Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Prompt Engineering or Prompt Fraud? Governance Challenges for Audit

Karishma Velisetty
Independent Researcher, USA

ARTICLE INFO

Received: 27 Feb 2026

Accepted: 02 March 2026

ABSTRACT

Generative Artificial Intelligence (GenAI) has rapidly become a transformative tool across business functions, including finance, internal audit, and compliance. However, its adoption introduces novel risks that existing frameworks are not fully equipped to address. This article defines prompt fraud as the intentional manipulation of AI prompts to produce outputs that bypass traditional internal controls and generate misleading or fraudulent artifacts. Unlike conventional fraud, which targets systems or personnel through established attack vectors, prompt fraud exploits linguistic controls at the reasoning layer of GenAI systems. The concept represents a paradigm shift in how fraud can be perpetrated, as it requires no system-level intrusion, no credential compromise, and no technical exploitation of software vulnerabilities. Instead, it uses the natural language features of large language models to create responses that sound convincing, include false information, or tell misleading stories meant to trick auditors and decision-makers. This article explores the evolving threat landscape surrounding prompt fraud, provides a structured audit framework for its detection and prevention, assesses the control weaknesses that make organizations vulnerable, and proposes mitigation strategies grounded in governance, technology, and human oversight. The paper further examines the roles of internal and external threat actors, the implications of Shadow AI, and the regulatory and ethical dimensions of AI-assisted fraud. It ends by suggesting that organizations should use better audit methods, strong AI management systems, and ongoing monitoring to deal with the fast-changing risks from GenAI in business settings.

Keywords: Prompt Fraud, Generative Artificial Intelligence, Internal Audit, AI Governance, Fraud Risk Management

1. Introduction and Background

1.1 The Rise of GenAI in Business and Audit

More and more businesses and auditors are using generative AI systems, such as transformer-based large language models (LLMs), because they can understand and process language, combine complex information, and automate everyday tasks. These systems have demonstrated remarkable proficiency in generating human-like text, summarizing voluminous documents, drafting correspondence, and producing analytical narratives that support decision-making across organizational functions. The adoption of GenAI within finance, compliance, and internal audit has accelerated as organizations seek efficiency gains, cost reductions, and enhanced analytical capabilities. However, the very characteristics that make these systems valuable also introduce unprecedented vulnerabilities. The ability of LLMs to produce fluent, contextually appropriate, and authoritative-sounding text means that their outputs can be readily accepted as credible by human reviewers, even when the underlying content is fabricated, misleading, or intentionally crafted to deceive. The effect of GenAI on internal auditing is considered an area that needs thorough study and understanding, especially since this technology is changing how audit evidence is created, checked, and trusted in organizations. These worries are increased because LLM-based applications are very susceptible to prompt injection attacks, and the Open Web Application Security Project (OWASP) lists harmful prompt exploitation as one of the top 10 risks for applications that use LLMs.

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1.2 Purpose, Scope, and Research Gap

Even though GenAI offers operational advantages, new findings show that these systems create new ways for fraud that current audit methods don't fully consider. Traditional internal controls assume that fraud requires either system-level access, credential compromise, or direct manipulation of records and documents. Prompt Fraud contests these assumptions by presenting a method of manipulation that functions solely at the linguistic interface between human users and AI systems. As many organizations rely on GenAI to support documentation, approvals, and decision-making, auditors and risk professionals must expand their methodologies to address this evolving threat. Professional organizations have created guidelines to help auditors evaluate risks related to AI, emphasizing the importance of having organized and systematic methods for managing AI within audit processes. The need for this response is highlighted by the OWASP Top 10 list for large language models, which points out 10 major weaknesses that can create serious security problems for applications using these advanced technologies, with prompt injection being the biggest worry. This paper intends to describe Prompt Fraud clearly as a specific type of fraud, look at the dangers it brings, find the weaknesses in controls that allow it to happen, and suggest a way to audit and a control system to reduce this new and possibly widespread risk.

2. Theoretical Foundations and Literature Review

2.1 Risk Landscape of GenAI and Fraud Theory

Research on AI-related fraud has mainly looked at how to spot deepfakes and fake identities created with Generative Adversarial Networks (GANs), showing how AI-made content can trick financial systems and security checks. Deepfake and deception risks have attracted regulatory attention globally, including calls for standards to detect manipulated content and protect the integrity of financial and identity verification systems. We have significantly advanced the theoretical understanding of fraud in the context of AI. The AI-Fraud Diamond extends the traditional Fraud Triangle by adding technical opacity as a fourth condition, alongside pressure, opportunity, and rationalization [3]. This addition was confirmed by talking to four auditors from two of the Big Four consulting firms, who agreed that technical opacity is becoming more important in auditing but also pointed out that many financial auditors still do not fully understand the IT and AI systems needed to evaluate current fraud risks. The AI-Fraud Diamond creates a system to categorize AI-related fraud into five types: changing input data, misusing models, manipulating algorithm decisions, creating fake information, and fraud based on ethics, giving a clear way to understand how fraud occurs in systems that use algorithms.

