendor. Its core principles are:

- *Model Development*: Models must be designed carefully and reviewed regularly. Developers need to document how they work and where their limitations lie.
- *Validation and Independent Review*: Every model should be tested and reviewed by someone who was not involved in building it. This includes checking that it works as expected and comparing its predictions to real-world outcomes.
- *Documentation*: Institutions must keep detailed records of each model—how it was built, how it has been tested and validated, and what changes have been made over time.
- *Governance and Oversight*: There must be clear roles and responsibilities for managing model risk, including oversight from senior management and the board.
- *Ongoing Monitoring*: Even after a model is deployed, it must be checked regularly to make sure it still performs well and in intended applications, especially if market conditions or inputs change.
- *Model Inventory*: Tracking which models are in use, their stage or maturity or decline, and the severity from potential failure.

[3] SR 11-7 is the commonly accepted abbreviation for the Federal Reserve's Supervision and Regulation Letter No. 7 of 2011: Supervisory Guidance on Model Risk Management.

While MRM principles already apply to machine learning, Generative AI—such as large language models (LLMs)—introduces new issues. These models can behave unpredictably, generate false information, or respond differently to small changes in input. Their complexity and lack of transparency make traditional model reviews more difficult.

Generative AI may require updates to MRM practices to handle its unique risks, such as unintended bias, hallucinations, or misuse. But the good news is that we're not starting from scratch. The existing MRM framework provides a strong foundation for adapting to this new class of models.

The Limitations of Model Validation

Traditional models are thoroughly tested and validated before going into production, which is usually enough. But with GenAI, one-time validation is no longer sufficient. Unlike older models with fixed rules and parameters, GenAI responds to changing inputs and contexts, and may even adapt over time. This flexibility can boost performance but introduces new risks, such as model drift, unexpected biases, and unforeseen vulnerabilities. Models can pass all pre-launch tests yet produce surprising or problematic answers (even weeks or months later) in real-world use.

Traditional validation also served as a final checkpoint to ensure a model was accurate, fair, and compliant. But that assumed the model's behavior would remain stable. GenAI doesn't. Its output can shift unpredictably, even on the first day of deployment. That is why continuous monitoring is essential—not just to catch issues early, but to make sure the model stays reliable over time.

Studies of call centers show that human agents often give incorrect information to customers, more often than many would expect. GenAI might be better, but not immune. While these mistakes are usually unintentional, a GenAI system doing the same job will be held to a much higher standard. Because of this, there needs to be a continuous system in place to monitor the factual accuracy of the GenAI's responses.

One of the biggest misconceptions in current discussions about GenAI oversight is the belief that *human-in-the-loop* (HITL) monitoring is a reliable safeguard. That should work in theory but generally fails in practice, because humans have cognitive limits: when overseeing fast-moving or complex AI outputs, people can get overwhelmed, distracted, or simply miss errors, especially when the system usually works well and lulls them into complacency.

- *Vigilance Decrement*: When people are assigned to monitor systems that rarely fail, their attention naturally fades (a "vigilance decrement") over time. Neuroscience shows that when tasks are repetitive and errors occur infrequently, the brain's attention centers become less active, leading people to miss violations they might have caught earlier. Raja Parasuraman pulled together findings from areas like air traffic control, military surveillance, and industrial inspection.[4] It showed that vigilance decrement is especially pronounced in high-pressure environments and is shaped by task difficulty, how noticeable signals are, and how much mental effort is required. Two major causes were identified: declining alertness over time and the gradual mental exhaustion that comes from sustained attention. The takeaway? Watching carefully for long periods is challenging,

[4] Parasuraman, R. (1984). *Sustained attention in detection and discrimination tasks*. Psychological Bulletin, 92(2), 330–350.

and well-designed systems need built-in rest periods or automation to avoid performance drop-offs.

