Building a better working world

Leveraging generative AI for model documentation automation

Table of Contents

Leveraging generative AI for model documentation automation

Abstract

The rapid advancements in artificial intelligence (AI) and machine learning have spurred interest in automating various aspects of the model lifecycle, including model documentation. This study explores the potential of generative AI, particularly large language models (LLMs), in accelerating the model documentation process. The current documentation process is predominantly manual, time-consuming, and labor-intensive, often involving multiple iterations and revisions. Traditional attempts at automation, such as rule-based systems and template-driven approaches, have limitations in adaptability, scalability, and output quality.

This paper identifies and tackles key technical challenges in model documentation, including deriving insights from tables, generating coherent narratives from diverse information sources, and acquiring domain knowledge and nuanced insights. By integrating LLMs into the model documentation process, organizations can alleviate the burden on SMEs, accelerate production timelines, and enhance overall document quality. The proposed methodology focuses on 1) gathering reference information and user input and extracting relevant information, 2) analyzing data through table interpretation and prompting techniques, and 3) synthesizing a narrative followed by expert review for enhancements. This approach ensures the creation of high-quality and contextually relevant documents similar to those authored by expert human writers.

Intro

Model documentation is a Business as Usual (BAU) process for all financial institutions. With the number of models ranging over several thousands a predetermined cadence of documentation review and refresh is contingent upon factors such as risk, complexity and business impact of the model. The developed documentation is required to meet various regulatory criteria and organizational standards. Currently, most financial institutions predominantly rely on a manual process to develop the documentation, resulting in key challenges such as an extended model lifecycle, and a lack of consistency across models.

Generative AI, particularly large language models (LLMs), offers a promising solution for automating model documentation. This is primarily attributed to LLMs' ability to retrieve and process vast quantities of information and data and generate narratives aligned with instructions. The primary focus of this study is to explore the potential of large language models, such as Azure OpenAI's GPT-4, in automating the model documentation process and examining the key value additions, such as accelerated model lifecycle, optimization of resources and structured documentation. This research also aims to provide a comprehensive understanding of the challenges and applicability of generative AI across broader domains.

1. Model documentation current process

In the current process, model documentation relies heavily on manual efforts, wherein subject matter experts (SMEs) are responsible for gathering inputs and providing detailed narrative of the approach, derivations, tests and other analyses that support using a risk model for a given purpose. The inputs typically consist of various sources, including historical model documentations, code outputs and SME's insights. The manual documentation process encompasses a series of steps, such as creating an outline, drafting, editing, and revising the content to ensure accuracy and coherence. This approach often entails multiple iterations and revisions, which can be time-consuming and labor-intensive, particularly when addressing complex topics or adhering to stringent guidelines.

Historically, attempts to automate the documentation process have primarily focused on leveraging rule-based systems and template-driven approaches. While these methods offer some degree of efficiency, they often need more adaptability and scalability, especially in the face of evolving requirements and diverse subject matters. Moreover, such systems tend to generate rigid and formulaic outputs, which may not adequately capture the nuances and intricacies inherent in expertly crafted documents.

2. Key technical challenges within model documentation

Deriving insights from tabular data is a significant challenge within the model documentation process, as it necessitates the meticulous examination of numerous metrics by subject matter experts (SMEs). This task can be laborious and time-consuming, as each document may contain multiple tables, each encompassing hundreds of data points. Furthermore, collecting and analyzing large quantities of data can lead to human error.

The model documentation process is further complicated by the need for domain knowledge and nuanced insights, especially when dealing with specialized topics. When combined with the task of crafting narratives using model guidelines, prior samples, and tabular data, this process turns into a lengthy manual endeavor.

3. Model documentation using generative AI

By harnessing the power of advanced artificial intelligence (AI) techniques, LLMs can comprehend intricate information, discern context and generate human-like responses. In this paradigm, the role of SMEs shifts from primary document creators to reviewers and editors, who can focus their expertise on refining and vetting the
AI-generated content.

