DEEP FUTURE
ANALYTICS

MARCH 08, 2025

# Effective Generative AI Model Risk Management

7

**Joseph L. Breeden, PhD**

CEO, Deep Future Analytics LLC, www.deepfutureananlytics.ai

President, Model Risk Managers' International Association

Deep Future Analytics LLC
1600 Lena St., Suite E3, Santa Fe, NM 87505 USA

breeden@deepfutureanalytics.com
1-505-670-7670

WWW.DEEPFUTUREANALYTICS.COM

# TABLE OF CONTENTS

# Executive Summary

This paper provides a foundational framework for the model risk management (MRM) of Large Language Models (LLMs) as applied in finance and insurance. While existing MRM principles remain relevant, the interpretation, metrics, and prioritization must be reconsidered. Unlike traditional predictive models, LLMs generate novel content dynamically, making pre-deployment validation insufficient and necessitating continuous monitoring as the primary risk control mechanism.

Traditional MRM frameworks assume a level of model stability that does not apply to generative AI (GenAI). Conventional models are validated against fixed datasets, whereas LLMs may respond in unexpected ways to outliers and changing contexts. Instead of relying on static validation, institutions must shift toward real-time monitoring. The predominant proposal is to employ human-in-the-loop (HITL) monitoring of AI systems. Unfortunately, psychology studies have proven that humans monitoring low-event rate situations become desensitized, rarely to spot unexpected failures, and become reliant on automated answers. A better approach is to leverage AI-assisted compliance monitoring to track accuracy and ensure regulatory and ethical compliance.

Financial institutions deploying GenAI models often have limited visibility into their underlying architectures, training data, and embedded biases. Since many LLMs are primarily developed by external vendors, their opacity complicates compliance with existing model documentation and explainability requirements. A refined approach will be needed where institutions document fine-tuning data, prompt engineering techniques, and post-training interventions rather than attempting to audit proprietary foundation models.

Monitoring is only useful if management has the will and ability to respond. AI models bear a greater risk of catastrophic failure. As such, disaster planning must establish metrics and limits by which the AI will be disconnected in favor of a clear fallback plan. Fallbacks must be viable on emergency timescales, meaning that if call centers are to replace AI with human agents, those agents must be trained and available on demand. When the fallback is an alternative model, that model must be deployed and monitored alongside the primary model, as in a champion-challenger approach. This shift is critical, as organizations are likely to be reluctant to disable malfunctioning AI systems due to operational dependencies. A champion-challenger approach is not new to MRM, but it has not been mandatory except in the most mission critical situations at large institutions. The failure risk of LLMs necessitates considering monitoring and fallback solutions as a standard part of LLM deployment.

Monitoring will take priority over validation for obvious reasons, but LLM monitoring is an emerging field. Retrieval-augmented generation (RAG) models constrain and simplify the problem and align better with a traditional validation process, but general LLMs are less amenable to traditional methods. For general LLMs, the most promising current approach is to deploy LLM oversight of LLM communications. Previous research has suggested that such LLM-as-judge systems may share the biases of the frontline system, and thus fail to spot problems. That mostly occurs from a flawed problem statement. Rather than asking the LLM overseer if a message is “ethical”, human operators should create lists of assertions that should be true if a message is compliant with applicable regulations and ethical guidelines. The LLM overseer compares each sampled frontline LLM communication with each assertion to determine the level of agreement. The compliant messages are archived, but questionable messages can be promoted to a dashboard for human review. This AI-augmented human-in-the-loop oversight is a viable solution to both the psychological weaknesses of humans and embedded biases of LLM systems, and results in tracking metrics as needed for performance assessment.

From this review, we see that the principles of model risk management still apply to LLMs, but the application of these principles must be deeply rethought. By shifting the emphasis from pre-deployment validation to continuous oversight, leveraging AI-assisted compliance mechanisms, and ensuring robust fallback planning, financial institutions can integrate GenAI into critical operations while mitigating its unique risks.