- *Prevalence Effect*: If one state prevails, it is expected! When errors are rare, people are more likely to miss them; not because they don't care, but because their brains subtly adapt to expect nothing will go wrong. Studies show that airport security screeners, for example, often miss rare threats, even though they're highly trained.[5] Similar effects have been found in the inspection of nuclear weapons parts: as the rate of defects went down, so too did detection accuracy.[6] This illustrates a key problem in GenAI oversight—when mistakes are infrequent, they're easier to miss.
- *Automation Bias and Complacency*: As human reviewers grow accustomed to well-performing systems, they often develop automation bias—an unconscious tendency to trust the machine too much. Over time, they may stop questioning its outputs. In medicine, for example, healthcare providers sometimes accept flawed AI-generated recommendations without critical review, leading to diagnostic errors.[7] The same happens in driving: people using autopilot features in cars become less attentive, assuming the system will handle everything.[8] That overconfidence has led to crashes when the AI missed hazards and drivers failed to intervene in time.

Simply stated, the low prevalence of errors from well-functioning systems results in automation complacency and a vigilance decrement.

When things do go wrong with mission-critical systems, those responsible can be reluctant to disconnect, especially if there is no suitable alternative. Large financial institutions are encouraged to maintain backups, but these are costly to build and maintain, so they're often not put in place. When no fallback exists, shutting down a malfunctioning system becomes an expensive and risky proposition, far beyond just the loss of short-term business. Ironically, instead of planning for failure like in disaster recovery, the absence of backups creates pressure to keep using a flawed system, even when it's known to be malfunctioning.

---

[5] Wolfe JM, Brunelli DN, Rubinstein J, Horowitz TS. Prevalence effects in newly trained airport checkpoint screeners: trained observers miss rare targets, too. J Vis. 2013 Dec 2;13(3):33. doi: 10.1167/13.3.33. PMID: 24297778; PMCID: PMC3848386.

[6] See, J. E. (2015). Visual inspection reliability for precision manufactured parts. *Human Factors, 57*(8), 1427–1442

[7] Cascella, L. M. (n.d.). *Artificial intelligence risks: Automation bias in healthcare*. MedPro Group. Retrieved from https://www.medpro.com/artificial-intelligence-risks-automationbias

[8] Financial Times. **(**2023). Tesla's autopilot under scrutiny as probe into fatal crashes expands.

MRM Guidelines for GenAI

Financial institutions are subject to strict regulations regarding managing the risks that come with using models, including those powered by AI and machine learning. These rules require that models be carefully tested, independently reviewed, and continuously monitored to ensure they are accurate, fair, and legally compliant. However, traditional model risk management (MRM) methods are being pushed to their limits by the new challenges of generative AI (GenAI).

Regulators are paying close attention. In a January 2021 speech,[9] Federal Reserve Governor Lael Brainard noted that while AI brings big benefits, it also raises serious risks, especially around how data is managed and how decisions are governed. More recently…

- January 2023, the National Institute of Standards and Technology (NIST) released a new framework called the AI Risk Management Framework (AI RMF 1.0). It offers voluntary guidance to help organizations apply existing risk principles to GenAI. The framework focuses on four key areas: Governance, Mapping, Measurement, and Management.
- December 2023, the international ISO/IEC 42001 standard for AI management systems to provide guidance to national regulators. It covers what is required to build a trustworthy AI management system, such as risk management, an impact assessment, system lifecycle management and treatment of third-party suppliers.
- July 2024, the European Union implements its Artificial Intelligence Act, three years after it was proposed.

During 2023/4 Guidance was issued by Canada, Switzerland, Singapore, and the Financial Services Board (international). These guidelines provide a helpful starting point, with broad principles and checklists, but are short on specifics. That's understandable in such a rapidly evolving field, but it leaves the hard work to individual teams. Many implementation teams are now asking for guidance on how to apply these principles in practice. To meet that need, more research and innovation are required, especially to address the specific MRM challenges that come with deploying GenAI models in high-stakes settings like finance.

