Google Cloud

air Alliance for Innovative Regulation

Generative AI Risk Management in Financial Institutions

Updated: October 2024

Google Cloud

air Alliance for Innovative Regulation

Table of contents

Google Cloud

air Alliance for Innovative Regulation

Executive summary

The introduction of Generative Artificial Intelligence ("generative AI") in the financial services sector promises to usher in an era of transformation for quality, accessibility, efficiency, and compliance in financial markets and services. Generative AI offers many benefits, including the potential to enhance individual productivity, strengthen security operations, and drive data-based decision-making and operational efficiencies. Ninety percent of senior leaders running generative AI in production report resulting revenue gains of 6% or more,[1] and aggregate efficiency, and productivity gains due to generative AI initiatives across the banking sector are forecast to be substantial.[2]

At the same time, generative AI introduces risks that must be managed and mitigated. Historically, regulators and the financial services industry have developed various model risk management frameworks to address the potential risks that arise from the use of models in decision-making. These principles-based frameworks typically provide for model validation (rigorous assessment of a model's accuracy, reliability, and limitations), governance and control (roles and responsibilities for model development, implementation, and monitoring), and risk mitigation (identifying and managing potential risks, such as model bias, data quality issues, and misuse). This paper advocates for the continued reliance on these well-established model risk management frameworks to address the emerging challenges posed by generative AI.

Developed jointly by the Alliance for Innovative Regulation and Google Cloud, this paper builds on our earlier joint publication, '*Applying Model Risk Management Guidance to Artificial Intelligence/Machine Learning Based Risk Models.*'[3] In this paper, we expand on that foundation by exploring how model risk management frameworks and established governance practices can be applied to manage risks in generative AI contexts.

Specifically, the paper proposes that regulators acknowledge best practices, provide enhanced regulatory clarity, and establish expectations in the following four areas: (A) model governance and controls; (B) model development, implementation and use; (C) model validation and oversight; and (D) shared responsibility in third-party risk management.

We highlight three key topics where additional regulatory clarity can benefit all stakeholders:

1. **Documentation Requirements** – We recommend updating and clarifying model risk management guidance to specify documentation expectations for generative AI models.
2. **Model Evaluation and Grounding** – We recommend that regulators take into account developers' use of practices such as grounding and outcome-based model evaluations, in addition to model explainability and transparency, in establishing the safety and soundness of generative AI-based models.
3. **Controls for Safe and Sound AI Implementation** – We recommend that regulators recognize a set of controls, including continuous monitoring, robust testing protocols, and human-in-the-loop oversight, that are appropriate for ensuring the responsible deployment of generative AI in financial services.

Responsible adoption of generative AI in financial services requires a collaborative approach among industry participants, regulatory bodies, and technology providers. With tailored application of model risk management frameworks and enhanced regulatory certainty, financial institutions can responsibly adopt technological advancements in a manner that aligns with regulatory expectations, while upholding high standards of compliance, accountability, ethics, and transparency.

[1] Google Cloud (2024) *The ROI of generative AI in financial services.* Available at: https://inthecloud.withgoogle.com/gen-ai-index-finance/dl-cd.html (Accessed: September 12, 2024).
[2] Chui, M. et al. (2023) *The economic potential of generative AI*, McKinsey & Company. Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier (Accessed: September 16, 2024).
[3] Google Cloud and Alliance for Innovative Regulation (Alliance for Innovative Regulation) (2023) *Applying model risk management guidance to artificial intelligence/machine learning based risk models.* Available at: https://services.google.com/fh/files/misc/wp_applying_existing_ai_ml_model_risk_management_guidance.pdf.

Introduction

Generative AI has the potential to contribute significantly to the economy, with estimates suggesting a potential addition of up to $340 billion annually to the banking sector alone.[4] Multinational corporations across many industries, including financial institutions, are exploring or already taking advantage of generative AI-based models.

Governments also recognize the strategic importance of AI, prioritizing partnerships with the private sector to foster innovation, and ensuring that proper regulatory frameworks are in place to manage risks. Importantly, many financial regulatory bodies and supervisory authorities around the world are actively exploring how to oversee financial institutions that are using advanced AI technologies.[5] For instance, the Financial Stability Board (FSB), the international body that monitors and assesses vulnerabilities affecting the global financial system, advises that many of the risks posed by AI are familiar to financial regulators. The FSB Standing Committee Chair recently highlighted that principles of model risk management, including managing data quality, design, and governance, provide a robust framework for assessing and mitigating risks.[6] Likewise, the Monetary Authority of Singapore has set specific guidelines to ensure the safe and responsible deployment of generative AI technologies within its financial sector.[7] Further, the European Banking Authority has issued a report on machine learning (ML) for internal ratings-based models, offering recommendations for regulatory compliance of ML techniques.[8]

These initiatives underscore a global commitment to ensuring that AI technologies are managed with appropriate safeguards, highlighting the need for the application of model risk management frameworks to adapt to rapid technological advancements.