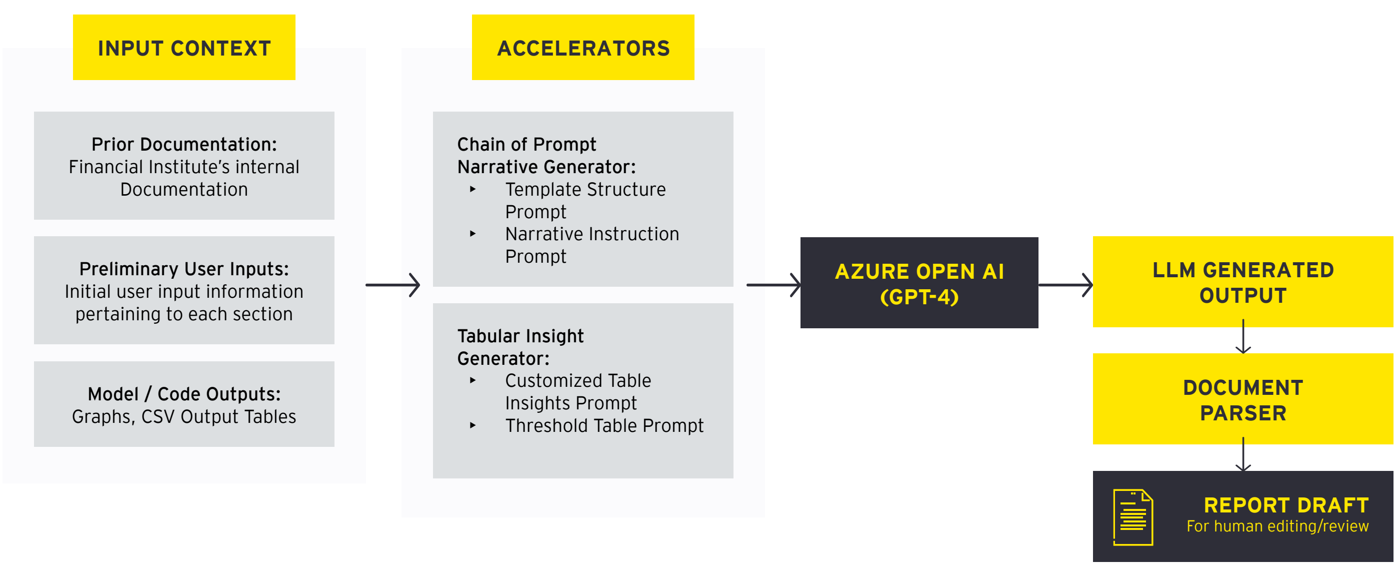
The approach outlined below aims to streamline content generation by leveraging AI to extract information, analyze data, and create coherent narratives. Its effectiveness lies in the collaboration between AI and subject matter experts, ensuring accurate and relevant output tailored to specific requirements.

- **Extract information from relevant sources taking additional user input as needed:** This phase entails the extraction of relevant data from various sources, including model guidelines, model code outputs, and other relevant repositories. Any missing information is provided to the LLM by the SME.
- **Reference historical samples:** Ingest examples of similar documents to acquire insights into best

practices, potential challenges, and established prototypes.

- **Understand document tone and structure:** Using historical samples and document guidelines, the LLM generates customizable templates and instructions to recreate similar documents.
- **Generate content using prompting techniques:** The implementation of advanced prompting strategies, such as few-shot learning and structured prompt templates, enables the generation of accurate and contextually appropriate content by the AI model.
- **Extract and analyze tabular data (reusable accelerator):** This stage necessitates the extraction of relevant information from tabular data sources and the subsequent analysis and interpretation of this data to support the development of coherent and insightful narratives.
- **Synthesize and generate document narrative (reusable accelerator):** Drawing upon the extracted information and table analysis, the AI model generates a cohesive and well-structured narrative that conforms to the established document structure.
- **Export to preferred document type:** Upon completion of the content generation process, the final output is exported to the desired document format (e.g., Word, PDF), ensuring compatibility with existing documentation workflows and facilitating seamless integration into the review and editing process.
- **SME review and enhancement:** The generated document is then submitted to subject matter experts (SMEs) for review and enhancement. SMEs provide their expertise to refine, validate, and ensure the accuracy and relevance of the AI-generated content, thereby elevating the overall quality of the document.

