A Guardrail Framework for Sensitive Financial Information Protection: A Taxonomy-Driven Approach

Mehdi Yekrangi[1], Houssem Chatbri[1], Claudia Beatrice Chianella[1], Owen O'Neill[1]

[1]AI Hub, The Bank of New York

mehdi.yekrangi@bny.com, houssem.chatbri@bny.com, claudiabeatrice.chianella@bny.com, owen.o'neill@bny.com

Abstract

The increasing adoption of large language models in the financial sector introduces significant challenges related to the handling of sensitive financial information (SFI). Existing general-purpose content safety solutions, or guardrails, often fall short in detecting domain-specific risks inherent in financial data processing. This study addresses these gaps by developing a comprehensive taxonomy of SFI, grounded in globally recognized financial, information security, and AI governance standards. Leveraging this taxonomy, we synthesized an extensive dataset encompassing diverse categories of SFI and trained GARD (Generative Adversarial network Risk Detection) model to detect sensitive content in both inputs and outputs of GenAI systems within the financial domain. Our evaluation compared GARD against commercial guardrail solutions, including the OpenAI Moderation API and Microsoft Azure Content Safety (ACS). The results demonstrated that while commercial solutions maintained high precision, their recall was substantially lower, indicating many risky instances went undetected. In contrast, our model achieved a recall score of 0.98, significantly outperforming the benchmarks and enhancing SFI detection. These findings underscore the necessity of domain-specific guardrails tailored to the financial sector to ensure robust AI safety and compliance. In conclusion, this work contributes (1) A detailed taxonomy of SFI tailored for GenAI applications, (2) A comprehensive synthetic dataset that encompasses a wide range of sensitive topics relevant to the domain and (3) A high-performance risk detection model that can be deployed independently or alongside existing solutions to improve content safety in financial services. This approach promotes trust, mitigates financial, legal, and reputational risks, and supports the responsible adoption of GenAI technologies in sensitive domains.

Introduction

Advancements in large language models (LLMs) have significantly expanded their applicability across a wide range of domains. The capacity of LLMs to generate coherent, contextually relevant, and fluent output has made them an important component across various domains including finance, healthcare, education, etc (Zhao et. al. 2023). However, this widespread adoption also introduces specific challenges, particularly when LLMs are employed to handle sensitive information (SI) (Bommasani et. al. 2021). SI in the context of information processing is defined as data that, if disclosed, altered, or destroyed without authorization, could result in harm to individuals or organizations, including but not limited to personally identifiable information (PII), financial records, health data, and proprietary business information. Such information require special handling and protection measures to ensure confidentiality, integrity, and availability in accordance with regulatory and organizational requirements (ISO/IEC 27001 2022, Security and Privacy Controls for Information Systems and Organizations 2020). Due to the extensive scale of their training datasets, LLMs may inadvertently encompass personal, confidential, or proprietary information. Thus, they may unintentionally generate responses that reveal such sensitive data (Lu et. al. 2023). Adversaries may leverage techniques such as model inversion or alternative approaches to extract SI from LLMs. For instance, an attacker could use a sequence of carefully designed prompts to retrieve social security numbers or private communications that the model may have acquired during its training process (Carlini et. al. 2020). To mitigate the aforementioned risks, guardrails are implemented to enforce constraints on the inputs and outputs of LLMs. Guardrails are mechanisms or frameworks that ensure LLMs' input/output are aligned with specified safety, risk, and policy requirements. In the rest of this paper, we use the term *guardrail* to refer to any AI safety mechanism applied to the inputs or outputs of generative AI (GenAI) systems. To assess safety of information processing in GenAI systems, we ground our study in a real-world use case of financial services and examine how commercial products may fall short in addressing domain-specific risks. Financial services are a primary area for GenAI systems, yet there has been limited discussion regarding AI safety in this context (Nie et. al. 2024). The security and privacy of financial data are paramount due to the substantial

risks associated with data breaches and regulatory non-compliance. Implementing LLMs in the financial sector introduces distinct challenges in ensuring strong data protection and effectively safeguarding SI (Nie et. al. 2024). Since financial data can be highly sensitive, deploying LLMs poses unique challenges in protecting this data from breaches and compliance violations. Hence, strong cybersecurity measures and privacy protocols are essential to mitigate risks of data leaks. To address this, we examine and apply relevant regulatory and risk frameworks in financial services (Gramm-Leach-Bliley Act-106th Congress Public Law 106–102 1999, Council and Standards 2024, Council and Examination, Information Security 2016), information processing (ISO/IEC 27001 2022, Security and Privacy Controls for Information Systems and Organizations 2020, McCallister, E; Grance, T; Scarfone, K 2010, Ross, R; Pillitteri, V 2024), and cybersecurity (OWASP Top 10 for LLM Applications 2025) toward developing a comprehensive taxonomy of sensitive topics within the domain. A well-defined taxonomy enables systematic categorization and understanding of various types of SFI, facilitating more effective monitoring and control.