In the United States, a broad set of financial laws and regulations govern activities that may involve AI technologies, including generative AI.[9] These comprehensive requirements help to safeguard consumers and market participants, and have proven durable in addressing evolving technologies over decades. In order to further mitigate the risk associated with such evolving technologies, regulators have promulgated financial regulatory guidance and standards, such as the 'Supervisory Guidance on Model Risk Management'[10] (SR 11-7). SR 11-7, issued by the U.S. Federal Banking Agencies, was updated by the Office of the Comptroller of the Currency in 2021 to expressly reference AI. These frameworks provide a structured approach to assessing and mitigating risks associated with advanced AI models.

In the first section of this paper, we provide a lay of the land for generative AI, delving into how generative AI works and differs from traditional AI, its key benefits and associated risks. This provides a foundation for Section II, which provides a discussion on applying model risk management frameworks to mitigate generative AI risks, focusing on effective governance; model development, implementation, and use; model validation; and third party risk management.

In applying model risk management frameworks to generative AI, the paper argues that generative AI models are not inherently high-risk just by virtue of the technology involved; instead, the risk they pose should be assessed based on the specific use-case or application at issue and the unique characteristic of such models that are implicated based on the use-case or application. Further, the paper argues that clear governance frameworks that define roles, responsibilities, and accountability will be essential for effective oversight of generative AI.

[4] Chui, M. et al., 2023.
[5] McCaul, E. (2024) *From data to decisions: AI and supervision*. Revue Banque, 26 February. Available at: https://www.bankingsupervision.europa.eu/press/interviews/date/2024/html/ssm.in240226~c6f7fc9251.en.html (Accessed: September 12, 2024).
[6] Liang, N. (2024) *Remarks by Nellie Liang, US Under Secretary for Domestic Finance, and Chair of the Financial Stability Board Standing Committee on Assessment of Vulnerabilities, at the OECD – FSB Roundtable onArtificial Intelligence in Finance, Paris, 22 May*. Available at: https://www.fsb.org/2024/06/remarks-by-nellie-liang-on-artificial-intelligence-in-finance/ (Accessed:September 12, 2024).
[7] Monetary Authority of Singapore (MAS) (2023) *Emerging Risks and Opportunities of Generative AI for Banks: A Singapore Perspective*. Available at: https://www.mas.gov.sg/-/media/mas/news/media-releases/2023/executive-summary---emerging-risks-and-opportunities-of-generative-ai-for-banks.pdf.
[8] European Banking Authority (EBA) (2023) *Follow-up Report from the Consultation on the Discussion Paper on Machine Learning for IRB Models*, EBA/REP/2023/28, August. Available at: https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf (Accessed: September 12, 2024).
[9] SIFMA & SIFMA Asset Management Group (2024) *Promoting Investor Success, Industry Innovation, and Efficiency with AI, SIFMA*. Available at: https://www.sifma.org/wp-content/uploads/2024/09/AI-Whitepaper-Promoting-Investor-Success-Industry-Innovation-and-Efficiency-with-AI.pdf (Accessed: October 1, 2024).
[10] Office of the Comptroller of the Currency (OCC) (2021) *Model risk management: New Comptroller's Handbook Booklet*, Office of the Comptroller of the Currency (OCC). Available at: https://occ.gov/news-issuances/bulletins/2021/bulletin-2021-39.html.

To this end, there is an opportunity for regulators to clarify documentation and ongoing testing and monitoring responsibilities, especially in the context of financial institutions partnering with third-party model developers. By applying and adapting model risk management frameworks to generative AI, all involved stakeholders can promote the responsible adoption of generative AI in financial services.

I. Understanding Generative AI

A. How Generative AI Works & Differences from "Traditional AI"

Generative AI is a subset of AI that uses pattern recognition and contextual and memory capabilities to generate new content. Generative AI models create new outputs from sample information in a dataset, or could determine the probability that new samples or inputs are from a given dataset.[11] Large language models (LLMs), a subset of foundation models that process and generate text, are trained on vast datasets to predict and generate language through probabilistic assessments of variables like the next word or phrase.[12]

Beyond LLMs, generative AI can also be multimodal, meaning it can interpret and combine different information formats including images, music, audio, and video, further enhancing its versatility and adaptability across various applications and use cases. These models excel at identifying complex patterns across different mediums, using pattern recognition to generate new outputs based on the inputs they receive, and simulating agent-like behavior. This shift to generative AI marks a notable evolution in AI technology, enabling new forms of human-technology interaction.