# Introduction

Artificial Intelligence (AI) is rapidly transforming industries, economies, and governance structures worldwide. Generative AI (GenAI), in particular, represents a paradigm shift, with its ability to produce human-like text and images at unprecedented speed and scale. While this technological advancement offers significant opportunities, it also presents a complex array of risks that must be proactively addressed.

Generative AI poses unique risks not found in the previous decade's machine learning (ML) algorithms in use across most industries today. Algorithms that could be called "optimization machine learning" provide solutions to highly nonlinear problems, often processing vast amounts of data. However, they are single-purpose algorithms optimized to solve a specific problem and are already overseen by the basic principles of model risk management (MRM). GenAI was not designed to solve a specific problem, but rather to forecast the next token (often a word) in a string of tokens across the broadest possible range of contexts. This training approach is distinct from optimization ML, resulting in its extraordinary generality. Consequently, GenAI is much more difficult to manage via existing model risk management principles.

The principles of model risk management still apply, but the interpretation and emphasis will need to change dramatically. Even more difficult is that different flavors of generative AI will require specific interpretations of the guidelines. This white paper provides a review of possible risks of Generative AI systems and suggests a basis for model risk managers to assess and mitigate these risks specifically for general Large Language Models (LLM) and Retrieval-Augmented Generation (RAG).

# Defining Generative AI

Generative Artificial Intelligence (GenAI) encompasses a broad category of techniques capable of producing content, including text, images, audio, video, or structured data, based on patterns extracted from extensive training datasets. At their core, generative AI systems learn to approximate the underlying distribution of input data, allowing them to generate outputs that mirror or recombine the styles, structures, or characteristics of their training examples. Among the most prominent forms of generative models are Large Language Models (LLMs), Retrieval Augmented Generation (RAG), Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models, each employing distinct computational techniques to create coherent and contextually relevant results. In financial services, text-based methods such as general LLMs and RAG are of primary interest.

## Large Language Models (LLMs)

LLMs are trained on vast amounts of text data and use deep neural networks (typically transformers) to generate human-like text. LLMs are trained to sequentially predict the next word in order to generate the most desirable response to an input prompt. Temperature controls within the API control the amount of variability from one session to another, but small changes in prompt design, context, or making implied instructions explicit can significantly alter the outputs. They are most often applied to text completion, summarization, translation, and conversational AI. While all LLMs are foundation models due to their broad applicability and pre-training on large text corpora, foundation models are generally meant to include a broader category of multimodal, vision, and speech models in addition to LLMs.

## Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) models are a specialized class of generative language models designed to improve the accuracy, relevance, and factual consistency of generated outputs by integrating an external retrieval mechanism.[1] Unlike standard generative models that rely solely on their pre-trained knowledge, RAG models dynamically fetch relevant information from a predefined corpus, such as databases or document repositories, before generating responses. This process ensures that the generated content remains grounded in approved sources, reducing the risk of hallucination. By incorporating retrieval into the generation process, RAG models provide greater transparency, adaptability, and domain-specific accuracy, making them particularly useful in applications such as question answering, customer support, and legal or financial document summarization. However, the creation of RAG models involves significant additional cost for the business and is only applicable where in-house data sources are available.

# 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.

## Societal Risks: Legal, Regulatory, and Ethical Failures

Societal risks arise from the use of GenAI creating misalignment with legal, regulatory, and ethical standards. Even a well-behaved AI that has been trained on data sourced internationally may not comply with local standards. AI governance will be required to assure that AI systems adhere to these standards at least as well as their human counterparts. This is 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 metrics of effectiveness.

- ***Misinformation***: Failing to accurately provide the information requested by consumers.
- ***Legal and Regulatory Violations:*** Communications or actions that fail to adhere to existing laws and regulations.
- ***Misalignment with Ethical Standards:*** AI communications that embed biases, reinforce discrimination, or produce outcomes that contradict ethical norms asserted by the institutions deploying these systems.

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

LLMs are trained on 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.