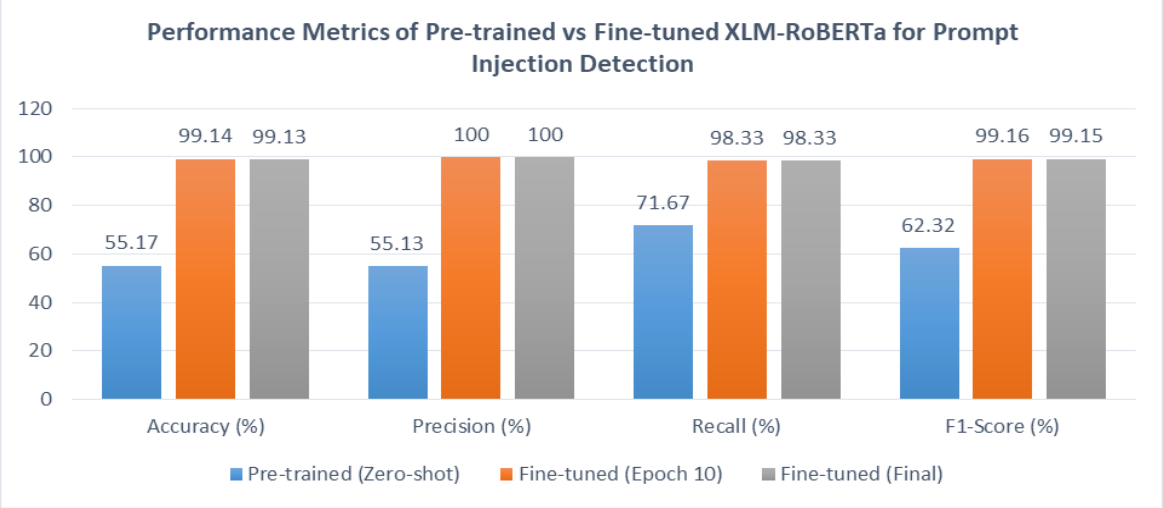

Fig. 1: Performance Metrics of Pre-trained vs Fine-tuned XLM-RoBERTa for Prompt Injection Detection [10]

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.2 Industry Perspectives and Detection Advances

Professional risk advisors and industry bodies emphasize that without rigorous governance, GenAI outputs can lead to inaccuracies, compliance violations, and legal exposure that may have far-reaching consequences for organizations. Internal audit and risk teams need to create new ways to assess AI systems based on how they are used in real situations, instead of using old risk assessment methods that were meant for straightforward, rule-based systems. The scale of AI-enabled threats is underscored by the fact that AI-enhanced phishing attacks surged by 600% during the COVID-19 pandemic, exploiting increased digital communication channels and demonstrating how AI-generated content can deceive targets at an unprecedented scale [3]. The increasing complexity of fake content created by AI has led to studies on financial fraud detection systems that use generative AI technology to spot tricks and unusual results in financial systems. These detection systems are an important tool in the battle against AI-driven fraud, but how well they work really relies on the quality of the rules and management around their use and how well organizations include them in their overall control systems.

3. Prompt Fraud: Definition, Characteristics, and Threat Landscape

3.1 Defining Prompt Fraud and Its Core Characteristics

Prompt fraud is when someone intentionally creates specific inputs, called prompts, for GenAI systems to trick them into generating fake results that can fool internal checks or auditors. It involves linguistic manipulation rather than direct system intrusion, making it subtle, quick, and potentially scalable across multiple organizational processes and functions. The core characteristics of prompt fraud distinguish it from conventional fraud typologies in several important ways. First, it relies on linguistic manipulation, whereby fraudsters create prompts that instruct AI to write plausible but false documentation, narratives, or evidence. Second, it pretends to have authority, making the results look professional and trustworthy by using the advanced language skills of modern LLMs. Third, it can get around controls without needing access to the system or stealing credentials, working only through the natural language interface that GenAI systems provide to users. The main new risk is that audit and internal control systems often accept AI outputs as trustworthy evidence without checking where they came from or understanding the purpose of the prompts that created them. Research into making LLM trust boundaries stronger has looked at different ways harmful inputs can trick the system, showing how these attacks can make AI outputs less reliable and challenge the basic security beliefs that organizations depend on when using these systems. The seriousness of this problem shows that prompt injection means changing the inputs by mixing harmful data with a stable prompt created by developers, which can completely or partially alter the output of the LLM. Research shows that using datasets with 546 training examples and 116 test examples, each having two parts that represent the prompt text and a label that marks it as harmful or safe, is enough to create effective detection models.