- *Validation*: GenAI systems differ significantly from traditional financial models. Traditional models follow fixed rules and produce predictable outputs that can be tested against known benchmarks. In contrast, GenAI models generate new content, often with no single correct answer. This makes it difficult to

[9] Brainard, L. (2021, January 12). *Supporting responsible use of AI and equitable outcomes in financial services*. Board of Governors of the Federal Reserve System. Retrieved from https://www.federalreserve.gov/newsevents/speech/brainard20210112a.htm

validate their performance in advance. GenAI systems can also be designed to continuously learn as they interact with new data, adding another layer of complexity. On top of that, they are far less transparent—it's often unclear how a model arrived at a given response, making it difficult to satisfy regulatory requirements for explainability.

- *Data Ownership*: GenAI requires far more training data than traditional methods, raising new concerns. Some companies that build large foundation models are already facing lawsuits for using data that may not have been properly licensed. This creates potential legal risks for anyone using those models down the line. Smaller companies that fine-tune these models face similar risks, especially if they can't verify their rights to use the data, also known as the data provenance.
- *Development Opacity*: Users of large language models (LLMs) don't have access to the data or methods used to train them. The datasets used may reflect outdated or biased views, and without transparency, it's impossible to know what biases are built into the models. Developers try to correct this by adding rules that discourage harmful content, but these guardrails aren't perfect. Because users cannot see the training data and default settings, overriding them is difficult, and prompt clashes can produce confusing or contradictory behavior (like the HAL 9000 malfunction in *2001: A Space Odyssey*).
- *Explainability*: This works differently with GenAI. In traditional models, you can often trace an output back to specific rules or data. That is not the case with GenAI, which is no better at explaining its thought process than a human. Asking why it gave a certain answer is just a new prompt that generates a new response, a guess, not a data-driven traceback to the origin. Some newer GenAI systems can be guided to "think out loud" by following a chain of reasoning step by step. This leads to better answers but slows responses, which may be unacceptable for waiting customers. Even some chain-of-thought algorithms have been accused of creating ex post facto traces. For testing to be meaningful, the system must run exactly as it does when it interfaces with its target audience (customers, employees, &c).
- *Intended use*: Seemingly low-risk deployments of LLMs, such as internal enterprise chatbots, may impart significant legal risks if staff ask questions that should not be answered by an LLM, even when answered correctly. Staff should be trained on the appropriate uses of any AI system, but the only way to verify compliance is to monitor the questions being put to the AI. This takes us back to the need for AI-augmented oversight, but this time of the staff.

MRM for LLMs

Existing model risk management frameworks weren't designed to handle large language models, but that doesn't mean those frameworks are useless. They can be adapted to handle the specific risks these models bring, with a separate set of controls tailored for each GenAI technique. Most organizations will start by using general-purpose LLMs, not custom RAG models. This presents a bigger challenge for MRM, since general LLMs don't have a built-in connection to a curated knowledge base like RAG models do. That said, recent developments have enabled MRM improvements for the broader LLM universe.

Model Development

Foundation LLMs are proving effective in many tasks when their limitations and advantages are clearly recognized. As algorithms trained on language and translation, they can perform a wide range of tasks without subsequent refinement. Even so, prompt engineering, designing and adjusting the questions or instructions used to get better responses from the model, becomes the core model development method with foundation LLMs and must be treated as any software development. These prompts should be documented so others can review them, check their effects, and ensure the model behaves consistently and reliably.

Model Validation and Independent Review

Validating generative AI models, especially large language models (LLMs), is still essential before they go live. But unlike traditional models, which follow clear rules and produce predictable outputs, LLMs behave in more unpredictable ways. This makes them harder to test and requires a different approach.