Given the complexity and potential risks associated with GenAI systems—such as data breaches, unintended financial harm, and legal accountability—risk detection models are necessary to identify and mitigate vulnerabilities arising from the use of these systems. These models can help detect anomalies, prevent data leakage, and ensure compliance with privacy regulations. The competitive and high-stakes nature of finance demands clear frameworks for responsibility and ethical use, which are supported by taxonomy-driven risk management. Together, these efforts promote trust, enhance the safe adoption of GenAI technologies, and safeguard stakeholders against financial, legal, and reputational risks inherent in handling SFI. Our study demonstrates that general-purpose content safety solutions, known as guardrails for GenAI systems, may be insufficient for detecting risks associated with SI processing in financial services. Therefore, we take a comprehensive approach to outline various types of SI in this area and develop a risk detection model, capable of identifying such information in both inputs and outputs of GenAI systems, thereby enhancing their safety in this context. Henceforth, we refer to texts and samples that contain sensitive financial information (detailed in the appendix) as *risky*. In the remainder of this paper, we begin by providing a brief overview of related work in Section 2. Section 3, details our methodology for developing a taxonomy of SI and constructing the corresponding risk detection model. In Section 4, we present and evaluate our results in comparison with commercial content safety solutions. Finally in Section 5, we conclude by summarizing our key findings and offering recommendations for enhancing GenAI systems within the financial services sector.

Background

Academic research to date has predominantly concentrated on a limited range of general risk categories as almost two thirds of evaluations of LLMs focus on misinformation, toxicity, and representational harm (Rauh et. al. 2024). However, when LLMs are deployed across various domains, domain-specific risks may emerge. Common approaches to prevent unauthorized disclosure of SI by LLMs encompass data sanitization, differential privacy, and output filtering. Data sanitisation, which involves removing SI from training data, is a key mitigation strategy to reduce the risk of LLMs memorising and regurgitating such information. However, this is challenging in practice due to the scale of training data and the difficulty of identifying all sensitive content (Carlini et. al. 2020). Another approach as output filtering is a post-processing step that inspects the model's outputs and discards or modifies those that are deemed unsafe or otherwise undesirable before they are shown to the user (Ganguli et. al. 2022). Differential privacy, as another approach by adding mathematical noise during training ensures that the output of a computation does not significantly change with the addition or removal of a single data point, thereby preventing the model from memorizing or revealing information about individuals (Wilson et. al. 2019). This paper focuses on both the inputs and outputs of GenAI systems in the context of their commercial deployment for domain-specific applications. Accordingly, our approach encompasses not only output filtering but also the examination and management of inputs to LLMs. The existing guardrails predominantly address risks such as prompt injection[1], jailbreaking[2], harmful content categories, and misinformation without focus on domain-specific risks (Nie et. al. 2024). While these areas are critical in establishing a foundational guardrail infrastructure for LLMs, a substantial gap persists in mitigating one of the most significant risk categories—SI disclosure, as identified second in the OWASP risk framework for LLMs (OWASP Top 10 for LLM Applications 2025).

Recent advancements in LLMs have raised significant concerns regarding the inadvertent disclosure of SI, prompt-

[1] Prompt injection is an attack technique in which an adversary crafts inputs designed to manipulate a language model's behaviour, often causing it to ignore original instructions or generate unintended, potentially harmful outputs

[2] Jailbreaking refers to the process of manipulating a language model through specially crafted prompts to bypass its built-in safety mechanisms and content restrictions, enabling the generation of outputs that would otherwise be blocked.

Source	Risk category	Sample
Gemini 1.5-Pro, Mixtral-8x7B, GPT 4o, Llama 3.1, Eliza prompts[3]	Stakeholders' Topics (see appendix 2), Prevalent Financial Topics (see appendix 3), Personally Identifiable Information (PII), Authentication and Security Data, Business and Proprietary Information, Customer Financial Data, Financial Account Information, Non-public Personal Information, Payment Card Data, Regulatory and Compliance Data, Health-Related Financial Data, Operational logs, Other Sensitive Data	Synthetic: 15,360 Real: 2,135
Total unique risky samples: 10,098	Total unique risky/non-risky samples:17,495	

Table 1: The high-level topics and categories of the dataset.