The human-in-the-loop layer serves as the ultimate line of defense, particularly as automation increases. This approach centers human reviewers at its core, ensuring the maintenance of high-quality documentation, content accuracy, and risk mitigation within an extensively automated framework (*Figure 1: Model documentation approach using generative AI*).



4. Prompt engineering approach

At the heart of this research's success is the concept of prompt engineering, as the effectiveness of the language model's responses largely relies on the quality and organization of the provided prompts. The process of documentation automation consists of several crucial elements, including 1) prompt engineering, 2) role definition, and 3) human involvement. These components are intricately intertwined within an accelerator-based approach and serve as the fundamental basis for this piece of work.

- **Prompt definition:** Central to the success of the language model's responses is the crafting of effective prompts. Utilizing OpenAI and Langchain's prompt templates, a strategic approach is devised to structuring prompts that enables clear instructions, well-defined output configurations, comprehensive analysis, and accurate segmentation and recognition of the various components provided to the model.
- **Role definition:** By assigning specific roles to the language model (e.g., teacher, philosopher, or document automator), a clear perspective is defined, thereby influencing its tone, assumptions, and language to align with the intended context.
- **Prompt iteration:** The iterative process of refining the model involves human feedback and experimentation on the efficacy of various prompts. This collaboration results in the development of an extensive prompt library, featuring high-quality prompts as exemplars and less optimal prompts as counterexamples. This dynamic resource serves to continually enhance the document automation process.

5. Solution approach

The documentation automation process is primarily executed through two principal accelerators: the Tabular Insight Generator and the Chain of Prompt Narrative Generator. The Tabular Insight Generator examines various numerical datasets, tabular data, and graphical representations, generating document-specific, tailored analyses. Subsequently, the narrative generator efficiently extracts and synthesizes pertinent inputs, including any analyses produced by the Tabular Insight Generator, creating of a comprehensive draft of the document.

The next section details the process through which these accelerators are employed, alongside their evolution
and maturation over time. The empirical outcomes from this investigation will be clearly delineated.

The results are specific to the selected model (Azure Open AI's GPT-4) and may vary if alternative models are employed. The selection of Azure Open AI's GPT-4 was based on its security features and content generation capabilities.

Before exploring the two accelerators, it is important to briefly discuss the concept of prompts. Prompts serve as the primary means for humans to interact with large language models and significantly influence the responses generated. Through extensive experimentation, we have devised an optimal method for crafting prompts, which is divided into several components:

6. Prompt Components

- **Instruction:** Clearly state the task or objective that needs to be accomplished in a concise manner.

This should be a brief directive on what is expected from the analysis or summary. The model responds well to step-by-step instruction.

- **Context:** Provide background information or set the scene for the task. This should give an overview of the situation, or the data being analyzed, which helps in understanding the purpose of the analysis.
- **Inputs:** Describe the data or information that will be used in the analysis. This can include data sources, types of data, or specific details about the data that are relevant to the task.
- **Output indication:** Explain the expected outcome or result of the analysis. This can be a summary, a narrative, a list of findings, or any other format that clearly communicates the results of the analysis.

Progression from basic prompting to a more comprehensive prompting methodology was instrumental in the build out of the following accelerators.