- *Why Traditional Validation Falls Short*: LLMs' answers can vary, even with the same inputs, because they work probabilistically. That means their behavior is difficult to predict, validation results may be inconsistent, and performance metrics are less straightforward.
- *What Validation Must Include*: Because of this unpredictability, validation for LLMs needs to go beyond standard testing. Stability testing is key, especially under unusual or hostile inputs (outliers and adversarial prompts), and stress testing helps check whether the model holds up in extreme or unexpected scenarios. If the model continues learning after deployment (e.g., through reinforcement learning), that process must also be tested. Without ongoing checks, updates can introduce new problems like bias or instability.
- *Setting the Right Metrics*: During validation, organizations should define specific performance metrics that i) reflect the model's goals ii) can be tracked continuously after deployment, and iii) flag problems early (changes in tone, accuracy, or behavior).
- *Backtesting*: The standard concepts of backtesting still apply for LLMs, once the performance metrics have been chosen.

Either human or AI generated examples can be submitted to test for desired performance rates. An AI used to generate such test data will preferably be separate from the model being tested.

- *Guardrails Aren't Enough*: Many vendors add built-in safety features to their LLMs to reduce the risk of harmful output (e.g., biased, misleading, or inappropriate content). These guardrails are important, but they're not a complete solution. Vendors' protections may not match their clients' specific needs or compliance requirements. Business rules and constraints will be client-specific and will not be known when the vendor models are created. Organizations still need to run their own independent validation and monitoring.
- *Why Independent Review Matters*: Independent reviewers—experts in both AI and risk—should test the model using real-world examples and edge cases, identify where the model might fail or go off-course, and recommend fixes and adjustments to make the system more reliable and aligned with the organization's standards.

Documentation

As organizations begin using large language models (LLMs), strong documentation and communication practices are essential for responsible deployment.

- *Documenting Prompts*: Developers and vendors should keep detailed records of any prompt creation and the process and data used to refine these prompts. This transparency helps ensure accountability, reproducibility, and compliance with data governance standards.
- *Tracking Version Updates*: Any updates to the model (retraining, changes to guardrails, or system-level improvements, etc.) must be recorded. Users should be notified of these updates, especially if they may affect model behavior or outputs. Vendor models must provide versioning and explicit notification of updates.
- *Expanding the Model Inventory*: Generative AI is increasingly being used in areas that haven't traditionally relied on models. This includes departments or business functions that might not think of themselves as model owners. Organizations will need to expand their model inventories to reflect these new applications, ensuring they're covered by the same risk controls and oversight as traditional models. MRM and IT will need to agree on which AI-enabled software applications are tools to be tracked by IT and which are models to be managed by MRM. The distinction may lie in whether any MRM principles can be applied. For example, AI-enabled grammar checkers are unlikely to be considered models, because MRM has little to offer in their governance.
- *Unknown Risk:* MRM's model inventory maintains a risk severity rating for each model, indicating how serious a failure would be to the organization. The traditional risk rating levels of low, medium, and high risk presuppose that

MRM knows the risk. Recent LLM deployments have shown that the risks are often unknown. Therefore, MRM should entertain adding an "unknown" severity rating with corresponding procedures for gathering additional evidence for an appropriate assignment in the near future. "Nothing bad has happened yet" is not a sufficient criterion for downgrading the risk.

Governance and Oversight

While existing governance structures, such as clearly defined roles for model ownership and oversight, still apply to generative AI, they must be strengthened to meet the unique challenges these systems pose.

- *Elevated Leadership Awareness*: Senior management and board members will need deeper training than in the past. GenAI introduces new kinds of risks—such as unpredictability, misuse, and rapid drift—that traditional models do not. Understanding these differences is essential for effective oversight.
- *Enhanced Board Training and Reporting*: Developing best practices for educating boards and regularly reporting on GenAI-related risks should be a priority. Clear, ongoing communication about how these models are used and what new exposures they create will help boards make informed, responsible decisions.