ing a surge in research focused on risk detection and mitigation. (Lehman et. al. 2021) explored the risks of sensitive data leakage in medical LLMs, finding that models trained on clinical notes can output confidential patient information, even when explicit filtering mechanisms are in place. These findings underscore the persistent challenge of reliably detecting and mitigating SI risks in LLM outputs. In another related work (Nasr et. al. 2023) investigated extractable memorization in language models, demonstrating that adversaries can efficiently extract large amounts of training data—including gigabytes—from both open-source and proprietary models through targeted queries. Their findings reveal that even alignment techniques, such as those used in ChatGPT, do not fully prevent memorization, as novel attacks can significantly increase the leakage of sensitive training data. To address SI disclosure risks, approaches such as user trust profiling have been proposed (Feretzakis, G; Verykios, V S 2024) in a trust framework, enabling policy-based access control tailored to users' roles and contextual factors. Although they have determined SI detection as a major module in their framework, but there is no empirically validated mechanism for detecting SI introduced. Some other efforts (Gehrmann et. al. 2025) have been made to create a risk taxonomy associated with GenAI in financial services, but they have not specifically focused on the granular topics related to SI within this sector. Also, the resulting taxonomies have not been grounded in the established risk frameworks for the domain. Hence, there remains a significant gap in the development of domain-specific risk detection models targeting this particular risk in GenAI systems with a granular categorization of SI.

One of the major challenges in this area is the absence of available datasets pertaining to SFI, which complicates the development of robust guardrails models in this domain. This stems from the reluctancy of organizations in sharing such data. To address this gap, we leverage our holistic taxonomy to create a comprehensive dataset encompassing all SI topics with a particular focus on the financial sector. Accordingly, this study makes three key contributions to the domain: it 1) develops a grounded taxonomy for SFI. 2) introduces a comprehensive synthetic dataset that encompasses a wide range of sensitive topics relevant to the domain, and 3) constructs an empirically validated risk detection model specifically focused on SFI detection.

Method

Our approach is structured around three primary components: 1) synthetic data generation, 2) embedding, and 3) predictive modelling. Each component plays a distinct and essential role within our proposed methodology, as detailed in this section.

Synthetic Data Generation

Accessing datasets containing SI, particularly within the financial domain, presents a significant challenge. This constraint motivates us to tackle the issue through two complementary approaches. The first approach is embodied in the taxonomy we have developed for synthetic data generation encompassing a wide range of topics in this area. The second involves the generative adversarial network (GAN) model we propose specifically for the task of not only detecting texts including SI but also generating vectors similar to those texts. In this section, we focus on synthetic data generation, with the GAN model and its role in addressing this challenge discussed subsequently in the predictive modelling section. To construct a comprehensive dataset of SFI, it is essential to develop a coherent taxonomy categorizing the different types of SFI, along with a precise delineation for each type. Additionally, our focus on financial information necessitates the identification of prevalent keywords in this domain, such as *Futures*, *Credit Cards*, *Swaps* etc. To address these requirements, we collate related guidelines and frameworks (Gramm-Leach-Bliley Act-106th Congress Public Law 106–102 1999, Council and Standards 2024, Council and Examination, Information Security 2016, ISO/IEC 27001 2022, Security and Privacy Controls for Information Systems and Organizations 2020, McCallister, E; Grance, T; Scarfone, K 2010, Ross, R; Pillitteri, V 2024, ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system 2023) in the financial domain as well as information management to protect SI. We also direct synthetic data creation via guiding LLMs by prevalent terms in financial and banking fields (Zuo, X; Jiang, A; Zhou, K 2024). Other resource we consider for the

[3] Eliza is a secure, enterprise-grade AI interface at Bank of New York (BNY) that facilitates responsible access to generative AI tools for authorized users. (https://www.bny.com/corporate/global/en/about-us/technology-innovation/artificial-intelligence.html)

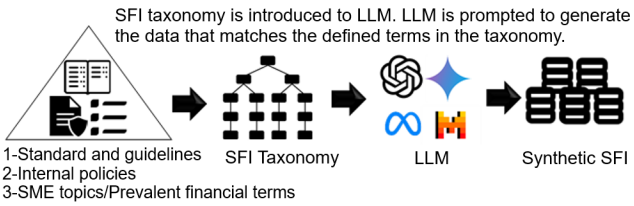

Figure 1: Synthetic data generation process. The approach utilizes diverse resources to construct a comprehensive taxonomy, which then serves as the basis for data generation.

development of this dataset are BNY's internal policies concerning information classification, handling, and records management. Besides the introduced standards and frameworks, the specified guidelines within the BNY's policies provide a foundation, upon which we can elaborate the principles and controls, particularly by defining the requirements and controls for information classification, records management, and data incident management. By structuring and building our dataset in accordance with the consolidated established resources on SFI, we further introduce a domain-specific taxonomy highlighted in Appendix 1. This taxonomy is intended to address different risks involved in SFI while also facilitating future research in the domain. With the taxonomy established, we utilized a range of LLMs—including Gemini 1.5-Pro, Mixtral-8x7B, GPT-4o, and Llama 3.1—to generate input and output data corresponding to the key topics identified in the taxonomy. Additionally, these models were employed to produce non-risky data within this domain while real prompts from Eliza (the BNY's GenAI system) also incorporated. In total, the dataset consists of 17,495 domain-specific samples, encompassing both risky and non-risky instances.