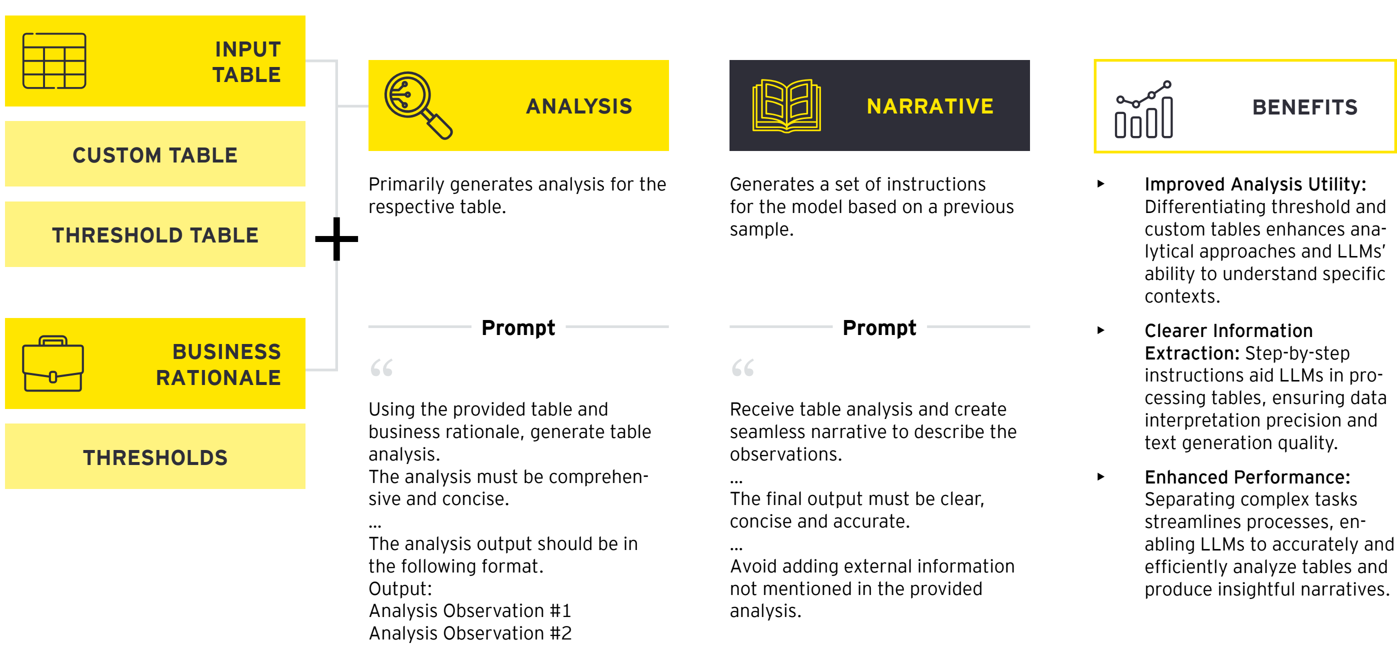
7. Tabular Insight Generator

High level strategy for table insight generation: An end-to-end strategy was developed to address the unique challenges associated with analyzing tabular data. This involved identifying table types, providing step-by-step instructions, and clearly specifying objectives.

Initial prompt engineering: Initially, a relatively simple and unambiguous prompt was attempted: Given a tabular representation of an X-type document, analyze the data and present the findings.

Key challenges: Firstly, while most of the information provided was accurate, much of it appeared to be of limited utility for the analysis. Secondly, it was evident that LLMs can effectively interpret numerical data. However, as the volume of data within the tables increased, the precision of the LLMs diminished. This led to an exploration of methods for converting tables into more comprehensible formats, such as transforming CSV files into Python data frames. Thirdly, it became apparent that LLMs exhibit a stronger affinity for textual data than numerical data, due to their optimization for natural language processing

Refined prompt engineering: *Figure 2: Tabular Insight Generator refined prompt flow and benefits*



Open challenges: While significant progress has been made in addressing the challenges associated with analyzing tabular data, several open challenges remain. These include exploring custom tables for increased scalability, further optimizing instruction-based prompting, and exploring Langchain's CSV agents to handle larger datasets. Research should continue to investigate these areas to maximize the effectiveness of LLMs in interpreting and analyzing tabular data.