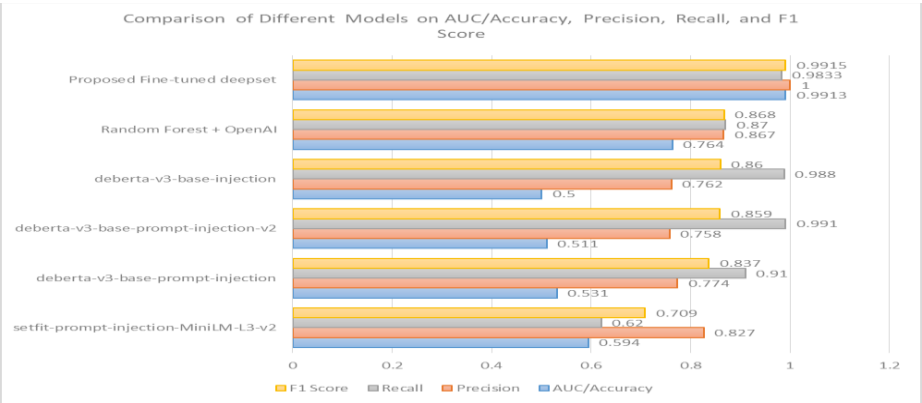

Fig. 2: Comparison of Different Models on AUC/Accuracy, Precision, Recall, and F1 Score [10]

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.2 Threat Actors, Use Cases, and Fraud Scenarios

The threat landscape for prompt fraud encompasses both internal and external actors who may exploit GenAI systems for different purposes and through different mechanisms. Internal actors, like employees or contractors with special access to AI tools might create prompts to make invoices just under the approval limit, write vendor change requests that sound like they come from a manager, or produce explanations that support unauthorized entries in financial systems. External actors, like third-party vendors or enemies, might send fake documents as part of vendor communications or clever phishing scams, using AI-generated deepfake emails or fake regulatory filings to trick finance or audit teams. Additionally, methods for categorizing prompt injection attacks have shown the wide range of techniques that attackers can use, from directly overriding instructions to changing context and sneaking in harmful content, all of which can be used for fraud in business settings. Indirect prompt injection attacks are especially dangerous because they take advantage of how LLMs are built. LLMs process external inputs and instructions without separating them, which lets attackers add harmful code to these inputs and causes LLMs to produce harmful or misleading outputs. Research has found that even models like XLM-RoBERTa, which don't require extra training, only achieved 55.17% accuracy in identifying these attacks without prior examples, showing how easily these changes can bypass basic detection. In financial reporting, a crafted prompt might instruct AI to produce narratives supporting aggressive accounting. In procurement and vendor management, Prompt Fraud can be used during the buying process to create fake invoices or approval emails that look like they follow company rules, which could result in unauthorized payments if there are weak checks in place, like fake job positions or made-up reconciliations.

4. Control Weaknesses and Risk Analysis

4.1 Gaps in Organizational AI Governance

Organizations currently lack widespread processes for validating the input and output quality of GenAI systems, creating significant vulnerabilities that prompt fraud can exploit. Common weaknesses include not keeping records of prompts that would allow us to look back and see how AI outputs were created, using AI outputs as proof without checking their sources for accuracy, and relying too much on AI-generated summaries without enough human checks to catch any false information or misleading details. Without governance frameworks that establish safeguards for both AI inputs and outputs, GenAI becomes a potential channel for fraud rather than a neutral tool that enhances organizational capabilities. The rise of Shadow AI, which means employees using AI tools without permission from IT or compliance, makes these problems much worse because it can lead to fraud by ignoring internal rules and putting organizations at risk of data breaches, GDPR violations, and not following laws like the EU AI Act. The AI-Fraud Diamond research, which included interviews with four experts on five key topics, such as fraud in complex systems and the flaws of the Fraud Triangle, revealed that these unregulated tools can cause technical issues like prompt injections and training data poisoning, damaging data security and system reliability, and that fraud is now more about tampering with IT systems than just altering numbers.

