Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 04 April-2026, Page No.-9494-9514

DOI: 10.47191/etj/v11i04.21, I.F. – 8.595

Mandatory Al Red Teaming in Cloud-Based Fintech Platforms: Implications for Regulatory Compliance and Fraud Prevention

Temitope Ibrahim Lawal[1], Ololade Zainab Adesokan[2], Onyinye Agatha Obioha-Val[3], Damilola Abidemi Akinwunmi[4], Omobolaji Olufunmilayo Olateju[5]

[1]Fintech and Digital Technology Researcher, Pace University, 78 N Broadway, White Plains, NY 10603, United States of America https://orcid.org/0009-0000-8785-839X

[2]Digital Marketing and E-commerce Researcher, Institution and Address American National University, Salem VA 1813 E Main St, Salem, VA 24153, United States https://orcid.org/0009-0006-6888-0552

[3]Information Technology Researcher, University of the Cumberlands, 104 Maple Drive, Williamsburg, KY 40769, United States of America https://orcid.org/0009-0001-6021-6124

[4]Data Analytics and Financial Systems Risk Researcher, Glasgow Caledonian University, Cowcaddens Road, Glasgow, G4 0BA, Scotland, United Kingdom https://orcid.org/0009-0000-0248-1264

[5]Agricultural Technology Researcher, University of Ibadan, Oduduwa Road, Ibadan, Oyo State, Nigeria https://orcid.org/0009-0003-6470-1025

ABSTRACT: This study examines the role of mandatory artificial intelligence red teaming in strengthening the security and regulatory compliance of cloud-based fintech platforms. A quantitative research design was employed using four open-access datasets: the IEEE Credit Card Fraud dataset, the IBM Credit Card Fraud dataset, the BankSim dataset, and the MITRE ATT&CK Enterprise dataset. The analysis integrated logistic regression vulnerability modelling, adversarial perturbation impact analysis, structural equation modelling, and a quantitative cybersecurity risk scoring model. The results indicate that key variables such as V14 significantly increase fraud probability ($\beta = 2.031$; odds ratio = 7.62), while adversarial manipulation reduced fraud detection accuracy from 0.986 to 0.863 before improving to 0.948 after red teaming mitigation. Governance oversight ($\beta = 0.56$, $p < 0.001$) also significantly strengthened fraud resilience. Based on these findings, the study develops and empirically operationalizes an AI Red Teaming Governance Framework consisting of four integrated components: vulnerability identification, adversarial red teaming, governance oversight, and regulatory compliance integration. The framework provides a structured approach for implementing and monitoring AI security testing within cloud-based fintech platforms. The study recommends mandatory AI red teaming regulations, continuous adversarial testing, standardized governance frameworks, and prioritization of high-risk attack vectors in fintech cybersecurity strategies.

KEYWORDS: Artificial Intelligence Red Teaming, Fintech Cybersecurity, Adversarial Machine Learning, Fraud Detection Systems, AI Governance and Regulatory Compliance.

1. INTRODUCTION

The rapid digital transformation of financial services has significantly reshaped the global financial ecosystem. Financial technology platforms now support a wide range of services including mobile banking, digital payments, peer to peer lending, automated investment management, and algorithmic credit assessment (Bhattacharjee et al., 2024). This transformation has been driven largely by advances in artificial intelligence and cloud computing, which enable financial institutions to process massive volumes of transactional data and automate financial decision making at unprecedented speed. Artificial intelligence models are increasingly used to detect fraud, evaluate credit risk, monitor transactions, and authenticate digital identities (Goyal et al., 2025; Kothinti, 2025). Recent studies indicate that the integration of artificial intelligence into financial services has significantly improved operational efficiency, predictive analytics, and financial risk monitoring capabilities within digital financial ecosystems (Aliyu & Iheonkhan, 2025; Vuković et al., 2025). As financial services continue to migrate toward digital platforms, artificial intelligence has become a central component of contemporary fintech infrastructures. These developments have been further supported by cloud based infrastructures that allow fintech companies to deploy scalable services capable of processing millions of transactions across geographically distributed networks. Such infrastructures rely on interconnected application programming interfaces, machine learning pipelines, and automated decision engines that operate continuously in real time (Kumar, 2025). While these systems

improve efficiency and expand access to financial services, they also create increasingly complex digital ecosystems in which vulnerabilities within one component can propagate across multiple systems, thereby increasing exposure to cyber exploitation and operational disruption (Chuang & Shrestha, 2025; Ogunmolu et al., 2025).

