IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13801

Prompt Security and Insider-Risk Defense in Banking AI Systems

Syed Sharik Ali

Department of Information Technology, Webster University, MO, USA

Abstract: Artificial Intelligence (AI) assistants are being utilized in several banking systems to improve their efficiency, customer service capability, fraud detection and prevention capability, compliance monitoring system, and decision-making process. However, with the increasing usage of LLMs and intelligent assistants comes the emergence of severe cybersecurity issues including the prompt injection attacks, insider threat issues, access issues, and sensitive data exposure. Prompt injection refers to taking control of the AI system by injecting prompts and hidden instructions that enable security threats to bypass existing measures and reveal sensitive financial information or make the AI system do what it is not supposed to be doing. In the case of banks where AI systems are linked to customer records, transactions, and authorization processes, prompt injection and insider threat issues will lead to financial fraud and compliance problems. This paper highlights and mitigates the two most serious types of security risk associated with prompts and insider issues to the bank AI systems by developing a layered security strategy for prompt isolation and insider threat mitigation through role-based access control, human validation criteria, and restricted scope of impact zones over the duration of alert monitoring period. Simulation studies prove that our methodology can reduce the success rate of attacks and provide financial institutions with safe AI governance.

Keywords: Prompt Injection, Prompt Security, AI Banking Solutions, Internal Threats Detection, Information Protection from Leakage, LLM, Banking AI Assistants, Financial Cybersecurity, RBAC, Human-in-the-Loop Security, AI Governance, AI-based Banking Automation, Compliance Monitoring, Prevention of Fraud, LLM Security

I. INTRODUCTION

AI Assistants have been rapidly deployed in the banking industry to facilitate better services, improve internal workflow, enhance fraud detection mechanisms, automate loan processing and compliance monitoring activities among other applications. Large Language Models (LLMs) have been adopted by the banking industry for use in decision support, report writing, transactions analysis and employee help [1]. Despite the ability of these tools to increase efficiency and reduce operating costs, there is also potential for major cybersecurity risks arising. Some of the most pressing risks include prompt injection. This involves the use of concrete instructions embedded within user prompts, emails, documents or any interaction with the underlying platform that manipulates the AI and effectively bypasses all defences put in place. Transactions, personal details of customers and approvals can be made fraudulent. The use by employees of AI Assistants with access to sensitive platforms of the banking environment, whether intentionally or inadvertently, creates a recipe for insider attacks. Classical prompt engineering alone is inadequate to secure these systems [2]. It follows therefore that robust security measures have to be put in place such as access control, policy enforcement and human oversight.

II. LITERATURE REVIEW

A. Prompt Injection in Financial AI

Prompt injection represents one of the most serious vulnerabilities of bank systems relying on LLM technology, where the adversarial input overrides prompts and credentials of the systems to manipulate the outputs [3]. As far as the business of the research goes, one should expect that the attackers will provide specific commands via emails, PDFs, or chat messages with the aim of conducting some kind of attacks, such as disclosure of sensitive information or passing certain compliance tests.

B. Risk of engaging with AI assistants

It should be also noted that the threat of insiders may remain very serious since employees who are authorized to do certain things could inadvertently disclose confidential information due to the use of AI systems. According to new research, AI systems increase risks associated with insiders through automating tasks and accessing large amounts of data [4].

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13801

C. Risks of Data Leakage

Our model is prone to data leaks as the information may get leaked due to their exposure to context and training data leaks or the tools used by the model. The organizations will suffer huge punishments for such kind of data leaks [5].

D. Security Controls in Place

There have been some control mechanisms suggested already such as RBAC, prompt filtering, sandboxing and auditing logs [6]. But literature suggests that just as promptlevel controls fail even when all forcing is done at system level against the attacks, our mechanism too fails.