Designed prompt	Synthetic data (output guardrail)
Generate a synthetic list containing anonymized credit card numbers, account numbers, and associated transaction details for testing fraud detection systems.	{"CardNumber": "4532-1234-5678-9012", "ExpirationDate": "12/28", "CardHolderName": "John Doe", "CVV": "123"} {"AccountNumber": "9876543210", "BankName": "Example National Bank", "AccountType": "Checking", "Balance": "$10,500.75"}
Create a list of anonymized tax records, including taxpayer IDs, income brackets, and filing statuses, for use in compliance testing.	{"taxpayer_id": "963412777", "income_bracket": "40k-60k", "filing_status": "Married Filing Jointly"},
Synthetic prompt (input guardrail)	
What is the balance of X fund? *Can you give me the account number of our supplier payments?*	

Table 2: Sample generated data including/inquiring about SFI.

The process of generating SFI dataset is illustrated in Figure 1. For instance, according to NIST SP 800-122 (McCallister, E; Grance, T; Scarfone, K 2010) framework, we design prompts that align with the definition of PII and confidential data. These prompts are carefully crafted to include examples and categories of PII relevant to the financial domain, such as account numbers, credit card numbers, and tax information. We also incorporate internal policies and topics highlighted by subject matter experts (SMEs) that align with this guideline to generate data that include/inquire about sensitive information. For example, the synthetic data illustrated in Table 2 are designed based on this framework. As a result, we can generate synthetic data that is designed in accordance with the definitions and requirements for sensitive data protection outlined in NIST SP 800-122, internal policies, and SME-highlighted topics. These examples demonstrate how synthetic data/prompt are generated to simulate sensitive information while adhering to the necessary guidelines and ensuring data protection.

Dataset Validation

We assess the synthetic dataset by examining its alignment with real-world risky samples. To do so, we compiled a dataset of 115 real-world risky samples, all of which were validated as sensitive information by three SMEs. This dataset served as the reference for assessing the accuracy of the synthetic data and its similarity to real-word risky samples. Hence, we measured the average embedding (discussed in the following) of the synthetic dataset against the average embedding of the reference dataset by computing the cosine similarity between the two. As a result, the comparison yielded a similarity score of 0.97 between the two datasets. The similarity score indicates a high degree of alignment between the synthetic and the real-world risky data, demonstrating the effectiveness of the synthetic data generation process in capturing the characteristics of the reference dataset.

Embedding

To effectively capture both prominent and nuanced distinctions between texts including SI (risky) and not including SI (non-risky) within the dataset, it is essential to employ an embedding technique capable of representing these variations in a high-dimensional vector space. For this purpose, we utilize OpenAI text-embedding-ada-002 (New and improved embedding model OpenAI. 2022.). This embedding method is selected over alternatives such as BERT due to its demonstrated enhanced performance in capturing semantic relationships and classification tasks (New and improved embedding model OpenAI. 2022., Muennighoff, N; Tazi, N; Magne, L; Reimers, N 2023). The vectors produced by this embedding have a length of 1,536 dimensions. Consequently, all samples of the dataset are transformed into these high-dimensional vectors, enabling their subsequent processing by the GAN model.