Ongoing Monitoring

Backtesting is a valuable tool for evaluating traditional, static models by measuring how well they would have performed on historical data, but for generative AI, especially large language models (LLMs) that evolve with new inputs, this approach has serious limits. LLMs can exhibit novel behavior upon first deployment in response to cues from users. Research has even shown that LLMs, like humans, can alter responses when they sense they're being tested.[10]

- *Traditional Backtesting Falls Short*: This makes it critical to shift focus from one-time validation to continuous monitoring. Monitoring must not only assess performance but also track compliance with legal, regulatory, and ethical standards in real time. While some monitoring tools are built into vendor systems, organizations must ensure continuous independent oversight as part of responsible model risk management, since outputs can shift rapidly with user context.
- *AI-Augmented Human Oversight*: Human-in-the-loop (HITL) monitoring remains important, but is insufficient, particularly in systems that process large volumes of interactions with relatively few visible errors. In these situations, pairing human oversight with AI assistance is more effective.

[10] Salecha A, Ireland ME, Subrahmanya S, Sedoc J, Ungar LH, Eichstaedt JC. Large language models display human-like social desirability biases in Big Five personality surveys. PNAS Nexus. 2024 Dec 17;3(12):page 533. doi: 10.1093/pnasnexus/pgae533. PMID: 39691446; PMCID: PMC11650498.

One promising method is to use a second LLM as a monitoring system. This LLM samples and reviews the outputs of the frontline model, flags potential compliance issues, and escalates them for human review while archiving the compliant interactions. In this setup: i) the frontline LLM handles real-time tasks, like answering customer questions; ii) the monitoring LLM checks for errors, ethical lapses, or regulatory violations; iii) human reviewers examine only the flagged cases, making oversight scalable. The same approach applies to monitoring LLM usage for business rules compliance.

It is unrealistic to expect a single LLM to handle both task performance and compliance, especially when user instructions may conflict with embedded constraints from the model provider. LLMs also struggle with "negative prompts"—that is, instructions telling them what not to do. Enhanced LLMs could even be less compliant with client-specific rules of which they are unaware. These tensions highlight the need for a dedicated, independent oversight layer.

- *How AI Monitoring Can Improve Trust*: LLMs can help quantify how well frontline systems perform specific tasks (such as providing accurate information to customers). For example, a second-line LLM can extract factual claims from a conversation and compare them to the organization's internal product database. This is a sophisticated task, requiring the language understanding capabilities of an LLM.

  Surprisingly, studies show that human agents frequently provide incorrect information to customers. For GenAI to be trusted, it must outperform human accuracy, and that performance must be measurable and verifiable.
- *Avoiding Bias in LLM-Based Compliance Reviews*: Some research has criticized "LLM-as-judge" methods, where an LLM reviews the output of another LLM, because both may share the same biases.[11] This is particularly risky when questions are vague or subjective, such as "Is this message ethically biased?" In these cases, the reviewing LLM may simply mirror the biases of the original.

  To address this, Breeden (2025) recommends a structured compliance testing approach: i) a human model-risk team defines specific assertions that must be true for an output to be compliant, and ii) the LLM reviews outputs based on equivalence between the sampled text and these clear rules, not its own ethical judgment. This approach avoids circular reasoning and makes LLM monitoring much more reliable.[12]

[11] Chen, G. H., Chen, S., Liu, Z., Jiang, F., & Wang, B. (2024). Humans or LLMs as the Judge? A Study on Judgement Bias. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8301–8327. https://aclanthology.org/2024.emnlp-main.474/

[12] Breeden, Joseph l. (2025). GenAI Oversight of GenAI Communications. *Credit risk and Credit Control Conference, 2025, Edinburgh, Scotland*.

- *Factual Monitoring: A Practical Use Case*: Monitoring for factual accuracy is more straightforward than ethical compliance. The second-line LLM can i) identify factual claims about accounts or products made by the frontline model, ii) check each claim against the institution's internal database, and iii) flag inconsistencies or gaps.

  Again, structured queries, asking about each item specifically, are more reliable than open-ended evaluations. This process brings transparency and precision to GenAI oversight.