[1] 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

## Misuse Risks: Malicious Actors

AI's capabilities make it a potent tool for both beneficial and harmful uses. Without strong safeguards, AI can be exploited for misinformation, fraud, and cyberattacks.

- ***Fraud:*** GenAI is already being deployed in identity theft and fake identity generation, highlighting the need for confirmed identity systems
- ***Unauthorized Agents:*** The self-direction in Agentic AI brings the risk of conducting unauthorized transactions. This mirrors the existing problem of verifying authority for transactions by humans.
- ***Security Threats:*** AI-driven cyberattacks pose risks to digital infrastructure, financial institutions, and critical national security systems.

Misuse Risks are serious but generally stem from bad actors exploiting AI, rather than inherent AI misbehavior as in societal risks. Defending against such attacks places new burdens on information technology infrastructure and requires new thinking around confirmed identities for both humans and AI agents.

## Control Risks: Loss of Oversight and Unintended Consequences

Control risks arise from the difficulty of predicting and regulating AI behavior, particularly as AI systems grow in complexity and autonomy. The ability to maintain human oversight is critical in preventing unintended or catastrophic outcomes.

- ***Unintended Consequences:*** AI systems can misinterpret human instructions or optimize in harmful ways faster than humans can react.
- ***Runaway AI:*** AI that escapes human control and autonomously pursues misaligned goals leading to existential threats or severe systemic disruptions.

Control Risks require proactive mitigation. Societal risks and misuse risks are already prevalent, whereas control risks are more theoretical. However, the extreme severity of control risks requires that actions be taken long before dangerous activity is observed.

# Model Risk Management Principles from Finance and Insurance

We have over a decade of experience in combatting societal risks from the use of statistical models through the adoption of model risk management principles. These methods were generalized to accommodate the nonlinearity and relative opacity of machine learning models. Generative AI presents unique challenges for model risk management.

Model Risk Management (MRM) has become an essential component of financial oversight in countries with mature or rapidly growing banking and insurance industries. While each jurisdiction tailors its approach to local regulatory and business environments, the foundational principles of MRM stem from guidance first formalized in the United States under the Federal Reserve's SR 11-7. These guidelines provide a comprehensive framework for managing risks associated with models used by regulated institutions, whether internally developed or externally acquired. The ultimate goal of MRM is to safeguard financial systems and society from harm arising from poorly constructed, incorrectly implemented, misused, or malfunctioning models.

The MRM guidelines define model risk as the potential for adverse consequences arising from errors in model design, implementation, or inappropriate use. Model risk can emerge from inaccurate assumptions, flawed methodologies, coding errors, or misinterpretation of results. To address these concerns, SR 11-7 outlines a framework centered on effective governance, rigorous validation, and comprehensive risk management practices.

***Model Development***: Institutions must establish strong model development practices. The models should be appropriate for their intended use, with regular assessment of their limitations and assumptions.

***Model Validation and Independent Review***: A fundamental aspect of MRM is model validation, which involves assessing conceptual soundness, conducting empirical testing, and comparing model outputs to actual outcomes. Validation should be performed by qualified personnel independent of the model development process.

***Documentation***: Institutions must maintain thorough documentation of model development, validation, and changes over time. This ensures transparency, facilitates audits, and supports regulatory reviews. Model inventories should be maintained, cataloging all models used by the institution along with their associated risks and limitations.

***Governance and Oversight***: Strong MRM requires clear governance structures, with defined roles and responsibilities for model risk oversight. Senior management and the board of directors are expected to establish and enforce model risk policies, allocate resources for validation, and ensure compliance with regulatory expectations.

***Ongoing Monitoring***: Banks must continuously monitor model performance to verify the accuracy and reliability of model outputs. This includes identifying potential biases, recalibrating as needed, and updating models to reflect changes in market conditions or underlying assumptions.

SR 11-7 emphasizes that model risk should be managed like other types of financial risks, requiring a structured risk management approach rather than reliance on model outputs without critical assessment. Banking institutions must foster a culture of model risk awareness, ensure independent challenge to model assumptions and integrate model risk considerations into overall risk management frameworks.