Predictive Modelling

Given the comprehensive nature of our dataset, which encompasses a wide range of SI pertinent to the financial and banking sector, GARD gets trained with a holistic perspective on risk categories and their associated feature vectors. As previously noted, addressing the scarcity of real-world sensitive datasets in the financial sector is a prerequisite for developing a model that can detect diverse types of SI. This is achieved, first, by developing a comprehensive taxonomy that encompasses sensitive topics relevant to the domain, and second, by utilizing synthesized data as input for the GAN model to generate similar vectors based on the input data. We employ GANs due to their capability for continuous learning in both data classification and data generation. This allows the model to produce vectors analogous to those representing sensitive text, including instances that may not be present in the original dataset, which ultimately makes the classification more robust in detecting SI. The advantage of utilizing GANs lies in their continuous generation and discrimination cycle, enhancing performance in risk detection. Our objective is to identify any text including SI. In this context, the GAN model's generator and discriminator engage in a competitive process to surpass one another in performance. This iterative cycle drives the generator to produce vectors that closely resemble those of including SI, while remaining undetectable by the discriminator. Concurrently, the discriminator is trained to distinguish between risky and non-risky vectors generated by the generator to learn to recognize distinguishing features associated with the risky vectors.The GAN model consists of a generator and a discriminator, both constructed as multilayer feedforward neural networks using PyTorch. The generator receives a random noise vector and outputs synthetic embeddings matching the dimensionality of the real data, while the discriminator distinguishes between real and generated embeddings using a series of linear layers with LeakyReLU activations and dropout for regularization. The dataset, comprising OpenAI-generated embeddings labelled as *risky*/*non-risky*, is balanced through resampling to address class imbalance. Labels are encoded numerically, and both embeddings and labels are converted to PyTorch tensors for model training. The models are trained adversarially using the binary cross-entropy loss function and Adam optimizer, with label smoothing applied to real samples to improve discriminator robustness. The training process alternates between updating the discriminator to better distinguish risky embeddings and updating the generator to produce more realistic embeddings. Model performance is monitored over 200 epochs, and the final trained discriminator is saved for subsequent evaluation and application. The GAN model's training/testing process is illustrated in Figure 2.

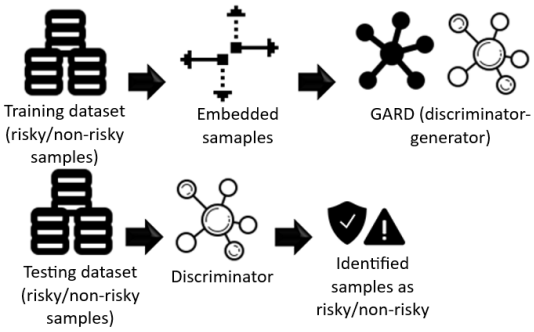

Figure 2: Training and testing GARD for risk detection using the generated dataset.

Results and Evaluation

In order to evaluate the performance of GARD, we consider guardrail systems provided by different vendors including OpenAI Moderation API (OpenAI, Moderation- OpenAI API 2023) and Microsoft Azure Content Safety (ACS) (Microsoft 2023). The OpenAI Moderation API is an automated content moderation tool that analyses text to detect and flag content that may violate safety policies, such as hate speech, violence, or SI. This tool is inherently integrated into OpenAI LLMs to automatically monitor and filter inputs and outputs for content that may violate safety policies. The other guardrail system as Microsoft ACS is a cloud-based service that uses advanced AI models to detect, classify, and filter harmful, offensive, or sensitive content—including text and images—in real time, helping organizations ensure compliance with safety and moderation policies across their applications. These guardrail systems are designed and developed to evaluate the inputs and outputs of the LLMs for potential violations of safety policies, including harmful content categories, hate speech, privacy breaches such as the disclosure of PII, and other sensitive or inappropriate material. Their implementation helps ensure responsible and secure deployment of GenAI in real-world applications. Hence, we conducted evaluations using both the ACS and the OpenAI Moderation API and compared their results with those obtained by GARD. In addition to its default AI classifier for content monitoring, ACS provides a blocklist feature. A blocklist is a customizable list of specific words, phrases, or patterns that an organization wants to automatically detect and block in user-generated content. When content matches any entry in the blocklist, ACS flags or filters it according to the configured moderation policies, helping to prevent the display or processing of undesirable or sensitive material. Therefore, the evaluation on ACS was conducted over multiple rounds of testing. In the initial round, no specific instructions or blocklists were provided to the ACS based on ACS's assertion that its default AI classifiers are generally adequate for most content moderation needs.

Guardrail	Performance				
	Acc	Prec	Rec	F1	Latency (X/sec)
ACS 1	0.47	1.00	**0.01**	0.02	0.3
ACS 2	0.66	1.00	**0.37**	0.54	
OpenAI API	0.62	1.00	**0.30**	0.46	0.5
GARD	0.92	0.86	**0.98**	0.92	50

Table 3. ML metrics of commercial guardrails vs. GARD.