Unlike "traditional" AI, which primarily analyzes and interprets existing data to make decisions or predictions, generative AI ventures into the realm of creation. "Traditional" or "predictive" AI is typically deterministic, requiring detailed, topic or domain-specific data and significant customization. Generative AI on the other hand, leverages pre-trained large models which create outputs that are probabilistic, offering a range of possible outcomes based on the learned patterns and the input provided.[13]

Definitions Used in This Paper[14]

AI system: Is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

AI models: Are a set of instructions or rules that enable machines to learn, analyze data and make decisions based on that knowledge.

Foundation model: A machine learning model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of downstream distinctive tasks or applications, including simple task completion, natural language understanding, translation, or content generation.

Large Language Model (LLMs): As a subset of foundation models, large language models (LLMs), are models that can help computers analyze, understand and respond to human inputs using speech and written text.

B. Benefits of Generative AI

In the financial services sector, generative AI can transform operations by enhancing efficiency, supporting decision-making support,[15] augmenting intelligence, and improving compliance through applications in areas such as fraud detection, risk management, and the automation of manual processes.[16]

[11] Google for Developers (n.d.) *Machine learning glossary*. Updated 2024. Available at: https://developers.google.com/machine-learning/glossary (Accessed: September 17, 2024).
[12] Zewe, A. (2023) *Explained: Generative AI*, MIT News | Massachusetts Institute of Technology. Available at: https://news.mit.edu/2023/explained-generative-ai-1109 (Accessed: May 30, 2024).
[13] Vassilev, A. et al. (2024) *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations*. Gaithersburg, MD: National Institute of Standards and Technology (NIST). doi: 10.6028/nist.ai.100-2e2023.
[14] See Asia Cloud Computing Association (2024) *AI Handbook, Asia Cloud Computing Association*. Available at: https://asiacloudcomputing.org/wp-content/uploads/2024/09/2024-09-13-ACCA-AI-Handbook-Final.pdf (Accessed: October 7, 2024).
[15] Sekhar, S. G. A. (2023) *How can financial services firms derive value from Generative AI?, EY*. MIT OpenCourseWare. Available at: https://www.ey.com/en_us/financial-services/is-financial-services-ready-for-generative-ai (Accessed: May 30, 2024).
[16] Hall, B. (2024) *101 real-world generative AI use cases from the world's leading organizations, Google Cloud Blog*. Google Cloud, 12 April. Available at: https://cloud.google.com/transform/101-real-world-generative-ai-use-cases-from-industry-leaders.

- **Enhanced Operational Efficiency and Automation:** Generative AI can automate repetitive tasks such as responding to requests for proposals, localizing multilingual content, and conducting compliance checks. This automation can reduce manual errors and frees up resources for higher-value initiatives, such as strategic planning and innovation. In some instances, AI systems are used to streamline the development of security and compliance features, improving both speed and accuracy in product development and operations.
- **Advanced Data Management and Insights:** Generative AI can help handle complex, unstructured data, enabling firms to extract valuable insights through advanced conversational interfaces and data summarization techniques. This capability supports informed decision-making and strategic planning. This helps financial institutions analyze large datasets quickly, uncover hidden patterns, and make informed decisions in real time. These capabilities are particularly valuable in risk management and strategic forecasting, where timely and accurate data analysis can provide a competitive edge.
- **Enhanced Customer Engagement:** Generative AI can transform customer interactions through advanced chatbots and enhanced search functionalities, providing personalized, efficient customer interactions, and setting new benchmarks for customer satisfaction.
- **Improved Security Operations:** Generative AI can enhance security operations by contextualizing the latest threats and support staff in detecting, investigating, and responding to cyber threats to ensure constant threat oversight. It prioritizes critical threats, automates routine tasks, and improves threat detection and vulnerability management. This enables analysts to focus on significant issues, reduces burnout, and shifts security operations from reactive to proactive, ultimately improving overall security efficiency.
- **Increased Individual Productivity:** Generative AI can improve employee productivity across a number of functions by: amplifying creativity for marketers, speeding up information finding for customer service agents, accelerating coding for developers, providing helpful summaries from internal knowledge repositories such as HR, and providing assistance in email and calendar applications as well as spreadsheets, presentations, documents, and others. These tools enable employees to focus on higher-value activities while reducing time spent on manual processes, ultimately increasing efficiency and driving innovation.