# MRM Guidelines for GenAI

Financial institutions operate under regulatory frameworks for model risk management that also apply to the use of AI and machine learning models. These regulations require that all models undergo rigorous validation, independent testing, and continuous monitoring to ensure accuracy, fairness, and compliance with banking laws. Traditional MRM approaches are straining to adapt to GenAI implementations.

Recognizing the challenges posed by AI, the Federal Reserve has emphasized the importance of robust governance and risk management practices. In a speech delivered in January 2021,[2] Governor Lael Brainard highlighted that while AI offers significant benefits, it also introduces model risks, particularly concerning data management and governance. The National Institute of Standards and Technology (NIST) also released its AI Risk Management Framework (AI RMF 1.0) in January 2023. Their voluntary guidance attempts to adapt general model risk management principles to GenAI. The framework is built around four key functions: Governance, Mapping, Measurement, and Management.

The NIST guidelines provide principles and checklists, but as with most other regulatory guidance, they have few specifics. Principles-based guidance is appropriate in such a rapidly changing field, but the hard work of specific application is left to implementation teams that are asking for suggestions. Much more work needs to be done to develop techniques specifically designed for the unique challenges of GenAI.

# Unique MRM Challenges with GenAI

Unlike conventional financial models, GenAI systems are dynamic, continuously adapting based on new data, making validation more complex. The opacity of such models challenges interpretability, complicating compliance with explainability requirements. Additionally, current MRM frameworks do not explicitly address generative capabilities, where models produce novel outputs that may not have predefined correctness measures or were simply never encountered during validation testing. New risk management practices must therefore be developed to extend MRM's scope to the governance of GenAI.

## Data Ownership

Generative AI requires dramatically more training data than prior modeling techniques. Having the legal right to use data for modeling is not a new requirement. Nevertheless, providers of foundation models are engaged in various lawsuits because of their expansive use of accessible data. This creates downstream legal risks for users of these models. Smaller vendors providing fine-tuned models bear similar risks if they cannot verify their rights to use their data, but have fewer resources to defend against accusations of misuse.

## Little Transparency in Development

The development data and process for large language models are not available to users. Historic language and documents were routinely biased by contemporary standards. By not knowing the data upon which LLMs are trained, users cannot know what biases are embedded within the models. Model developers may layer instructions on top of the base model to discourage inappropriate language, but such guardrails are never perfect and unavailable for scrutiny by users.

[2] 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

Because users cannot see the training data and default settings, overriding them can be quite difficult. From this lack of visibility, users cannot know when given prompts will conflict with internal instructions, creating Hal 9000-style malfunctions of conflicting instructions.

## A Shortage of Explainability

Explainability also takes on a different meaning with GenAI. No single output can be traced back to a responsible structure or set of parameters. Asking the GenAI to explain an output is an *ex post facto* exercise. The GenAI can no more see its inner workings than humans can. Asking for an explanation is a new query with the previous output as part of the data. The provided answer is a new, generative response with no more explainability than the original response.

Chain-of-thought reasoning is the latest enhancement to GenAI systems. Asking the algorithm to follow a sequence of steps to obtain an answer can lead to better outputs, but it is too computationally expensive to generate and archive for the millions of responses requested from a customer-facing GenAI system. The added response time is likely to be unacceptable to waiting consumers. Further, chain-of-thought reasoning cannot be turned on just for running diagnostics, because it changes the output from the algorithm. For testing, the GenAI system much be run exactly as consumers will interact with it.

## The Limitations of Model Validation

Because of the adaptive nature of GenAI, the traditional focus of model risk management on validation before deployment must shift to continuous monitoring. Validation on a snapshot date is a necessary but insufficient condition for ensuring reliable performance once the system is in production. Unlike conventional models, which operate on fixed parameters and predefined relationships, GenAI systems respond dynamically to new inputs, changing contexts, and iterative learning processes. This adaptability introduces both opportunities for improved performance and significant risks related to model drift, unintended biases, and unforeseen vulnerabilities. A model that passes all validation checks on a given day may still generate responses that deviate from intended behavior weeks or months later, or even when first deployed and encountering unexpected consumer inquiries.