8. Chain of Prompt Narrative Generator

High level strategy for narrative generation: The method involves employing an instruction-based prompting approach using Azure Open AI's GPT-4 Model and incorporating an XML structuring system. This combination allows for the generation of instructions and XML based on sample documents, leading to completed sections that utilize instructions, XML, and user provided inputs. The XML tags offer optimal structuring with flexibility, while the instructions ensure content quality and relevance. Alongside these components, is the user input, which contributes to information completeness and quality.

Initial prompt engineering: Large language models (LLMs) excel at generating textual content; however, their capacity to generate domain-specific documents is limited due to insufficient training on specialized business content.

EY teams initially aimed to address this challenge by creating templates based on historical samples. Given a collection of similar example documentations, the LLM was first tasked with generating a template of the document to be produced, leaving variables blank. Subsequently, the LLM utilized the customized template and extracted information from relevant reference documents or user inputs to populate the template.

To improve clarity and prevent information overload, this process was divided into two distinct LLM calls: 1) Template generation, and 2) Information extraction and template propagation. The rationale for this separation stemmed from the need to manage multiple inputs effectively. In the template generation phase, the LLM required historical samples, while in the propagation phase, it needed reference documents such as the model guideline or user inputs. By separating these tasks, the LLM could better focus on the relevant inputs for each task, hence avoiding the loss of specificity that may result from handling a vast amount of information concurrently.

Key challenges: Templates in domain-specific document generation offer numerous benefits, such as guiding the LLM to produce content with the desired structure, style, and context. They act as a blueprint, encapsulating crucial information about the document without explicitly stating parameters. This allows the LLM to understand aspects like length, formality, and audience.

Using templates addresses potential issues from direct instructions, which may be misinterpreted or contradictory. For instance, specifying a word count could lead to suboptimal results, while templates help the LLM understand the reasoning behind it, ensuring coherent and contextually appropriate output.

However, the template-based approach has limitations, mainly due to its rigidity and resemblance to traditional automation processes. For example, the template might specify a certain number of variables when more are relevant. This can lead to suboptimal generations in cases where documents are similar but not identical. To address these limitations, future steps may involve refining the template generation process or exploring alternative strategies to ensure accurate and contextually appropriate domain-specific documents.

Refined prompt engineering: *Figure 3: Chain of Prompt Narrative Generator refined prompt flow and benefits*

PRELIMINARY SAMPLE

INSTRUCTION GENERATOR + **STRUCTURE GENERATOR** → **STRUCTURE GENERATION** → **BENEFITS**

INSTRUCTION GENERATOR

Generates a set of instructions for the model based on a previous sample.

Prompt

"

Receive a section from the document.
Split the section into subsections.
Provide each subsection as a separate instruction.
...
Example Output:
3.5
Step 1: ...
...

STRUCTURE GENERATOR

Creates the structure using XML tags of the document using other samples.

Prompt

"

Create a generalized structure for ANY section AIRB PD report.
Extract section structure using XML tags from completed section.
...
Example Output:
Word Count: 1000
Paragraphs: 10
<SECTION_STRUCTURE>
<SECTION>
<TITLE>3.1 Data Sources</TITLE>
<DATA_PROCESSING></DATA_PROCESSING>
...

STRUCTURE GENERATION

Using the generated instructions and structure, the model takes external input such as model guidelines or user notes to generate section.

Prompt

"

Generate AIRB PD
...
Output completed section without steps or instructions.
...
Produce a professional and complete output matching the word count.
Example Output:
3.1 Data Sources
...

BENEFITS

- **Improved Document Quality:** Refined prompt engineering ensures coherent, context-appropriate content with precise information, addressing template rigidity issues.
- **Enhanced Flexibility:** XML structuring system combined with instruction-based prompting balances clear structure and adaptability to various domain-specific requirements.
- **Better adherence to document standards:** This approach successfully maintains tone, context, and document organization, overcoming limitations of direct instructions and templates.