C. Risks Associated with Generative AI

While generative AI applications offer significant potential benefits, the technology also has unique characteristics and risks that should be assessed and mitigated. The risks associated with generative AI fall into two primary areas: inherent risks and external risks. Inherent risks stem directly from the technology itself, such as unintended bias that can skew model decision-making outcomes and the overall complexity of generative AI systems that may affect model transparency. External risks arise from the broader application of generative AI within the financial ecosystem, encompassing data security challenges, regulatory compliance issues, and the complexities involved in integrating new technologies into existing infrastructures.

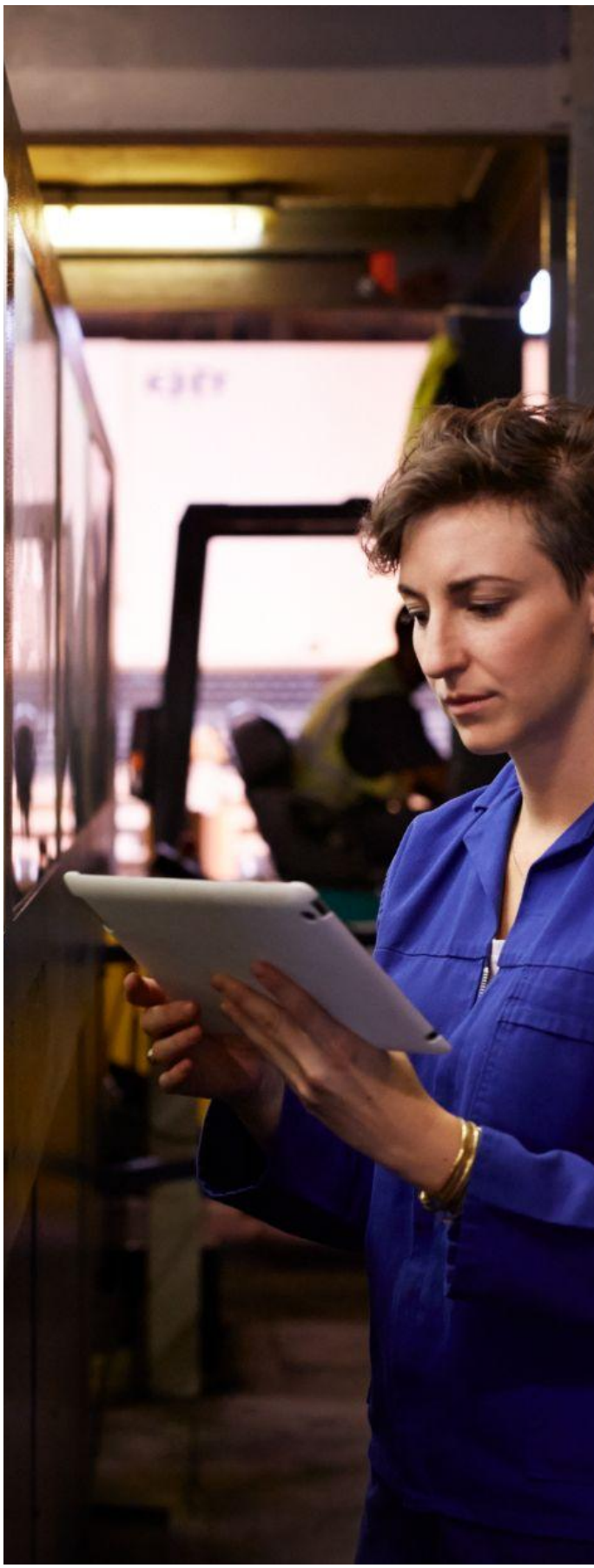
In terms of intrinsic risks, the following are particularly relevant:

- **Hallucinations:** Hallucinations in generative AI models occur when models generate outputs that appear coherent but are factually incorrect or nonsensical. These can be caused by data quality issues, or the complex learning architecture of these models. Unlike bias, which reflects skewed or unfair outcomes, hallucinations involve the model producing incorrect or misleading information, potentially damaging decision-making processes within financial institutions.

- **Amplification of Bias:** Generative AI systems can unintentionally amplify harmful biases in the data they are trained on, leading to skewed outcomes which can be unfair in certain settings. These biases can disproportionately affect certain demographic groups or customer segments, potentially resulting in unfair decisions or outcomes.
- **Explainability:** Generative AI systems are inherently complex and often lack transparency in how they arrive at decisions or outputs. This aspect may make it difficult for stakeholders—both technical and non-technical—to fully understand the rationale behind the AI's behavior, which can challenge efforts to ensure regulatory compliance, maintain public trust, and establish accountability within financial institutions.