In traditional model risk management, validation serves as a critical gatekeeping function before deployment, providing assurance that a model meets regulatory, operational, and ethical standards. However, this paradigm assumes relative model stability, where a validated model remains fundamentally unchanged in structure and behavior over time. GenAI systems, on the other hand, require continuous evaluation because their outputs can shift unpredictably in response to new data. Given this variability, a one-time validation process offers only a temporary guarantee of performance, making it inadequate as a long-term risk control measure.

## Misinforming Consumers

Call center studies show that human agents provide incorrect information to consumers in response to inquiries at a surprisingly high rate. These errors are assumed to be unintentional, but a GenAI system filling the same role will have a much higher performance standard. Consequently, an ongoing system for monitoring the factual accuracy of the GenAI system is required.

# Failure of Human-in-the-Loop Monitoring

One of the great fallacies in current discussions around GenAI oversight is the reliance on Human-in-the-Loop (HITL) control. Human-in-the-loop (HITL) monitoring is intended to ensure oversight and intervention when automated systems make errors. However, several factors contribute to its failure in critical applications. These failures typically arise due to the cognitive limitations of humans.

## The Vigilance Decrement Effect

Studies have shown that humans tasked with monitoring low-failure-rate systems exhibit a "vigilance decrement"—a decline in detection accuracy over time. Neurological research reveals that as tasks become repetitive and errors occur only rarely, human monitors experience reduced neural activity in attention-related areas, leading to missed violations.

Parasuraman's meta-analysis synthesized findings from numerous vigilance studies across domains such as air traffic control, industrial inspection, and military surveillance.[3] The review established that vigilance decrement is more pronounced in high-workload environments and is influenced by factors such as signal salience, task complexity, and cognitive demands. The study also identified two primary explanations for vigilance decrement: a decline in arousal over time and a progressive depletion of cognitive resources. This work reinforced the idea that sustained attention is mentally taxing and that task design must incorporate breaks or automation to mitigate performance declines.

## The Prevalence Effect in Oversight

The "prevalence effect" describes how infrequent errors are more likely to be missed by human reviewers. Studies have demonstrated that when humans expect an error rate to be low, they subconsciously lower their detection efforts, increasing the probability of failing to identify the error. For example, studies have examined the effectiveness of Transportation Security Administration (TSA) screeners over time[4] finding that screeners are more likely to miss rare targets. A study of the visual inspection of nuclear weapons parts had similar results.[5] The researchers found that human inspectors' ability to detect defects decreased as the defect rate lowered, highlighting the challenges in maintaining high inspection accuracy in low-prevalence scenarios.

## Automation Bias and Complacency

HITL monitors also exhibit "automation bias," where they develop undue trust in AI-generated outputs over time. As AI-generated communications consistently pass compliance checks, human reviewers tend to reduce scrutiny, assuming the system is functioning correctly. In medical settings, healthcare providers become overly dependent on automated systems,

[3] Parasuraman, R. (1984). *Sustained attention in detection and discrimination tasks.* Psychological Bulletin, 92(2), 330–350.
[4] 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.
[5] See, J. E. (2015). Visual inspection reliability for precision manufactured parts. *Human Factors, 57*(8), 1427–1442

accepting recommendations without critical evaluation.[6] This over-reliance can lead to diagnostic errors, especially if the system provides incorrect suggestions. Drivers using advanced driver-assistance systems have exhibited reduced attention to the driving task, assuming the system can handle all situations.[7] This complacency has led to accidents when the system failed to recognize obstacles or hazards, and drivers did not intervene in time.