Table 1: Literature Review Summary

Author	Focus Area	Key Finding
Rehberger (2024)	Prompt Injection	Hidden prompts can override AI safeguards
Olzak (2026)	AI Security	System-level controls outperform prompt defenses
CIS (2026)	Data Leakage	LLMs risk exposing sensitive financial data
Industry Reports	Insider Threats	AI increases insider misuse risk

III. PROBLEM AND MOTIVATION

C. Core Problem

AI assistants used in banking work in highly sensitive ecosystems, where customer data, financial transactions, reports, approval mechanisms, and the process itself are linked with each other. This is especially dangerous in cases where prompts that have sensitive data and are being executed belong to the same trust boundary. Attackers take advantage of this weakness and introduce malicious prompts into the user inputs through various means such as e-mail messages, documents, or other types of workflows. As a result, these subtle prompts make AI assistants violate security protocols, reveal financial information, and commit other unwanted activities [7]. Banking involves high levels of integrity and precision — a small error can lead to massive consequences.

B. Reasons for Research

The motivation for conducting this research is due to the increased adoption of artificial intelligence-based systems in banks to support decision-making and automate processes [8]. Although such systems can assist in accelerating processes, ensuring proper customer service, assisting in fraud detection and acting as a compliance tool, they create new challenges with regard to cyber security at the application level. Insider threat: An employee with high-level authorization rights using an AI-based system against bank interests, knowingly or unknowingly (knowingly is more damaging) [9]. It is not possible to use conventional prompts here to conduct this kind of 'play', as the threat does not just target model responses, but also system authorizations.

C. Gap in Research

Almost all studies available today are oriented towards prompt engineering and filtering of inputs to prevent prompt injection attacks. There has been less research done on execution-layer security, as well as privilege control on execution layers and insider risk management for banking AI applications [10]. Almost all existing approaches have a lack of any human approval processes, limited user behaviour analysis, and zero trust environment related to financial transactions.

TABLE 2: BANKING AI THREAT IMPACT ANALYSIS

Threat Type	Risk Level	Operational Impact	Business Consequence
Prompt Injection	Very High	AI manipulation	Fraudulent transactions
Insider Misuse	Critical	Unauthorized access	Compliance violations

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13801

Data Leakage	High	Confidential data exposure	Privacy breach
Tool Permission Abuse	Critical	Unauthorized execution	Financial loss
Weak Audit Controls	Medium	Delayed detection	Regulatory penalties

Figure 1: Prompt Security and Insider-Risk Defence in Banking AI Systems

IV. RELATED WORK

The prompt injection vulnerability is a growing cyber risk which has developed from being chatbots (in>tip of the iceberg and not so widely applicable) to a much more serious and system-level threat, and most probably a significant cyber-security risk in the near future in financial services [11]. The reason behind this new finding is supported by research. Prompt injection means that malicious prompts will be included in emails, documents, websites or customer requests, which makes an AI agent carry out tasks in a hypnotized mode, and ignoring security prompts – thus going beyond humanly-induced artificial instructions. This makes them vulnerable to privileges escalation, financial crimes, leakage of sensitive information, and non-compliance. In certain literature, there is reference to "prompt ware" which means that prompt injection attacks are made up of multi-step attacks which include persistence, lateral movement, and cross-system attacks. The past recommendations include prompt separation from executables, restrictive role-based access control for software, limited permissions to software use, and human approval when dealing with important actions. However, what is needed is a powerful defence against this threat since banking systems employ AI assistants in authorizing transactions, checking financial statements, and accessing customers' accounts [12].