In terms of external risks, some of the key concerns raised by financial institutions with respect to Generative AI pertain to data security and global regulatory uncertainty.

- **Data Security:** Typically these concerns are focused on ensuring that confidential or proprietary data is safeguarded from unwarranted access or potential exposure – for example, to avoid training data being provided verbatim in response to users' prompts, or queries, or to avoid manipulation of prompts to trick the generative AI model into revealing sensitive data. Yet other risks exist, such as model poisoning, where the training data set is corrupted with mislabeled or misleading information that can lead the large language model to learn incorrect patterns and produce inaccurate results. Robust data security and integrity measures are needed to safeguard against these kinds of risks.
- **Regulatory Uncertainty:** Financial institutions face a rapidly evolving regulatory landscape when integrating generative AI. The complexity of analyzing, reconciling, and interpreting regulatory attitudes and requirements can strain resources, impacting the ability of institutions to adapt swiftly and focus resources on model risk management. These risks underline a deep need in the industry for regulatory clarity.



II. Adapting Model Risk Management Frameworks to Generative AI

Established model risk management frameworks, such as those outlined by the Supervisory Guidance on Model Risk Management (SR 11-7) and the NIST AI Risk Management Framework, provide a structured approach to assessing and mitigating the risks associated with financial models. These frameworks emphasize robust validation, comprehensive governance, and ongoing monitoring to ensure the reliability and transparency of financial models.[17] Specifically, the NIST AI Risk Management Framework offers guidance on managing AI risks, including integrating AI governance into existing governance structures and processes such as privacy and security. The recent draft AI Risk Management Framework Generative AI Profile addresses unique risks posed by generative AI but is notably industry-agnostic and therefore not specifically targeted at financial institutions; nevertheless, it can serve as a valuable resource.

The U.S. framework has similar counterparts around the world. For example, the principles set forth in the UK's SS1/23 (Supervisory Statement 1/23) on model risk management mirror those established by the U.S. regulatory bodies. SS1/23 mandates similar protocols for model identification, classification, and governance, ensuring that models are clearly defined, inventoried, and assessed based on risk. Additionally, SS1/23 emphasizes the importance of model testing, ensuring data quality, and verifying the appropriateness of models for their intended uses—foundational principles for deploying generative AI technologies effectively and safely.

Historically, these kinds of model risk management frameworks have been essential in managing financial model risk by ensuring models are appropriately developed, implemented, and maintained. A recent U.S. Treasury AI Cyber Risk Report highlights that "most institutions identified utilizing the Office of the Comptroller of the Currency's Model Risk Management guidance to drive their underlying controls related to model risk... their existing practices may already align with many aspects of these frameworks, though implemented in different guises through existing risk management policies and frameworks."[18] The frameworks establish protocols for identifying, assessing, and mitigating risks, thus maintaining operational integrity and regulatory compliance in the financial sector.

The principles-based nature of model risk management frameworks allows firms to adapt them to new technologies such as generative AI, and avoids the need for entirely new regulatory frameworks. However, the unique complexities and potential impacts of generative AI may require regulatory clarifications to help firms effectively manage the new types of risks introduced by these advanced AI systems.

A. Robust Governance and Program Oversight

Effective program governance, management, personnel, and oversight policies and procedures are core to model risk management. They play a crucial role in shaping and controlling the model development process. Regulators should consider and amplify the importance of the following considerations and strategies when firms are establishing these foundational processes:

- **Socio-technical Considerations:** Managing AI risks requires a comprehensive approach that integrates both technical and social factors. NIST has identified a series of "trustworthiness characteristics" — including (1) valid and reliable, (2) safe, secure and resilient, (3) accountable and transparent, (4) explainable and interpretable, (5) privacy-enhanced, and (6) fair with harmful bias managed. According to NIST, "[c]reating trustworthy AI requires balancing each of these characteristics based on the AI system's context of use. While all characteristics are socio-technical system attributes, accountability and transparency also relate to the processes and activities internal to an AI system and its external setting. Neglecting these characteristics can increase the probability and magnitude of negative consequences."[19]

[17] *See* Google Cloud and Alliance for Innovative Regulation (Alliance for Innovative Regulation) 2023, p. 7-8. for a detailed discussion on SR 11-7 model risk management guidance requirements.
[18] U.S. Department of the Treasury (2024) *Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector*, p. 45. Available at: https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf (Accessed: September 12, 2024).
[19] *See* National Institute of Standards and Technology (NIST) (2023) *Trustworthy & Responsible AI Resource Center: AI Risks and Trustworthiness, National Institute of Standards and Technology (NIST)*. Available at: https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Foundational_Information/3-sec-characteristics (Accessed: October 9, 2024).

