Managing the Model Risks of Generative AI Productivity Tools in Banking

Abstract

The recent advancements in Generative AI (Gen AI), and Large Language Models (LLMs) in particular, have resulted in the proliferation of multipurpose Gen AI-powered tools that promise to enhance the productivity of professionals in the financial sector. Though the outputs of these tools might not directly impact financial statements, they are often employed for information processing and may influence decisions that impact business operations. For this reason, and to align with regulatory obligations, such as those outlined in SS1/23 for UK-based financial institutions, the risks of these tools need to be appropriately managed.

In this paper, we introduce a novel validation approach specifically designed for GenAI-powered productivity tools. These tools assist people by drafting content, summarising documents and email threads, and answering questions on documents. Our methodology is tailored to comply with SS1/23 principles, ensuring regulatory adherence. It involves bespoke, use-case-specific testing that leverages domain-specific data to uncover subtle errors and limitations that might otherwise lead to misinformed decision-making.

We describe a structured testing framework which combines both quantitative metrics and qualitative assessments, utilising other LLMs to both scale up the generating of synthetic testing data and to automate the evaluation process. This large-scale evaluation, covering over a thousand individual outputs, has enabled us to assess performance in real-world scenarios across critical dimensions, such as accuracy, coverage, bias, toxicity, privacy, relevance, redundancy, robustness, and style and tone. The integration of automated analysis with subject matter expert reviews further enhances the reliability of our findings.

Our work highlights the importance of combining automated, high-volume testing with qualitative insights in the validation of GenAI-powered tools. This dual approach not only provides a robust framework for Model Risk Management but also ensures that these productivity tools continue to enhance operational efficiency while mitigating potential risks in complex, high-stakes environments.

The validation exercise underscores the challenges of assessing “black-box” AI models where independent verification for specific banking business contexts is limited. Through a combination of automated testing and subject matter expert review, our analysis provides insights into both the strengths and inherent risks of deploying such tools in high-stakes environments. Our findings advocate for continuous, scenario-specific validation to support robust Model Risk Management practices, ultimately ensuring that GenAI-powered productivity tools enhance operational efficiency while mitigating potential risks.

Authors & Affiliations

Dr Eleimon Gonis[1], Dr Maria Kalantzaki[1], Miss Olivia Nowicka[1], Mr Xiaodong Yi[1], Mr Chris Heys[2], Mr Tim Chapman[2], Mr Kyriakos Nikiforou[2], Mr Simone Pedemonte[2], Miss Fran Wilkinson[2]
[1]Virgin Money, London, United Kingdom. [2]PwC, London, United Kingdom