The growing reliance on digital financial infrastructures has coincided with a significant rise in cybercrime and financial fraud targeting the financial sector. Global assessments indicate that consumers lost approximately $442 billion to financial scams and fraud in 2024, demonstrating the scale of financial exploitation occurring within digital economies (Harris, 2025). At the same time, the financial sector has become one of the most targeted industries for cyberattacks, accounting for nearly 27% of reported global data breaches, while the average cost of a financial sector breach is estimated at about $5.9 million per incident (Bonderud, 2024; Elgan, 2024). As financial services become more dependent on automated digital platforms, ensuring the security and resilience of these infrastructures has become a critical priority for financial institutions and regulatory authorities.

Artificial intelligence has emerged as a key technological tool for combating these threats by enabling financial institutions to detect anomalous transaction patterns and identify fraudulent behaviour in real time. Machine learning algorithms can analyse vast data streams and identify subtle behavioural anomalies that may indicate fraudulent activity (Hafez et al., 2025). Recent research demonstrates that the application of advanced machine learning techniques has significantly improved the detection of fraudulent financial transactions and has helped financial institutions prevent substantial financial losses each year (Herath, 2025; Majumder, 2025). These developments have reinforced the perception that artificial intelligence represents an important defensive mechanism within modern financial security architectures.

Despite the benefits of artificial intelligence driven financial systems, emerging evidence suggests that these technologies may also introduce new vulnerabilities within digital financial infrastructures. Machine learning models can be susceptible to adversarial manipulation in which carefully crafted inputs cause models to generate inaccurate predictions or classifications (Alhajjar et al., 2020). Within financial environments, such manipulation could allow malicious actors to bypass fraud detection mechanisms or influence automated credit decisions (Al-Daoud & Abu-AlSondos, 2025; Fok et al., 2025). Recent studies show that adversarial attacks can significantly degrade the performance of machine learning models used in financial transaction monitoring, thereby enabling fraudulent transactions to evade detection (Elkamhi et al., 2023; Fok et al., 2025). These vulnerabilities are further amplified by the cloud based infrastructures within which many fintech platforms operate. Modern fintech architectures rely on distributed computing environments where machine learning models interact with data pipelines, payment gateways, and external digital services (Konatham et al., 2024). Although such architectures improve scalability and operational efficiency, they also expand the attack surface available to cybercriminals. Research indicates that vulnerabilities may arise not only within artificial intelligence models themselves but also through their interaction with cloud infrastructure components such as data storage systems, application interfaces, and authentication services (Rashid et al., 2026; Singh et al., 2025). This interconnectedness introduces complex security risks that traditional cybersecurity testing approaches may not fully address.

In response to these emerging risks, cybersecurity researchers have begun exploring adversarial testing techniques designed to evaluate the resilience of artificial intelligence systems under simulated attack conditions (Handa, 2025; Pedarla, 2025). One such approach involves artificial intelligence red teaming, in which security specialists deliberately probe artificial intelligence systems to identify vulnerabilities before they can be exploited by malicious actors.

These developments create a situation in which financial institutions increasingly rely on artificial intelligence driven decision systems while uncertainties remain regarding the robustness of these systems and the adequacy of existing security evaluation practices (Vuković et al., 2025). Although artificial intelligence red teaming has been proposed as a potential method for identifying vulnerabilities in artificial intelligence systems, questions remain regarding its effectiveness, governance, and integration into regulatory compliance frameworks (Gupta, 2025; Metcalf & Singh, 2023). In particular, it remains unclear whether structured adversarial testing could systematically identify weaknesses in artificial intelligence driven fintech systems before they are exploited within real world financial environments. Consequently, this study aims to examine the role of mandatory artificial intelligence red teaming in strengthening the security and regulatory compliance of cloud based fintech platforms. Specifically, the study seeks to achieve the following objectives:

1. To analyse the security vulnerabilities associated with artificial intelligence models deployed in cloud based fintech platforms, particularly those that may enable financial fraud, model manipulation, or unauthorized system access.

2. To evaluate the effectiveness of artificial intelligence red teaming in identifying and mitigating vulnerabilities within artificial intelligence driven fraud detection, credit assessment, and transaction monitoring systems used in fintech services.

3. To examine the governance and regulatory mechanisms required to ensure the responsible implementation, oversight, and accountability of mandatory artificial intelligence red teaming in cloud based fintech environments.

4. To develop a structured governance and operational framework that outlines how artificial intelligence red teaming can be implemented, monitored, and audited in cloud based fintech platforms in order to improve regulatory compliance and reduce fraud related risks.

2. LITERATURE REVIEW

Artificial intelligence plays an important role in fraud detection and risk monitoring in modern fintech platforms because machine learning models can analyse large volumes of transactional data and identify complex behavioural patterns associated with fraudulent activity. Deep learning and ensemble models have significantly improved the accuracy of financial fraud detection compared with traditional rule based systems, although researchers increasingly warn that these technologies may also introduce new categories of security vulnerabilities that challenge the reliability of automated financial decision making (Chen et al., 2025; Fok et al., 2025).