In the subsequent round, a blocklist including the SI categories (see appendix) was introduced to the ACS as system prompts, enhancing its effectiveness for our particular use case involving SI. Consequently, we evaluated each tool using 3,552 samples as test dataset, comprising 1,902 *risky* (comprising various SI) and 1,650 *non-risky* instances. Having the test conducted on the dataset, the performance of different models is illustrated in Table 3. For the vendor solutions (ACS and the OpenAI Moderation API), precision remains high—indicating that all filtered instances are indeed risky—while recall is as low as 0.30 and 0.37 for OpenAI and ACS, respectively. This suggests that a significant portion of risky instances are not being detected or filtered. In contrast, GARD achieves a high recall score, indicating that it successfully detects most risky instances; however, this comes at the expense of precision, as it also produces some false positives. Nevertheless, the recall score is of utmost importance in risk detection application, as missing risky instances can have significant negative consequences (Chandolla, V; Banerjee, A; Kumar, V 2009). As demonstrated by the results, the primary objective in developing the risk detection model is to maximize coverage of risky instances through the introduction of a comprehensive taxonomy. This taxonomy allows the model to be exposed to a diverse range of vectors that include SI (e.g., payment card data) or are associated with SI (e.g., inquiries pertaining to payment card data). Consequently, the discriminator is capable of distinguishing between features linked to risky and non-risky vectors. The results indicate that commercially developed guardrail systems may not be ideally suited for domain-specific applications, such as those involving SFI. By training GARD on our domain-specific dataset, the model is optimized to address the particular risks associated with SFI in GenAI systems within the financial and banking sector. Another notable improvement is the reduced latency of GARD in comparison to the ACS and OpenAI Moderation API guardrails. Although the latency of commercial guardrail solutions may fluctuate depending on network conditions and implementation specifics, our experiments demonstrated average processing rates of 0.3 and 0.5 inputs per second for ACS and the OpenAI Moderation API, respectively. In contrast, GARD achieved a substantially higher throughput, processing an average of 50 inputs/sec.

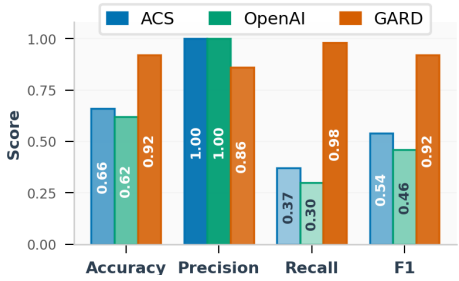

Figure 3: ML metrics of different guardrails.

Conclusion and Future Work

While there have been different guardrails provided for GenAI inputs/outputs, the topic of SI remains to be a challenge from an AI governance framework particularly for the organizations working with sensitive financial data. Furthermore, the scarcity of publicly available datasets in this domain complicates the development of effective risk detection models. To address these challenges, this study proposed a comprehensive taxonomy of SFI, grounded in globally recognized guidelines and standards for financial services as well as information security and AI systems. This taxonomy facilitated the synthesis of an extensive dataset encompassing a wide range of SI topics relevant to the financial sector. Consequently, we successfully trained a GAN model to detect data including SI with a 0.98 recall score outperforming commercial benchmarks. Hence, in this study we 1) introduced a broad SFI taxonomy tailored to GenAI inputs and outputs within the financial sector, 2) developed a comprehensive dataset of SFI and 3) trained GARD as a risk detection model designed to identify SFI in content processed by GenAI applications in this domain. The SFI taxonomy enables organizations to systematically monitor GenAI inputs and outputs according to the associated risk topics. Additionally, GARD can be deployed either independently or in conjunction with other solutions to enhance overall content safety. This study has been primarily conducted using synthetic data. Although extensive care was taken to generate data that closely resembles real-world cases, there remain inherent limitations associated with synthetic datasets. These include potential gaps in capturing the full complexity and variability of actual SI encountered in practice. As a result, further validation using real-world data is necessary to fully assess the generalizability and robustness of the proposed model. Another important consideration for future models is their scalability with respect to user personas. Incorporating this feature into GenAI applications within organizations would enable differentiated access to GenAI systems based on users' roles and associated attributes. Through such user profiling, not only can SI be distinguished from other data types, but specific personas may also be granted privileged access to certain information, while access for others remains restricted according to their profile.

Appendix 1

A comprehensive list of information types classified as sensitive in the finance and banking sector is provided in Table 1. This taxonomy is based on major guidelines and frameworks including GLBA, PCI DSS, FFIEC, ISO 27001, and NIST.