## Reluctance to Disconnect

The largest financial institutions are encouraged to have backup plans for their most mission-critical systems, but this is expensive, and such systems are not widely maintained. When no fallback exists, the decision to shut down a malfunctioning system becomes extremely expensive, beyond the short-term loss of business. This is the opposite of disaster planning, causing a tendency to continue using a system well beyond the point at which it is identified to be malfunctioning.

# Model Risk Management for Large-Language Models

The integration of LLMs into financial services does not necessitate a complete overhaul of model risk management frameworks, but it does require new adaptations to account for the unique risks these models introduce. Each technique requires a separate set of MRM controls tailored to that approach.

For most organizations, their first attempt to employ GenAI will be with general LLMs rather than creating custom RAG models. General LLMs present a greater challenge for model risk management, because they lack the explicit retrieval mechanism tied to a curated knowledgebase as with RAG. Nevertheless, recent developments also allow for effective model risk management of the general case.

## *Model Development*

After initial use of a large language model (LLM), developers may determine that a custom-trained model is necessary to better align with the specific requirements of their application. This decision may stem from the need for improved accuracy, domain specialization, compliance with regulatory standards, or enhanced performance on proprietary datasets. Any fine-tuning of an LLM should be carefully assessed for its suitability to the problem at hand, ensuring that adjustments do not introduce biases, reduce generalization capabilities, or compromise the integrity of the model's responses.

To maintain transparency and accountability, developers should retain fine-tuning data, not only to verify the appropriateness of modifications but also to establish proof of ownership and compliance with data governance standards. This record-keeping is particularly important when dealing with proprietary or sensitive data, as it enables auditability and

[6] Cascella, L. M. (n.d.). *Artificial intelligence risks: Automation bias in healthcare.* MedPro Group. Retrieved from https://www.medpro.com/artificial-intelligence-risks-automationbias
[7] Financial Times. **(**2023). Tesla's autopilot under scrutiny as probe into fatal crashes expands.

mitigates legal and ethical risks. Additionally, prompt engineering—the process of crafting and refining inputs to guide the model's behavior—should be documented and made available for review. This ensures that prompt modifications can be evaluated for their impact on model outputs and that best practices are followed to achieve consistency and reliability in responses.

## Model Validation and Independent Review

Pre-deployment validation remains a critical step for generative AI models, ensuring that they operate reliably and within acceptable risk parameters. However, unlike traditional optimization models, where validation methods are well-established and often yield reproducible performance metrics, LLMs introduce an inherent level of unpredictability due to their probabilistic nature. This makes their validation fundamentally less reliable and more complex. As a result, validation processes should place a strong emphasis on stability testing, particularly under outlier interactions and adversarial inputs. Given that generative AI models are designed to generate novel responses rather than optimize for a fixed objective function, stress testing under extreme or unexpected conditions is necessary to assess robustness.

If a continuous updating or reinforcement learning process is employed, this mechanism must be tested as part of the pre-deployment validation to ensure that updates do not introduce drift, unintended biases, or destabilization over time. Performance metrics should be defined during the validation process, with a focus on those that can be deployed for ongoing monitoring and anomaly detection in production. This proactive approach helps organizations identify potential issues early and adjust the model accordingly.

While vendors of LLM systems are incorporating guardrails into their models to mitigate risks such as misinformation, bias, and inappropriate content generation, independent review remains an essential component of model risk management. Relying solely on vendor-implemented safeguards is insufficient, as these measures may not fully align with an organization's specific risk appetite or regulatory obligations. Therefore, independent validation should be conducted by experts with a deep understanding of both the strengths and limitations of LLM. These experts should assess the model's behavior across a range of real-world scenarios, evaluate potential failure modes, and recommend necessary adjustments to enhance reliability and compliance.

## Documentation

LLM developers and vendors should document the source of fine-tuning data, use of such data, and prompt engineering. Version updates need to be documented and users notified of any such updates. The model inventory will need to expand to include novel uses of GenAI in areas that have not previously used models and may not think of themselves as model owners.