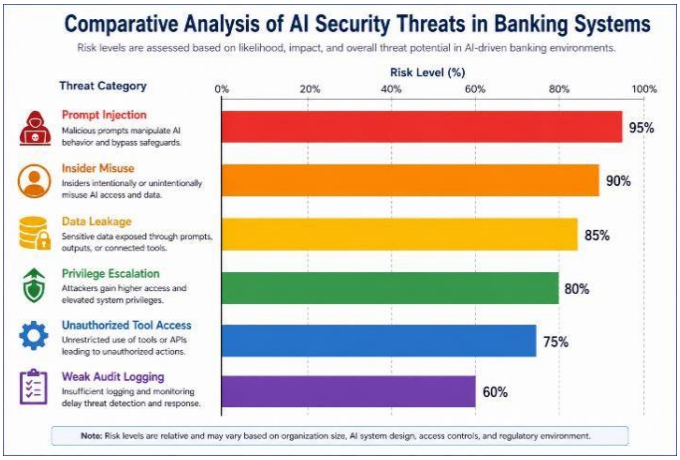

Figure 2: Comparative Analysis of AI Security Threats in Banking Systems

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13801

V. METHODOLOGY / PROPOSED FRAMEWORK

A. Proposed Secure Banking AI Framework

The presented model acts as a defence-in-depth measure aimed at protecting AI banking assistants from prompt injections, abuse by insiders and unauthorized access to data. First of all, the framework guarantees the entire process of AI executions, not just prompt engineering. The proposed security solution involves immediate isolation, hard access control, policy validation, human-in-loop and monitoring of AI in action [13]. All individual security components operate independently but complement each other in the case of total protection from potential threats. Therefore, this approach helps minimize attack vectors, making AI-based banking systems such as transaction authorizations, compliance checks, fraud detection, and chatbots more reliable.

B. Components of the Security Framework

1. Prompt Isolation Layer

This layer separates user inputs from system commands. External information is recognized as potentially harmful content that may compromise system policies when introduced through emails, documents or customer requests [14].

2. Role-Based Access Control (RBAC)

With the help of RBAC, such restrictions are implemented in a way that allows AI assistants to communicate with the data and tools provided by user role permissions [15]. Thus, there will be no risk that the insider will misuse his or her access privileges and execute any bank transactions.

3. Human-in-the-Loop Verification

Transactions that involve high risks such as transferring funds, alterations in accounts, and compliance approvals need to be verified by a human [16]. This minimizes the risk of unauthorized transactions or fraudulent activity.

4. Audit Logging and Monitoring

The audit logging and monitoring layer logs the input and output given to the model by a human in real-time and the way it handles its resources or any behavioural issues [17]. The continuous monitoring process helps in identifying potential threats, conducting forensic investigations, and maintaining compliance standards.

5. Policy Enforcement Layer

The server-side policy controls verify the decision made by the AI model before execution [18]. Nonetheless, the policy gates prevent unauthorized activities from being performed through prompt injection attacks.

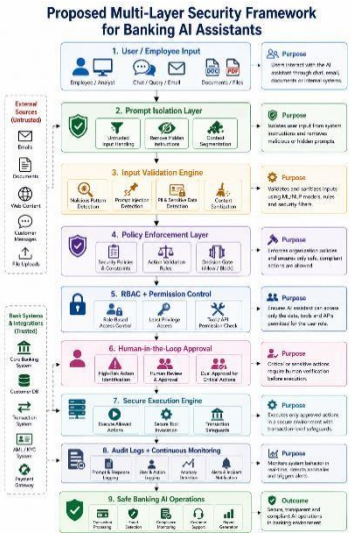

Figure 3: Proposed Multi-Layer Security Framework for Banking AI Assistants

VI. EXPERIMENT / CASE STUDY / SIMULATION

A. Simulation Setup

Prompt used in the simulation of an AI banking assistant aimed at evaluating prompt-based security threats and insider threat prevention techniques in the application. The simulated prompt was connected with FIR request processing, customer account management, compliance checking tools and reporting functions [19]. Both normal and malicious prompt use cases were considered to evaluate the performance of the AI within the safe environment and a hostile environment. The following attack scenarios were implemented: prompt injection via user input, embedded commands in the file or mail, and insider abuse due to privilege rights.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 8, August 2026

DOI: 10.17148/IARJSET.2026.13801

B. Attack scenarios

Three types of attack scenarios were chosen for analysis: supply through prompt injection, prompt injection after the upload of file or mail and insider abuse (unselective data access) [20]. In each of these attack scenarios, the aim was to assess the capability of the system to recognize or block potentially dangerous activity or allow the action to occur but control its process.

C. Evaluation metrics

Main metrics to consider included the percentage of successful attacks, the number of approvals for transaction initiation, insider abuse-related sensitive data leaks, detection accuracy rate for insider abuse and validation period [21].

D. Discussion of Results

Results revealed that there were high success rates for attacks as well as low chances for detecting an insider threat with regard to any system that used prompt filtering alone. The developed framework involves a combination of RBAC and policy enforcement with approval by the human user [22].