Type	Information
Personally Identifiable Information (PII)	●Full name ●Address (home, business, mailing) ●Date of birth ●Social Security Number (SSN) or National ID ●Driver's license or government-issued ID numbers ●Phone numbers ●Email addresses
Financial Account Information	●Bank account numbers (checking, savings, investment) ●Credit card and debit card numbers- Account login credentials (usernames, passwords, PINs) ●Routing numbers ●Account balances ●Transaction histories ●Loan and mortgage account numbers
Payment Card Data (PCI DSS)	●Primary Account Number (PAN) ●Cardholder name● Expiration date ●Service code ●Card Verification Value (CVV/CVC) ●Sensitive authentication data (magnetic stripe data, PIN blocks)
Customer Financial Data	●Credit reports and credit scores ●Income and employment information ●Tax identification numbers ●Tax returns and related document ●Investment and portfolio information ●Insurance policy numbers and details
Non-public Personal Information (GLBA)	●Any information provided by a consumer to obtain a financial product or service ●Any information resulting from a transaction or service performed for the consumer ●Any information otherwise obtained about a consumer in connection with providing a financial product or service
Business and Proprietary Information	●Trade secrets ●Business plans and strategies ●Internal financial statements and reports ●Mergers and acquisitions data ●Intellectual property
Regulatory and Compliance Data	●Suspicious Activity Reports (SARs) ●Know Your Customer (KYC) documentation ●Anti-Money Laundering (AML) records
Authentication and Security Data	●Security questions and answers ●Biometric data (fingerprints, facial recognition) ●Digital certificates and cryptographic keys
Health-Related Financial Data	●Health insurance policy numbers ●Medical billing and payment information
Other Sensitive Data	●Beneficiary information ●Power of attorney documents ●Trust and estate documents

Table 4: Detailed taxonomy of sensitive information in financial and banking sectors.

Appendix 2

To develop a comprehensive dataset of SFI and its associated categories that may result in SFI disclosure, subject matter experts (SMEs) were consulted to identify and define the relevant terms and keywords to be included in the dataset. The list highlighted by SMEs includes the following terms.

liquidity management	cash reserves	credit rating
creditworthiness	credit risk	loan default
credit risk assessment	value at risk	market trend
hedging strategy	crisis management	stakeholder trust
intrusion	exploit	phishing
zero-day	backdoor	malware
DDoS	SQL injection	brute force
password	social security number	SSN
username	pin	confidential
identity		

Table 5: Common terms and keywords in sensitive information, highlighted by SMEs.

Appendix 3

The synthetic data generation process was further guided by incorporating commonly used terms from the financial and banking sectors (Zuo, X; Jiang, A; Zhou, K 2024). Consequently, the curated list was provided to the LLMs to facilitate the generation of content relevant to SFI.

business valuation	book value	market value
liquidation value	replacement value	market research
quantitative research	qualitative research	market segmentation
target market	financial derivatives	credit card
options	futures	swaps
forward contract	counterparty risk	annuities
ordinary annuities	annuity due	perpetuity
present value of annuity	personal finance	budget
savings	expenses	income
emergency fund		

Table 6: Common terms in the financial and banking documents.

Acknowledgements

This work was carried out within the BNY AI Hub.

References

Bommasani, R; Hudson, D A; Adeli, E; Altman, R; Arx, S V; Bernstein, M S. 2021. "On the Opportunities and Risks of Foundation Models." *arXiv preprint. arXiv:2108.07258.* doi:https://doi.org/10.48550/arXiv.2108.07258.

Carlini, N; Tramer, F; Wallace, E; Jagielski, M; Herbert-Voss, A; Lee, K; Roberts, A; Brown, T; Song, D; Erlingsson, U; Oprea, A; Raffel, C. 2020. "Extracting Training Data from Large Language Models."*arXivpreprint.arXiv:2012.07805.* https://doi.org/10.48550/arXiv.2012.07805.

Chandolla, V; Banerjee, A; Kumar, V. 2009. "Anomaly detection: Asurvey."*ACMComputingSurveys*41(3).doi:10.1145/1541880.1541882.

Council, and Federal Financial Institutions Examination. 2016. "Information Security." https://ithandbook.ffiec.gov/it-booklets/information-security. Accessed 2025-6-15.

Council, and Payment Card Industry Security Standards. 2024. "PCI DSS: v4.0.1." https://docsprv.pcisecuritystandards.org/PCI%20DSS/Standard/PCI-DSS-v4_0_1.pdf. Accessed 2025-6-15.

Feretzakis, G; Verykios, V S. 2024. "Trustworthy AI: Securing Sensitive Data in Large Language Models." *arXiv preprint. arXiv:2409.18222.* doi:https://doi.org/10.3390/ai5040134.

Ganguli, D; Lovitt, L; Kernion, J; Askell, A; Bai, Y; Kadavath, S; Mann, B. 2022. "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned." *arXiv preprint.arXiv:2209.07858.* doi:https://doi.org/10.48550/arXiv.2209.07858.

Gehrmann, S; Huang, C; Teng, X; Yurovski, S; Shode, I; Patel, C C; Bhorkar, A; Thomas, N; Doucette, J; Rosenberg, D; Dredze, M; Rabinowitz, D. 2025. "Understanding and Mitigating Risks of Generative AI in Financial Services." *arXiv preprint. arXiv:2504.20086.* doi:https://doi.org/10.48550/arXiv.2504.20086.