Human judgment is needed to determine the specific metrics related to AI trustworthiness and set appropriate thresholds. In addition, tradeoffs often arise between characteristics, and rarely do all apply equally in every setting. For example, optimizing privacy may reduce predictive accuracy, requiring model developers and deployers to carefully balance tradeoffs between these metrics and consider the appropriate requirements based on the model's use case or application. Having the right governance teams and personnel involved at the outset—and throughout the model development and implementation life cycle—continues to be a key component of demonstrating adherence to risk management best practices.

- **Data Governance:** Generative AI models are limited by the recency, source, and comprehensiveness of their training data. Enhancing the overall quality and quantity of input data for fine-tuning models is critical for ensuring the robustness and reliability of generative AI models. This process includes how an enterprise conducts data selection and cleansing, enriching datasets to cover a broader spectrum of scenarios, and augmenting data with synthetic examples.

  To maximize the accuracy and relevance of AI outputs, it's important to utilize high-quality, current datasets. Accurate and representative data reduces the likelihood of hallucinations. Problems with data quality can be exacerbated or become more apparent in generative AI due to the nature of the outputs they produce. Implementing comprehensive data management practices—such as data cleansing and validation—can help promote the accuracy and reliability of input data, which in turn helps mitigate the risks associated with data bias, inaccuracies, and inconsistencies.

- **Model Documentation:** With generative AI, it is even more important that financial institutions and model developers and deployers understand regulators' documentation expectations for all categories of AI risk management. We recommend that existing model risk management guidance be enhanced to include discussion of documentation requirements in the context of advanced AI models, including generative AI. To this end, regulators might consider creating financial sector use-case-based guidance on proper documentation, including for the context of third-party developed models.

  One approach that has gained traction since it was introduced in a 2018 Google research paper[20] is the use of model cards. Model cards provide a structured way to present the essential facts of machine learning models and have been utilized within the industry. Model cards can take various forms depending on the use case. Given the rapid evolution of AI technologies, it is crucial that model card structures and content remain flexible to accommodate emerging industry benchmarks for performance and safety. Model documentation should be risk-based, protect intellectual property and security, and should not follow a one-size-fits all approach.[21]

  Sufficiency of documentation should be determined by what financial institutions need to use, validate, and understand the model's design, theory, and logic. Disclosure of proprietary details, such as model code, is unnecessary and unhelpful in verifying the sufficiency of a model and would have the unintended consequence of deterring model developers and deployers from sharing best-in-class technology with financial institutions.

[20] Mitchell, M. et al. (2018) *Model cards for model reporting, arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1810.03993 (Accessed: September 12, 2024).
[21] Google (no date a) Model cards: The value of a shared understanding of AI models, Google. Available at: https://modelcards.withgoogle.com/about (Accessed: September 12, 2024).

Although the existing regulatory model risk management guidance[22] requiring developing and maintaining "adequate documentation" that "explains in detail" the design, theory, and logic of the model is understandable and expected, few concrete parameters are provided regarding what would be deemed appropriate by examiners. Clear regulatory expectations relating to documentation sufficiency – including recognition of industry solutions, such as model cards – would be helpful to governance teams, developers, and deployers of AI models.

- **Explainability:** Explainability is useful for the purposes of understanding specific outcomes of AI/ML models. However, there are two points that regulators should bear in mind. First, the level of required explainability varies significantly across different applications of AI, based on the specific operational context in which it is used. A risk-based approach should be used to assess the level of explainability needed for a particular application. Second, while explainability is a helpful concept in the context of assessing the suitability of particular outputs, it may be insufficient or ineffective to establish whether the model as a whole is sound and fit for purpose. For those purposes, it will be important to look to other factors – such as ongoing model testing and monitoring programs (see below). These factors should be prioritized, relative to explainability, in the MRM assessment process.

B. Reliable Model Development, Implementation and Use

Model risk management frameworks underscore the importance of robust processes surrounding the development, deployment, and use of generative AI models. This includes implementing techniques to support the reliability and accuracy of AI outputs. We propose that regulators recognize techniques such as grounding and model evaluation based on outcomes, as being appropriate methods to enhance explainability, accuracy, and transparency of generative AI models.