## Governance and Oversight

The existing governance structures around defined roles for ownership and oversight still apply. However, senior management and the board of directors will require detailed training around the sensitivities and risks of GenAI, indeed more so than with traditional models. Developing best practices in board training and reporting of GenAI model risks takes on heightened importance.

## *Ongoing Monitoring*

Traditional backtesting, while invaluable for assessing the historical accuracy of static models, offers limited insight into the performance of LLMs that continuously learn and adapt from new data. For LLMs, the reliability of backtesting diminishes because the models can develop new patterns of behavior not present during the initial validation phase or may respond in unexpected ways to changes in context initiated by consumers. In fact, research has shown that LLMs can alter their responses when they detect that they are being tested.[8] This calls for a refocusing on a robust system of continuous monitoring that not only tracks task performance but also rigorously evaluates compliance with legal, regulatory, and ethical standards in real-time. Although vendors may build monitoring into their systems, users will need independent monitoring as an essential part of model risk management. Considering how dramatically GenAI outputs can change with changes in context, users will need to develop their own testing protocols.

Human-in-the-loop oversight is an essential part of GenAI model monitoring, but to be effective on low failure rate, high bandwidth systems, AI-augmented review will likely be required. For example, LLM can be deployed to sample communications and actions from an operational AI system, scan them for compliance with regulatory and ethical standards, and promote only the questionable content for review by a human manager. In this case, a GenAI system is designed specifically to serve as the second-line oversight of a front-line AI system. We should not expect an operational AI to succeed at incorporating all necessary constraints simultaneously with performing its primary objectives. Such conflicts can already be observed between default settings provided by the foundation model developers and the prompts provided by users. In addition, the current generation of GenAI models are well-known for difficulties incorporating negative commands – what not to do. They cannot be expected to properly balance such competing objectives in the various contexts created by users and consumers.

Such AI monitors can quantify the accuracy rate of the frontline system in situations such as providing requested information to customers. Extracting factual statements from a conversation and comparing to stored product information requires sophistication such as is found in large language models. Testing of human communications shows surprisingly high failure rates, but GenAI will only be trusted if it provably exceeds human performance.

Previous research studies have suggested that LLM-as-judge approaches will share the same biases as the frontline LLM.[9] This is true when the testing is unstructured, “Is this message ethically biased?” Generic questioning leaves the interpretation of regulation and ethics to the LLM, thus creating the possibility for a shared bias. The approach described in Breeden

[8] 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):pgae533. doi: 10.1093/pnasnexus/pgae533. PMID: 39691446; PMCID: PMC11650498.

[9] 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/

(2025)[10] requires the human MRM team to create a specific set of assertions that must be true for in order for the message to be compliant. This breaks the circular reasoning problem, allowing for effective compliance monitoring.

LLM monitoring for factual accuracy is more straightforward. The second-line LLM must identify any specific account or product terms from the communications and compare those against the institution's internal database. Rather than leave this as an open-ended problem, but best approach is again to query the message specifically for each verifiable informational item and compare successively to institutional databases.

## Disaster Planning

The risk of unexpected, aberrant behavior in LLMs is high enough that real fallback plans must be created, with predefined triggers according to performance metrics from monitoring systems. For a call center LLM, a fallback could be hiring more human agents, but that it only practical if humans can be hired and trained on the timescales required.

Where LLMs are replacing other models, simpler alternative models need to be kept in a state of readiness. This usually means that the fallback model is actually in use for a very small percentage of tasks, with ongoing monitoring, just as one would do with any challenger model.

Champion – challenger approaches are not new, but they have not been mandatory as part of model risk management. With LLMs, regulators will need to consider making tangible disaster planning or challenger models a requirement for deployment of LLMs. Otherwise, model owners will face extreme pressure to leave malfunctioning LLMs in operation even when monitoring shows problems beyond acceptable limits.

# Model Risk Management for Retrieval-Augmented Generation Models

RAG models are designed to restrict the LLM in ways that allow for validation that is more statistical and closer to the philosophy of validating traditional models.