Gramm-Leach-Bliley Act-106th Congress Public Law 106–102. 1999.https://www.govinfo.gov/content/pkg/PLAW106publ102/pdf/PLAW-106publ102.pdf. Accessed 2025-6-15.

ISO/IEC 27001. 2022. https://www.iso.org/standard/27001. Accessed: 2025-07-15.

ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. 2023. https://www.iso.org/standard/81230.html. Accessed: 2025-07-15.

Lehman, E; Jain, S; Pichotta, K; Goldberg, Y; Wallace, B C. 2021. "Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?" *arXiv preprint. arXiv:2104.07762.* doi:https://doi.org/10.48550/arXiv.2104.07762.

Lu, Y; Yao, Y; Ton, J F; Zhanng, X; Guo, R; Cheng, H; Klochkov, Y; Taufiq, M F; Li, H. 2023. "Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment." *arXiv preprint. arXiv:2308.05374.* https://doi.org/10.48550/arXiv.2308.05374.

McCallister, E; Grance, T; Scarfone, K. 2010. "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)." https://doi.org/10.6028/NIST.SP.800-122. Accessed 2025-6-15.

Microsoft. 2023. "Azure AI Content Safety." https://learn.microsoft.com/en-us/azure/ai-services/content-safety. Accessed: 2025-07-15.

Muennighoff, N; Tazi, N; Magne, L; Reimers, N. 2023. "MTEB: Massive Text Embedding Benchmark." *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics.* Dubrovnik: Association for Computational Linguistics. 2014–2037. doi:10.18653/v1/2023.eacl-main.148.

Nasr, M; Carlini, N; Hayase, J; Jagielski, M; Cooper, A F; Ippolito, D; Choquette-Choo, C A; Wallace, E; Tramèr, F; Lee, K. 2023. "Scalable Extraction of Training Data from (Production) Language Models." *arXiv preprint. arXiv:2311.17035.* doi:https://doi.org/10.48550/arXiv.2311.17035.

OpenAI. 2022. "New and improved embedding model."https://openai.com/index/new-and-improved-embedding-model. Accessed 2025-6-15.

Nie, Y; Kong, Y; Dong, X; Mulvey, J M; Poor, V H; Wen, Q; Zohren, S. 2024. "A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges." *arXiv preprint. arXiv:2406.11903.* doi:https://doi.org/10.48550/arXiv.2406.11903.

OpenAI. 2023. "Moderation- OpenAI API." https://platform.openai.com/docs/guides/moderation. Accessed 2025-6-15.

OWASP Top 10 for LLM Applications. 2025. https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-v2025.pdf. Accessed 2025-6-15.

Rauh, M; Marchal, M; Manzini, A; Hendricks , L A; Comanescu, R; Akbulut, C; Akbulut, T; Mateos-Garcia, J; Bergman, S; Kay, J; Griffin, C; Bariach, B; Gabriel, I; Rieser, V; Isaac, W; Weidinger, L. 2024. "Gaps in the Safety Evaluation of Generative AI." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* San Jose, California. 1200-1217. https://doi.org/10.1609/aies.v7i1.31717.

Ross, R; Pillitteri, V. 2024. "Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations." https://doi.org/10.6028/NIST.SP.800-171r3. Accessed 2025-6-15.

Security and Privacy Controls for Information Systems and Organizations. 2020. https://doi.org/10.6028/NIST.SP.800-53r5. Accessed 2025-6-18.

Wilson, R J; Zhang, C Y; Lam, W; Desfontaines, D; Simmons-Marengo, D; Gipson, B. 2019. "Differentially Private SQL with Bounded User Contribution." *arXiv preprint. arXiv:1909.01917.* doi:https://doi.org/10.48550/arXiv.1909.01917.

Zhao, W X; Zhou, K; Li, J; Tang, T; Wang, X; Hou, Y; Min, Y; Zhang, B; Zhang, J; Dong, Z; Du, Y; Yang, C; Chen, Y; Chen, Z; Jiang, J; Ren, R; Li, Y; Tang, X; Liu, Z; Liu, P; Nie, J Y; Wen, J R. 2023. "A Survey of Large Language Models." *arXiv preprint. arXiv:2303.18223v16.* https://doi.org/10.48550/arXiv.2303.18223.

Zuo, X; Jiang, A; Zhou, K. 2024. "Reinforcement prompting for financial synthetic data generation." *The Journal of Finance and Data Science* 10. doi:doi.org/10.1016/j.jfds.2024.100137.