- **Grounding:** Grounding[23] is a technique that anchors model responses to verified data sources, reducing the likelihood of models generating content that isn't factual ("hallucinating") and enhancing the trustworthiness and applicability of the generated content. In simple terms, grounding gives a model a connection to a trusted data source, allowing it to access and process information beyond its initial training data. This makes the model more reliable, informative, and capable of handling real-world tasks. We propose that regulators recognize grounding techniques, alongside model evaluation based on outcomes (see below), as appropriate methods to anchoring on verified information.
- **Model Evaluation:** Model evaluation is a critical development element that helps organizations have a clear understanding of how well generative AI models and applications align to their intended task. Model evaluation often incorporates both quantitative and qualitative assessments. From a quantitative standpoint, model evaluations can leverage metrics such as Bilingual Evaluation Understudy (BLEU) to measure the similarities between machine-generated translation and one or more human-written reference translations or Inception Score (IS) to evaluate the quality and diversity of generated images. From a qualitative standpoint, human evaluation is key, often requiring the development of human evaluation protocols and the incorporation of domain expertise. In addition, model evaluation suites are increasingly available from technology providers allowing financial institutions to compare evaluation metrics across multiple models to help inform which model should be deployed using visualization tools and/or evaluation metrics and, after deployment, to help maintain model performance.[24]

[22] Office of the Comptroller of the Currency (OCC), 2021.
[23] Google Cloud (updated 2024) *Generative AI on Vertex AI Documentation Guide on Grounding. Google Cloud.* Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/grounding/overview (Accessed: May 30, 2024).
[24] Google Cloud (updated 2024) *View and interpret evaluation results. Google Cloud.* Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/view-evaluation (Accessed: October 1, 2024).

C. Model Validation and Ongoing Monitoring and Testing

Existing MRM frameworks underscore the importance of model validation: ensuring the model is conceptually sound and operates as expected, including through ongoing monitoring and testing during model use. These concepts remain relevant for generative AI models. We urge regulators to recognize the following strategies as being particularly important to mitigating risk in the context of generative AI:

- **Human-in-the-Loop Oversight:** To support the safe and sound deployment of generative AI in financial services, we recommend that regulators evaluate and confirm the adequacy of controls such as continuous monitoring, robust testing protocols, and human-in-the-loop oversight. This includes protocols for manual reviews in high-impact scenarios.[25] Processes to support effective human intervention can include:[26]

  - **Risk Ranking.** Financial institutions can rank the risks of AI use cases based on agreed-upon criteria such as internal versus external use, involvement of sensitive data, impact on individuals, whether the application is mission-critical, and the level of uncertainty in AI outputs.

    Importantly, the mere use of generative AI should not automatically render a model high-risk—rather, that risk should be assessed based on the particular application to which generative AI is applied.

    Additionally, how the particular model is developed—whether in-house or through a third-party—impacts the risk profile. Depending on whether an institution builds its own generative AI applications, customizes existing models, tailors and integrates third-party-made models, or uses third-party out-of-the-box solutions, the nature and scope of risks will vary.

    The overall risk ranking, taking into account these various factors, should drive considerations of how and to what extent a human-in-the-loop is required to ensure proper model oversight.

  - **Implementing Triggers.** Once risks are identified and ranked, technical or operational thresholds can be implemented that necessitate human review, approval, or rejection of AI-generated decisions and actions. These controls often include manual review processes, confirmation prompts, or the ability to override AI decisions.

- **Advanced Fairness Metrics:** Ensuring that decisions made by generative AI are transparent, equitable, and accountable is crucial. This involves understanding fairness metrics specifically designed to identify and mitigate unintended or unwanted biases. Continuous monitoring for disparities in model outputs across different groups helps maintain fairness and prevent biases that could undermine trust. Many financial products are intentionally designed to target specific populations and may involve a level of statistical bias to provide tailored recommendations. This bias must be well-understood and managed, however, to monitor that it is not contributing to unfair or discriminatory outcomes. Data and model bias metrics help with efforts to evaluate and manage biases, ensuring models perform equitably across different demographics.[27]

- **Safety Filtering:** Utilizing safety filtering mechanisms helps prevent the generation of inappropriate or harmful content. Implementing built-in content filters and safety attribute scoring can help financial institutions detect and mitigate potential risks associated with generative AI outputs. For example, safety filters can screen for offensive or insensitive language, while safety attribute scoring can define confidence thresholds for acceptable outputs based on specific use cases and business contexts.