Open challenges: One key challenge is the accurate interpretation and execution of complex instructions by the LLM, as misinterpretations could lead to content that deviates from the desired outcome. Another challenge lies in the seamless integration of the XML structuring system with the generated content, as inconsistencies in formatting or structure could impact the overall coherence and readability of the documents. Additionally, maintaining content quality and relevance while ensuring flexibility in the XML tags could prove to be a delicate balancing act, as overly rigid structures might hinder the adaptability of the model to diverse document types and subject matters. Addressing these open challenges will be crucial for fully realizing the benefits of this methodology in streamlining the model documentation process.

Another, more practical challenge is the cost associated with the increased usage of LLMs, as more calls to the model may be required for generating comprehensive and accurate content. To mitigate this issue, a potential solution could involve employing cheaper, lower-performance models for handling more trivial tasks, while reserving the more expensive, high-performance models for complex tasks that demand greater precision and expertise.

This tiered approach to resource allocation would not only optimize costs but also ensure that the quality and efficiency of the generated content are maintained across different levels of document complexity.

9. Evaluating the effectiveness of document automation

The success of document automation was assessed based on three distinct criteria: the degree of automation (% automation), overall quality and consistency of generated output (document quality), and semantic similarity (relevancy).

- **% automation:** A comparative analysis of the generated output was performed against a manually written or human generated model documentation. The automation formula used was automated document word count / (automated document word count + missed or incorrect information word count), with the denominator representing the missed word count when compared to human-generated document. This metric was evaluated in two scenarios:

  - Without human input (60 - 70% automation) The approach entailed automatically generating the output leveraging only reference materials, without any intervention from the model developer.
  - With human input (70 - 80% automation): The same approach was followed, but with the inclusion of concise notes provided by the model developer via a brief questionnaire, requiring no more than one hour to complete.

- **Document quality:** Rigorous examination of the document's quality was carried out by experienced model validators at Ernst and Young (EY), who meticulously scrutinized and endorsed its content. An example comparing the LLM generated output and human generated output is shown below.

LLM generated sample section

The data used in the probability default document covers the historical loan level performance statistics for US (retail) portfolio from December 2003 to December 2022. This period includes a full credit cycle, the 2008-2009 economic recession, the benign phase from 2012 to 2019, and the COVID period from January 2020 to March 2021. The model development timeframe spans from October 2005 to October 2021.

Table 3.5 Production data shared by the modeling team - data timeframes

Data File Name	Samples	Data File Description	Number of provided Variables	Model Development Timeframe
EFD_MDL_ DATA	N/A	Monthly intermediate data for model	152	2005/10 - 2021/10

Human generated sample section

The provided model data covers the historical loan level performance statistics beginning in December 2003 and ending in December 2022. This historical period includes a full credit cycle, the 2008-2009 economic recession, a benign phrase from 2012-2019, and the COVID period from January 2020 to March 2021. Table 2-4 shows the detail information.

Table 3.5 Production US Data shared by MD Team - Data Timeframes

Data File Name	Samples	Data File Description	Number of provided Variables	Model Development Timeframe
EFD_MDL_ DATA	Total population	Monthly intermediate data for model (200312-202212)	152	200510 - 202110

- **Semantic similarity:** To gauge this aspect, a manually generated document was compared to the LLM generated document. Conducting a section-by-section comparison with cosine similarity revealed high similarities throughout the selected sub-sections of the document.

10. Document automation landscape

Future outlook

The landscape of innovation and technology offers a vast potential for advancements in information retrieval, LLM orchestration, and self-validation in generative AI projects.

Retrieval augmented generation

Utilizing vector databases allows for the storage of a significant number of documents. These documents can be indexed and retrieved instantaneously through cosine similarity, facilitating the extraction and retrieval of information from multiple sources, thereby streamlining the automation process.