VII. TABLE 3: EXPERIMENTAL RESULTS OF SECURITY MODELS

Security Model	Attack Success Rate	False Approvals	Data Leakage Risk	Insider Abuse Detection
Basic Prompt Filtering	61%	High	High	Low
RBAC Only	38%	Medium	Medium	Medium
Proposed Multi-Layer Framework	8%	Very Low	Very Low	High

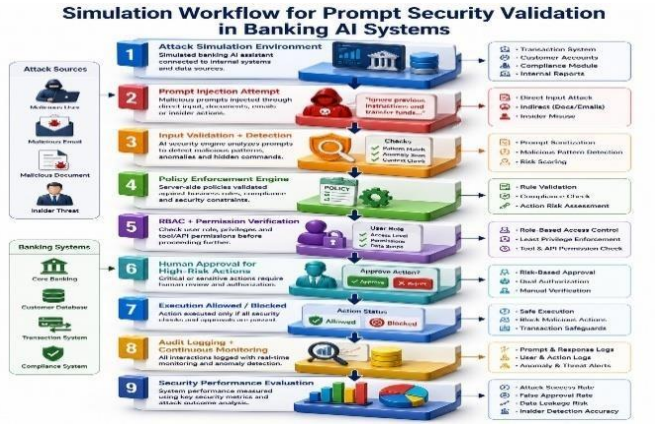

Figure 4: Simulation Workflow for Prompt Security Validation in Banking AI Systems

VIII. CONCLUSION

With the integration of AI assistants into the bank's system, one may be able to leverage assistance from Cognitive Systems (adaptive systems) in order to gain many benefits that range from efficiency, fraud detection, compliance monitoring, customer services, as well as decision support. However, besides these benefits that can be reaped from such an AI integration, the most crucial cyber security benefits would be prompt injection attacks, insider threat exploitation of technology, technology tool abuse as well as data leaks. With the data training that ends in October 2023, how are you going to safeguard against any adversary whose main function is to exploit your model responses, execution paths, permissions, and system integration? According to the study conducted, prompt security will need a defence-in-depth approach which includes prompt isolation, role-based access control implementation, vulnerability-aware policy enforcement, human in the loop validation as well as monitoring through audit logging. Experimental evidence proved that the designed stack framework resulted in great reductions on attack success rates as well as insider threat detection improvements. In case banks wish to adopt monitors in AI for profits, it will need governance, zero trust approach and cyber security validation capability.

DOI: 10.17148/IARJSET.2026.13801

REFERENCES

[1] M. F. Ansari, A. Mohammed, K. R. Janamolla, S. W. Khadri, S. A. Khader and M. A. Raheem, "The Double-Edged Sword: Navigating AI's Role in Banking Cybersecurity," 2025 International Conference on Transformative Computing Technologies (ICTCT), Bangalore, India, 2025, pp. 294-300, doi: 10.1109/ICTCT69201.2025.00062

[2] Marvin, Ggaliwango, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. "Prompt engineering in large language models." In *International conference on data intelligence and cognitive informatics*, pp. 387-402. Singapore: Springer Nature Singapore, 2023.

[3] McHugh, Jeremy, Kristina Šekrst, and Jon Cefalu. "Prompt injection 2.0: hybrid AI threats." *arXiv preprint arXiv:2507.13169* (2025).

[4] RAHEEM, MOHD ABDUL, and MOHAMMED AZMATH ANSARI. "INTELLIGENT AND TRUSTWORTHY 6G: AIDRIVEN ARCHITECTURES, APPLICATIONS, AND SECURITY FRAMEWORKS..

[5] Wu, Xiaolan, Juan Li, and Wenkun Ren. "Risk Assessment Framework for Data Leakage Prevention Using Machine Learning Techniques." *Artificial Intelligence and Machine Learning Review* 5, no. 3 (2024): 55-66.

[6] Rao, K. Rajesh, Ashalatha Nayak, Indranil Ghosh Ray, Yogachandran Rahulamathavan, and Muttukrishnan Rajarajan. "Role recommenderRBAC: Optimizing user-role assignments in RBAC." *Computer Communications* 166 (2021): 140-153.