[25] The high-impact scenario approach is in line with the EU AI Act, which emphasizes the importance of human oversight in critical decision-making processes.
[26] Chuvakin, A.. and Kaganovich, M. (2024) *Generative AI governance: 10 tips to level up your AI program, Google Cloud Blog*, 1 February. Available at: https://cloud.google.com/transform/gen-ai-governance-10-tips-to-level-up-your-ai-program.
[27] Google Cloud (updated 2024) *Introduction to model evaluation for fairness*. Available at: https://cloud.google.com/vertex-ai/docs/evaluation/intro-evaluation-fairness (Accessed: May 30, 2024).

- **Robust Testing & Monitoring:** Implementing continuous testing, validation, and monitoring of generative AI outputs against expected results and known data points is crucial to mitigating risks. For instance, overfitting – a subset of model quality issues – occurs when a generative AI model over-indexes on details and noise in the training data to the detriment of its performance on new data. This could be particularly problematic in financial modeling, where models must perform well across diverse and unforeseen market conditions. Beyond these risks, in the context of generative AI models, robust testing and monitoring procedures are necessary to detect and mitigate critical issues, such as data leakage and adversarial attacks. The following are key risk mitigation approaches:

  - **Penetration testing and red teaming.** Incorporating penetration testing and red teaming to identify vulnerabilities and test the organization's incident response can help enhance security measures. Stress tests and scenario analyses evaluate model performance under various conditions. These tests help identify potential weaknesses and make necessary adjustments to enhance the model's accuracy, reliability and security. Implementing systems to continuously track model performance and detect deviations from benchmarks is essential. For example, anomaly detection systems can help spot unusual predictions that might suggest the model is drawing conclusions based on spurious relationships rather than valid insights.[28]

  - **Automated systems.** Automated systems can adjust models promptly when performance deviates from expected or acceptable outcomes. This includes adapting and retraining models as new data becomes available and market conditions evolve. Automated monitoring systems can provide real-time alerts when model performance deviates from benchmarks, prompting immediate reviews and adjustments.

D. Shared Responsibility in Third-Party Risk Management

When a financial institution works with a third-party model developer, there are unique issues that can arise in applying model risk management frameworks and techniques. In these circumstances, it becomes even more important for all stakeholders to understand their respective roles and expectations. A collaborative approach between financial institutions, AI developers and deployers, and regulators is essential for navigating the complexities of generative AI implementation. Shared responsibility ensures comprehensive risk management, from model development to integration into an application, aligning technological advancements with operational and regulatory needs along the AI value chain.

In addition, when working with a third-party, the financial institution should consider data portability and interoperability requirements or ensure vendor adherence to international standards such as ISO 19941[29] to avoid vendor lock-in—with respect to both the model developer and/or any underlying model infrastructure required to operate the model. As more data is stored in the cloud and enables the use of generative AI, interoperability is a critical factor in ensuring financial institutions can facilitate data porting and working across a multi-cloud ecosystem. Data portability and interoperability are central to innovation and help to boost competition; without it customers may face higher prices, less security, and less innovation in their cloud and AI services due to limited cloud choices.

[28] *Evaluate model and system for safety* (updated 2024) *Google AI for Developers*. Available at: https://ai.google.dev/responsible/docs/evaluation (Accessed: October 1, 2024).
[29] ISO (2017) *ISO/IEC 19941:2017*. Available at: https://www.iso.org/standard/66639.html (Accessed: October 1, 2024).

III. Conclusion

The integration of generative AI into the financial services sector presents a transformative opportunity to enhance operational efficiencies, customer engagement, and overall decision-making capabilities. As with any technology, this is not without its risks and challenges, necessitating a robust and dynamic approach to model risk management. Existing frameworks are robust yet sufficiently flexible to meet the innovative demands of generative AI, and can be fine-tuned to effectively manage specific risks within financial institutions.

The adaptability of existing model risk management frameworks to these new technologies mitigates the need to create entirely new regulatory structures. Accordingly, the focus should be on optimizing, clarifying and adapting these established frameworks to ensure comprehensive risk evaluations. This includes incorporating thorough assessments to maintain operations within accepted boundaries and standards.

These standards and guidelines will only be achieved through collaboration between industry participants, regulators and governmental bodies. While the path forward involves navigating complex regulatory and ethical landscapes, the collective commitment to responsible innovation and adherence to robust model risk management practices will be pivotal in realizing the full potential of generative AI in financial services and beyond.

Looking forward, the discourse on generative AI in financial services will continue to evolve. We will continue to examine external risks associated with generative AI, including systemic and operational challenges, and the integration of these technologies into global financial systems.