Fraud Category	Underlying Mechanism	Manifestations
Input Data Manipulation	Data manipulated before or during AI model training to influence output	Data Poisoning, Synthetic Data Generation
Model Exploitation and Evasion	Attackers directly manipulate, steal, or bypass the AI model	Model Stealing, Adversarial Attacks, Evasion Attacks
Algorithmic Decision Manipulation	AI-driven decisions manipulated through bias or systematic influence	Bias Exploitation, Automated Redlining

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Synthetic Misinformation and Deception	AI used to generate misleading content that spreads false information	Deepfakes, Bot-Generated Content, AI-generated spam, and Phishing
Unregulated AI and Ethics Fraud	AI deployed without oversight or regulation beyond regulatory controls	Shadow AI, AI Ethics Washing (Fake Compliance)

Table 1: AI-Fraud Taxonomy Categories and Manifestations [3]

4.2 Audit Readiness and Competency Gaps

Beyond governance gaps, the readiness of internal audit functions to detect and respond to prompt fraud remains a significant concern across industries and organizational sizes. Practical insights from internal auditors involved in AI assurance activities show that many audit functions are still just starting to build the skills, tools, and methods needed to effectively audit AI systems. There are still big problems with finding risks, being ready to adopt new methods, and ensuring proper coverage, which means that even if there are controls in place, how well they work can This readiness gap is particularly concerning because prompt fraud, by its nature, requires auditors to possess skills that extend beyond traditional financial and operational auditing into areas such as natural language processing, AI system architecture, and prompt engineering. Research shows that a pre-trained XLM-RoBERTa model, used for detecting prompt injection without any adjustments, only achieved 55.17% accuracy, 55.13% precision, 71.67%. This means that without specific training and changes, even advanced language models struggle to reliably find harmful prompts. If we don't work on improving these abilities, audit functions might miss detecting prompt fraud, even in the areas they monitor, which could leave them open to being taken advantage of by skilled attackers.

5. Proposed Audit Methodology and Control Framework

5.1 Risk Assessment and Audit Planning

An effective audit method for dealing with prompt fraud should start with a thorough risk assessment that finds areas where GenAI creates controlled documents, writes approval narratives, or helps with decisions in financial reporting and other risky processes. This risk assessment should outline how the organization uses GenAI in all areas, pinpoint where AI-generated information influences decisions or controls, and assess the possible effects of any altered outputs at those points. Many people agree that using GenAI to make audits more efficient is important, but planning for this task must also include finding and reducing the fraud risks that come with the tools being used to do so [9]. Audit planning should include specific steps to verify that AI-generated evidence is trustworthy, like developing testing methods that focus on the unique aspects of prompt fraud and its language tricks. Studies indicate that enhancing these methods can be done rapidly within the first 10 training cycles, with a well-adjusted XLM-RoBERTa model reaching its highest performance of 99.14% accuracy and a 99.16% F1 score by the 10th cycle, showing that effective tools for detecting audits can be developed using specific datasets

Control Category	Control Type	Recommended Measures
Governance	Preventive	AI use policies, Prompt input standards, Role-based permissions
Technical	Preventive	Prompt firewalls, Output watermarking, Retrieval-augmented generation with evidence tracking
Detective	Detective	NLP-based anomaly detection, Audit trails for AI outputs, Manual sign-off on high-risk outputs
Monitoring	Continuous	Real-time prompt logging, Automated output verification, Cross-referencing with source data

Table 2: Proposed Control Framework for Prompt Fraud Mitigation [9, 10]

Journal of Information Systems Engineering and Management
2026, 11(2s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.2 Testing Procedures, Controls, and Regulatory Considerations

Testing procedures for Prompt Fraud should include three important areas that together ensure the reliability of AI-generated results. Prompt validation means that auditors need to check the prompts used to create important documents to make sure they came from approved sources and followed the organization's rules and standards. Output verification means checking AI-generated content against data from other sources and noting any differences between the outputs and known system facts or supporting evidence. Logging and traceability mean keeping detailed records of the input prompts, context information, and model versions to ensure that audits can verify both the language and data accuracy, making sure that outputs can be traced back to approved inputs and were not intended to mislead reviewers or decision-makers. The suggested control framework includes rules for managing AI use, such as policies, standards for input prompts, and permissions based on roles, as well as technical measures like barriers for prompts, marking outputs, and tracking evidence in generated content. Detective controls should include NLP-based anomaly detection, comprehensive audit trails for AI outputs, and mandatory manual sign-off on high-risk outputs. Using specially trained LLMs as detective controls has been shown to work well, with a fine-tuned XLM-RoBERTa model achieving 99.13% accuracy, 100% precision, 98.33% This model was trained on 546 examples and tested on 116 samples over 50 training cycles, performing much better than its non-fine-tuned version and previous methods like Multilingual BERT, which only achieved 96.55% accuracy on the same. Regulators are increasingly This study scrutinizes the impact of AI on audit quality and risk, with reviews of auditing firms revealing a lack of formal assessment metrics for evaluating how AI tools affect audit outcomes. This issue signals a pressing need for stronger regulatory oversight and the development of industry-wide standards for AI governance in audit contexts.

