Effective Generative AI Model Risk Management

Abstract

This paper provides a 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.

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.

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. Monitoring and fallback solutions must be a standard part of LLM deployment.

Monitoring will take priority over validation, but LLM monitoring is an emerging field. 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 monitoring, leveraging AI-assisted compliance mechanisms, and ensuring robust fallback planning, financial institutions can integrate GenAI into critical operations while mitigating its unique risks.

Authors & Affiliations

Joseph Breeden[1]
[1]Deep Future Analytics LLC, Santa Fe, USA