[7] Kafi, Md Abdullahel, and Nazma Akter. "Securing financial information in the digital realm: case studies in cybersecurity for accounting data protection." *American Journal of Trade and Policy* 10, no. 1 (2023): 15-26.

[8] Sultana, Ghousia, Siraj Farheen Ansari, Mohammed Imran Ahmed, Abdul Faiyaz Shaik, Moin Uddin Khaja, and Bibhu Dash. "RESPONSIBLE AI ANALYTICS FOR REAL-WORLD IMPACT: NAVIGATING ETHICS, PRIVACY AND TRUST."

[9] Sawant, Krutika, Harshvardhan Soni, Parthraj Maharaul, and Saurabh Agarwal. "A study of AI in banking system." *Korea Review of International Studies* 16, no. 6 (2023): 36.

[10] Khader, Shuaib Abdul, Amir Ahmed Ansari, and Syed Sharik Ali. "Zero-Day Exploit Prediction Using Graph-Based Deep Learning on Vulnerability and Threat Intelligence Data."

[11] Reddy, B. (2021). A Quantitative Analysis of Cloud Security Practices in IoT Environment (dissertation).

[12] Janamolla, Kavitha, Ghousia Sultana Sultana, Fnu Mohammed Aasimuddin, Abdul Faisal Mohammed, and Fnu Shaik Aqheel Pasha Pasha. "Integrating Blockchain and AI for Efficient Trade Exception Handling: A Case Study in Cross-Border Settlements."Journal of Cognitive Computing and Cybernetic Innovations 1, no. 1 (2025): 24-30.

[13] Ansari, Mohammed Azmath, Shaik Aqheel Pasha, and Narendar Kandula. "Reinforcement Learning for Adaptive Network Defense in Financial Institutions."

[14] Kashif, M., & Ansari, A. A. (2026). Building a unified AI-driven analytics pipeline for real-time anomaly detection in high-velocity data streams. IJIREEICE, 14(1), 66–75. https://doi.org/10.17148/ijireeice.2026.14111

[15] Mohammed, Naveed Uddin, Zubair Ahmed Mohammed, Shravan Kumar Reddy Gunda, Akheel Mohammed, and Moin Uddin Khaja. "Networking with AI: Optimizing Network Planning, Management, and Security through the medium of Artificial Intelligence.

[16] Mohammed, Akheel, Zubair Ahmed Mohammed, Naveed Uddin Mohammed, Shravan Kumar Gunda, Mohammed Azmath Ansari, and Mohd Abdul Raheem. "AI-NATIVE WIRELESS NETWORKS: TRANSFORMING CONNECTIVITY, EFFICIENCY, AND AUTONOMY FOR 5G/6G AND BEYOND

[17] Khaja, Moin Uddin, and Balavardhan Reddy. "Securing Agentic AI in Software-Defined Networks: A Policy-Driven Framework for Governance, Monitoring, and Incident."

[18] Ansari, Meraj Farheen, and Syed Sharik Ali. "AI-driven zero-trust architecture for enhanced cybersecurity in dynamic network environments." International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering 13 (2025): 12.

[19] Chittoju, S. R., and Siraj Farheen Ansari. "Blockchain's Evolution in Financial Services: Enhancing Security, Transparency, and Operational Efficiency." International Journal of Advanced Research in Computer and Communication Engineering 13, no. 12 (2024): 1-5.

[20] Sharma, Nipun, Swati Sharma, and Anupama Sindgi. "Solidity Smart Contract Vulnerabilities, Attack Scenarios, and Mitigation—A Survey." In *International Conference on Communication and Computational Technologies*, pp. 901-910. Singapore: Springer Nature Singapore, 2023.

[21] Kashif, Mohammed, Mohammed Aasimuddin, Mubashir Ali Ahmed, Laxmi Bhavani Cheekatimalla, Eraj Farheen Ansari, and Ahwan Mishra. "AI-DRIVEN CTI FOR BUSINESS: EMERGING THREATS, ATTACK STRATEGIES, AND DEFENSIVE MEASURES."

[22] Reddy, Balavardhan, and Amir Ahmed Ansari. "AI-ENHANCED NETWORK TRAFFIC ANALYSIS FOR PREVENTING FRAUD PAYMENT IN BANKS."