### *Model Development*

Ensuring that a RAG model is suitable for financial applications requires careful scrutiny of its retrieval mechanism and generative capabilities. The selection of knowledge sources is critical, as incomplete, biased, or outdated information can undermine the reliability of generated responses. Unlike foundation models, which generate responses from internalized knowledge, RAG models dynamically fetch documents, meaning developers must document retrieval logic, ranking mechanisms, and source reliability metrics.

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

## Model Validation and Independent Review

The greatest risk in any GenAI approach comes from queries that trigger outlier responses or user-created changes in context leading to unexpected results. The retrieval mechanism provides traceability and a connection to ground truth that allows for a more statistical testing process. The primary validation approach is to test the model under a wide range of inputs including stressed situations, ensuring that the model is evaluated across diverse scenarios.[11] The evaluation process should apply quantitative measures to assess key attributes such as relevance, factual accuracy, and completeness of responses, ensuring that generated content remains consistent with its source material. Assessments of bias, privacy violations, and inappropriate content are also required.

## Documentation

Unlike traditional models, where documentation focuses on training data and model parameters, RAG models require documentation on retrieval strategies, ranking algorithms, and external data dependencies. Version updates must include changes to retrieval mechanisms, knowledge source modifications, and updates to prompt structures.

## Ongoing Monitoring

While traditional backtesting is useful for point-in-time assessments, it offers limited value compared to traditional models, due to the dynamical nature of the models and their sensitivity to changes in context post-deployment. Effective model risk management for RAG models requires continuous monitoring to ensure reliability, accuracy, and compliance. Ongoing evaluation should identify performance degradation, emerging risks, and unintended biases by systematically tracking key indicators such as accuracy, relevance, groundedness, and potential concerns related to bias, privacy violations, and content toxicity. Additionally, because RAG models enable AI-driven search, model owners must track whether proprietary or confidential financial information is being indexed, retrieved, and exposed to end users.

Monitoring should integrate automated evaluation metrics with human oversight to detect discrepancies between expected and actual performance. Targeted analysis methods, including marginal and bivariate assessments, can help pinpoint specific weaknesses, enabling proactive mitigation strategies. The RAG model deployment should include methods for estimating response uncertainty so that these situations can receive additional review.

[11] 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

# Conclusion

Deployments of generative AI systems for consumer communications bring model risk management requirements to groups who may not have previously had exposure to these. This means that deploying GenAI systems is more than just an IT integration activity.

The principles of model risk management still apply, but their application to GenAI needs to be rethought. Real-time monitoring will be emphasized over pre-deployment validation. Human-in-the-loop oversight will require AI augmentation in order to be successful. Documentation and validation will have little access to general LLM model details but will need to provide detailed reviews of the fine-tuning data and process. LLM vendors will need to learn core principles of model risk management, such as a second line of defense performing independent review and monitoring. Some LLM vendors are already learning that proper model risk management can aid in building trust in their systems, thus widening their adoption.

# Bio

Dr. Breeden has been designing and deploying risk management systems for loan portfolios since 1996. He has been the Founder and CEO of Deep Future Analytics since 2011, which focuses on portfolio and loan-level forecasting solutions for pricing, account management, stress testing, CECL / IFRS9; and AI monitoring; serving banks, credit unions, and finance companies.

He is a board member of Upgrade, a San Francisco-based FinTech; President of the Model Risk Managers' International Association (mrmia.org); and an Associate Editor for the *Journal of Credit Risk*, the *Journal of Risk Model Validation*, the *Journal of Risk and Financial Management*, and the journal *AI and Ethics*. He is also the founder of auctionforecast.com, which predicts the values of fine wines using a proprietary database with over 4.5 million auction prices.

Dr. Breeden earned a Ph.D. in physics and has published over 90 academic articles, 9 patents, and 6 books, including *Redesigning Credit Risk Modeling to Achieve Profit and Volatility Targets* (2024), available on Amazon.com.

DEEP FUTURE

ANALYTICS