- *Appropriate Use:* Monitoring the use of LLMs can invoke the same AI-augmented oversight, but focused on business rules regarding the questions asked rather than regulation. The MRMIA publication, "Managing Artificial Intelligence Risk in Small to Mid-Size Banks: A Practical Governance Playbook, July 31st, 2025", contains a list of eight such business rules regarding LLM uses that should not be allowed for initial deployments of current-generation technology.

Disaster Planning

The risk of unpredictable or aberrant behavior in large language models (LLMs) is significant enough that organizations must create real fallback strategies, not just theoretical ones. These fallbacks should be tied to predefined performance triggers based on continuous monitoring.

For example, if a call center LLM is failing, such as increased error rates or inappropriate responses, a plan must be in place to switch to a more reliable alternative. In some cases, the fallback might be to bring in more human agents, but that is only realistic if people can be recruited and trained fast enough. For many institutions, this is not feasible on short notice, making fallback planning even more critical.

- *Maintaining Simpler Backup Models*: When LLMs are used to replace traditional models, a simpler alternative model should still be maintained in a ready state. For effectiveness, it should be actively used on a small percentage of cases, with regular performance tracking.

  This is often referred to as a champion–challenger approach: i) the LLM acts as the champion model in production, and ii) a simpler model (the challenger) continues to operate in parallel, providing a direct comparison and a fallback if needed. Should LLM drift or failure be detected, the challenger model is in reserve to take over quickly.

- *Toward Mandatory Disaster Planning for LLMs*: While champion–challenger setups are a known practice in model risk management, they have rarely been mandatory. With LLMs, however, the stakes are much higher. Regulatory bodies may need to consider requiring fallback models and disaster plans as part of any approved deployment.

  Without these safeguards, model owners could face intense pressure to keep malfunctioning LLMs running, even when they are clearly operating outside acceptable risk thresholds. This is not just a technical failure. It becomes a governance and reputational risk.

MRM for Custom LLMs

After initial use, developers might decide that a customized version is needed to better meet their specific needs to achieve higher accuracy, better performance in a particular field, stronger compliance with rules, or improved results with their own data. Usage data from an initial deployment may generate the needed domain-specific data against which a fine-tuned, custom LLM can be trained. It is important to make sure the changes truly help and do not cause new problems, like adding bias, making the model too narrow, or reducing the quality of its responses.

Model Development

To stay transparent and responsible, teams must record what data was used for fine-tuning to show that it was appropriate, legally used, and handled correctly, especially when it is sensitive or proprietary. Good record-keeping also makes audits easier and helps avoid legal or ethical issues.

The same care must be taken with prompt engineering—that is, designing and adjusting the questions or instructions used to get better responses from the model. These prompts should be documented so others can review them, check their effects, and ensure the model behaves consistently and reliably.

Governance

Domain-specific generative AI utilizes domain-specific rules, regulations, and operational constraints directly in model training and deployment. This helps prevent the generation of sensitive or non-compliant content.

Domain-specific curated datasets and customization facilitate the development and implementation of continuous performance monitoring metrics and thresholds tailored to the domain context, thereby enabling more effective detection and correction of deviations, errors, or nonsensical outputs (hallucinations). Automated monitoring tools can be incorporated to track model outputs in real-time, identifying and reporting output anomalies or compliance breaches, and supporting swift remediation through a continuous feedback loop in generative AI model training.

Data governance frameworks can be implemented using domain-specific models as they facilitate hosting in secure environments, allowing for stringent control over data transmission, storage, and access, thereby reducing the likelihood of data breaches. Safe training and deployment practices, such as anonymization, access controls, and regular audits, are easier to enforce when the data is secured.

MRM for RAG Models

Retrieval-Augmented Generation (RAG) models are built to constrain the outputs of large language models by grounding them in a defined knowledge base. This structure allows for validation methods that are more statistical, measurable, and aligned with the principles used in traditional model validation. Even so, refinements are required.