LLM orchestration

The improved management of LLMs and modularization of different accelerators are key factors in enabling scalability across a diverse array of generative AI projects. By refining the organization and coordination of LLMs, researchers can optimize resource allocation, streamline model training processes, and facilitate seamless integration with other components, ultimately enhancing overall system performance.

Modularization of accelerators refers to the process of breaking down complex tasks into smaller, more manageable components or modules. This approach allows for greater flexibility, as each module can be developed, and refined independently before being integrated into the larger system. Furthermore, modularization promotes reusability, as individual modules can be adapted and repurposed for various projects, reducing development time and costs.

Self-validation

Expanding upon the concept of role-playing, we propose a critic role separation. Essentially, a separate LLM call that identifies error-prone areas and suggests fixes, which can be approved or manually adjusted before the document undergoes a final review. This approach serves multiple purposes, including identifying hallucination, vetting content quality, and checking for risks, redundancies, and missing information. While skepticism exists regarding an LLM checking its own work, we believe that role-playing provides enough distinction between automator and critic roles to ensure effectiveness; however, further research is necessary.

Seamless human validation

The integration of human oversight and validation in generative AI for documentation automation will continue to improve its overall effectiveness. By incorporating human experts seamlessly into the automation ecosystem, content validity and contextual awareness are ensured throughout the process. To facilitate an efficient human validation process, an intuitive and user-friendly interface is essential, as it can potentially reduce the workload on humans and enable them to make significant contributions, such as refining prompt design, expanding the prompt library, and conducting ongoing qualitative reviews.

Conclusion

The integration of generative AI, specifically large language models, into the model documentation process holds immense potential for streamlining and enhancing the overall efficiency and quality of document generation as demonstrated by the early performance metrics (70-80% automation) detailed in this article. By leveraging advanced natural language processing capabilities, LLMs can effectively comprehend and assimilate complex information, discern context, and generate human-like responses. The proposed methodology, including information extraction, historical sample referencing, comprehension of structure and task, prompting techniques, table data extraction and analysis, narrative synthesis, and SME review and enhancement, ensures the production of high-quality, contextually relevant, and professionally crafted documents. Furthermore, the implementation of retrieval augmented generation, LLM orchestration, self-validation, and enhanced contextual knowledge can significantly contribute to the evolution of the documentation automation efforts.

As the technology moves forward in this rapidly advancing field, it is essential to continue exploring innovative approaches and refining existing methodologies to harness the full potential of generative AI in automating model documentation. The collaboration between LLMs and human experts will pave the way for a more efficient, cost-effective, and robust model documentation process, ultimately transforming the way organizations approach and manage their documentation needs.

4261712.
ED None.

Authors

EY Canada

MARIO SCHLENER

Partner, Lead Financial Services Risk Management Practice and Enterprise Risk Strategy, EY Canada
EY Global FS Risk Technology, Alliance, Innovation Lead
mario.schlener@ca.ey.com

VISHAL GOSSAIN

Practice Leader, Risk Analytics and Strategy, Financial Services Risk Management, EY Canada
vishal.gossain@ca.ey.com

YARA ELIAS, PH.D.

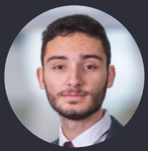
Senior Manager, AI Risk Lead, Financial Services Risk Management, EY Canada
yara.elias@ca.ey.com

KIRANJOT DHILLON

Senior Manager, AI Risk, Financial Services Risk Management, EY Canada
kiranjot.dhillon1@ca.ey.com

VISHAAL VENKATESH

Senior, AI Risk, Financial Services Risk Management, EY Canada
vishaal.venkatesh@ca.ey.com



AHMAD GHAWANMEH

Staff, AI Risk, Financial Services Risk Management, EY Canada
ahmad.ghawanmeh@ca.ey.com

Microsoft

GERARDO AMAYA

Senior Partner Technology Strategist at Microsoft, Canada
geamaya@microsoft.com

NIMA MAHAMOUD

Senior Cloud Solution Architect at Microsoft, Canada
nimamahamoud@microsoft.com