Conclusion

Prompt Fraud changes the way we think about fraud risk, shifting the focus from exploiting systems to taking advantage of how AI understands and uses language. This change fundamentally questions the basic ideas behind current risk frameworks, which were created for a time when fraud needed direct access to systems, stealing credentials, or altering physical or digital records. As GenAI systems become more involved in how organizations operate, the risk of prompt fraud affecting the accuracy of financial reports, purchasing, compliance, and other important functions will increase in both size and complexity. Today's risk management strategies need to change to include strong rules for how AI data is handled, ensure that people check important decisions, and use special audit methods to spot language tricks and fake information created by fraudulent prompts. Organizations that set up strong rules for using AI, have people check important decisions, and use systems to monitor things in real-time will be in a better position to handle the new risks of GenAI and keep their internal controls reliable. Without these adaptations, AI may inadvertently become a conduit for fraud, undermining both internal controls and stakeholder trust. Moving ahead will need teamwork between auditors, risk experts, tech specialists, regulators, and company leaders to create and put in place the rules, safety measures, and skills needed to tackle this new type of fraud risk effectively and in advance.

References

[1] Maria-Alessia Feleaga, "The Impact of Generative Artificial Intelligence on Internal Auditing: A Conceptual Literature-Based Analysis," in European Journal of Accounting, Finance & Business, June 2025. Available: https://www.researchgate.net/publication/400263743_THE_IMPACT_OF_GENERATIVE_ARTIFICIAL_INTELLIGENCE_ON_INTERNAL_AUDITING_A_CONCEPTUAL_LITERATURE-BASED_ANALYSIS

[2] The Institute of Internal Auditors (IIA), "Artificial Intelligence Auditing Framework," in IIA Global Knowledge Brief, 2025. Available:

https://www.theiia.org/globalassets/site/content/tools/professional/aiframework-sept-2024-update.pdf
[3] Benjamin Zweers, et al., "The AI-Fraud Diamond: A Novel Lens for Auditing Algorithmic Deception," in arXiv (Cornell University), 19 August 2025. Available: https://arxiv.org/abs/2508.13984
[4] Sagarika Behera, et al., "Generative AI-Based Financial Fraud Detection System," in ResearchGate / International Journal of Intelligent Systems, January 2025. Available: https://www.researchgate.net/publication/389869974_Generative_AI-Based_Financial_Fraud_Detection_System
[5] Surender Suresh Kumar, et al., "Strengthening LLM Trust Boundaries: A Survey of Prompt Injection Attacks," in IEEE International Conference on Human-Machine Systems (ICHMS), 19 June 2024. Available: https://ieeexplore.ieee.org/abstract/document/10555871
[6] B. Schneier, "A Taxonomy of Prompt Injection Attacks," in Schneier on Security / arXiv, March 2024. Available: https://www.schneier.com/blog/archives/2024/03/a-taxonomy-of-prompt-injection-attacks.html
[7] Cimcon / Compliance Week, "Shadow AI: What is it and How to Manage the Risk from it?," in Compliance Week/SlideShare Whitepapers, 2025. Available: https://www.slideshare.net/slideshow/shadow-ai-what-is-it-and-how-to-manage-the-risk-from-it/272013995
[8] Doong Yee Jiun, et al., "Auditing Artificial Intelligence in Practice: Insights from Internal Auditors on Risk, Adoption, and Assurance," in ResearchGate, December 2025. Available: https://www.researchgate.net/publication/399039379_Auditing_Artificial_Intelligence_in_Practice_Insights_from_Internal_Auditors_on_Risk_Adoption_and_Assurance
[9] Chelson Chong, et al., "Harnessing GenAI to Improve Audit Work Efficiency Through Proper Planning," in ISACA Industry News, 20 May 2025. Available: https://www.isaca.org/resources/news-and-trends/industry-news/2025/harnessing-genai-to-improve-audit-work-efficiency-through-proper-planning
[10] Md Abdur Rahman, et al., "Fine-tuned Large Language Models (LLMs): Improved Prompt Injection Attacks Detection," in 2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), July 2025. Available: https://www.researchgate.net/publication/394981773_Fine-tuned_Large_Language_Models_LLMs_Improved_Prompt_Injection_Attacks_Detection