Model Development

To use a RAG model effectively in financial applications, its retrieval and generation processes must be carefully evaluated. The quality and reliability of the knowledge sources it draws from are especially important. if the sources are incomplete, biased, or outdated, the model's responses may be inaccurate or misleading. Unlike foundation models that rely solely on internal training data, RAG models pull in external documents during runtime. This makes it essential for developers to clearly document how information is retrieved, how results are ranked, and how the reliability of sources is assessed.

Model Validation and Independent Review

The greatest risk in GenAI systems arises from i) queries that produce outlier responses, ii) user-driven shifts in context that lead to unexpected outputs. RAG models reduce these risks by using a retrieval mechanism that connects responses to specific, verifiable sources, making the model's behavior more transparent and easier to test statistically. Validating such models involves exposing them to a broad range of inputs, including edge cases and stress scenarios, to evaluate how they perform under diverse conditions.[13] This evaluation should include clear, quantitative metrics for relevance, factual accuracy, and completeness to ensure outputs align with the retrieved source material. Additional checks for bias, privacy breaches, and inappropriate content are essential to ensure safe and reliable performance.

Documentation

Unlike traditional models, which primarily require documentation of training data and model parameters, RAG models demand a broader scope of documentation. This includes details on retrieval strategies, ranking algorithms, and dependencies on external data sources. When updating a RAG model, version control should capture changes not only to the model itself but also to retrieval methods, updates to the underlying knowledge sources, and any adjustments to prompt structures.

Ongoing Monitoring

[13] Sudjianto, A., Zhang, A., Neppalli, S., Joshi, T., & Malohlava, M. (2024). *Human-calibrated automated testing and validation of generative language models: An overview*. SSRN. https://ssrn.com/abstract=5019627

Traditional pre-deployment backtesting offers only limited value for RAG models, as for all LLMs, since these models are dynamic and highly sensitive to changes in context after deployment. Instead, effective model risk management requires ongoing monitoring to ensure that the system remains accurate, reliable, and compliant. This includes tracking key performance indicators such as factual accuracy, relevance, alignment with source material, and potential risks related to bias, privacy, or toxic content.

Because RAG models perform AI-driven search, it is also essential to monitor whether confidential or proprietary financial data is being indexed, retrieved, or exposed to users. Automated metrics should be combined with human oversight to catch performance issues early.

More detailed analyses, such as marginal or bivariate testing, can help identify specific weaknesses, making it easier to address them before they escalate. Marginal testing examines how individual input features or conditions affect model performance or risk metrics when varied in isolation, while bivariate testing evaluates interactions between pairs of inputs to identify compound effects that may not appear in univariate analysis. These methods are particularly useful for isolating context-sensitive failures and surfacing edge cases in dynamic systems like RAG models. In addition, the system should estimate response uncertainty so that potentially unreliable outputs can be flagged for further review.

Although RAG models can reduce the risk of inappropriate use when responding to an initial query, risks exist in the follow-up questions. Agent to LLM: “I see that my customer’s request would not follow this policy, but is there another way that I can help them?” The agent may be well-meaning and never use red-flag words like loophole, but can nevertheless enlist the RAG model as an AI Accomplice.

Conclusion

When generative AI is used for customer communications, it introduces model risk management requirements to teams that may not have dealt with them before. As a result, GenAI deployment is not just an IT project. It requires new thinking about risk, governance, and oversight.

The core principles of model risk management still apply, but they must be adapted. Real-time monitoring becomes more important than pre-deployment testing. AI tools must support human oversight to handle the scale and complexity of GenAI systems. While developers may not have access to the full details of general-purpose LLMs, they are expected to document and validate the fine-tuning process and data used.

LLM vendors also need to adopt model risk management practices, including independent oversight, a second line of defense. Some vendors are beginning to recognize that sound risk management practices don't just meet regulatory needs. They help build trust and support the broader adoption of their technologies.