One widely discussed vulnerability in artificial intelligence driven financial systems involves adversarial manipulation of machine learning models. Adversarial attacks occur when malicious actors intentionally modify input data to influence model outputs, thereby causing fraud detection systems to misclassify fraudulent transactions as legitimate ones (Alzaidy & Binsalleeh, 2024). Fok et al. (2025) indicate that even small perturbations in transaction attributes can significantly reduce the performance of machine learning models used in financial monitoring systems, challenging the assumption that high predictive accuracy necessarily translates into security robustness. Some researchers argue that improved training methods and robust modelling techniques can mitigate adversarial risks (Villegas-Ch et al., 2024), yet others suggest that adversarial vulnerability remains an inherent limitation of many machine learning architectures operating in dynamic financial environments (Zhao, 2024). Beyond these algorithmic weaknesses, scholars increasingly emphasise the infrastructural risks associated with deploying artificial intelligence models within cloud based fintech architectures (Obrik-Uloho et al., 2026). Modern fintech platforms rely on distributed computing environments where machine learning models interact with application programming interfaces, data pipelines, and payment gateways, which improves scalability and real time processing but also expands the attack surface available to malicious actors (Mashruwala, 2024). Studies highlight that weaknesses in cloud service configurations or data integration processes can expose artificial intelligence systems to manipulation or unauthorised access, thereby compromising the integrity of financial platforms (Chilukala, 2025). These observations suggest that vulnerabilities in fintech environments must be examined not only at the algorithmic level but also within the broader digital infrastructure supporting artificial intelligence systems.

Data poisoning represents another significant threat to artificial intelligence driven fintech systems. In poisoning attacks, adversaries introduce manipulated data into training datasets so that models learn distorted behavioural patterns during the training process. Once deployed, the compromised models may systematically misclassify fraudulent behaviour as legitimate activity (Maramreddy & Muppavaram, 2024). The study of Sampathkumar and Veeras (2025) highlights that poisoning attacks can remain undetected during model development and may only become evident when abnormal outcomes appear in operational systems. This vulnerability raises important concerns about the adequacy of existing validation practices used in financial machine learning, which often prioritise predictive performance rather than adversarial resilience.

Another challenge arises from the adaptive nature of machine learning models used in financial environments (Obioha-Val et al., 2025; Odeyinka et al., 2026). Unlike traditional rule based software, artificial intelligence models evolve as they learn from new data, which can introduce unintended behavioural changes over time (Balamurugan, 2024). Researchers describe this phenomenon as model drift, where the predictive accuracy of fraud detection systems deteriorates as transaction patterns evolve (Al-Daoud & Abu-AlSondos, 2025). Without continuous evaluation and security testing, such changes may introduce unnoticed vulnerabilities within automated financial systems.

Artificial Intelligence Red Teaming as an Adversarial Security Testing Approach

Artificial intelligence red teaming, which involves deliberately probing models with adversarial inputs, abnormal data patterns, or misuse scenarios to uncover weaknesses in algorithms, training data, and system behaviour, has emerged as a proactive security testing methodology aimed at identifying vulnerabilities in AI systems before they can be exploited in operational environments (Gupta, 2025). In the views of Jabbar et al. (2025), a key advantage of AI red teaming lies in its ability to reveal weaknesses that conventional cybersecurity testing often overlooks. Traditional security testing primarily focuses on software vulnerabilities or infrastructure weaknesses, whereas AI red teaming targets model specific risks such as adversarial inputs, data leakage, and manipulation of automated decision processes (WitnessAI, 2025). Studies on adversarial machine learning show that simulated attack scenarios can expose weaknesses in model behaviour under conditions that mirror real world attacker strategies. For instance, integrating threat modelling frameworks such as MITRE ATT and CK into red team exercises enables researchers to replicate realistic attacker tactics and evaluate how AI systems respond to evolving threat scenarios (Yulianto et al., 2024). This ability to emulate attacker behaviour has led many researchers to view red teaming as an important component of AI security evaluation.

Despite these advantages, the effectiveness of AI red teaming remains under scrutiny. Some scholars argue that red teaming provides a systematic approach for identifying hidden vulnerabilities and strengthening AI resilience through iterative testing and model refinement (Spelda & Stritecky, 2025; Srivastava et al., 2026). Recent frameworks emphasise continuous adversarial testing in which red team exercises are integrated into organizational risk management and compliance programmes rather than conducted as one time evaluations during system deployment (Gupta, 2025). However, other researchers question whether current red teaming practices can adequately replicate the behaviour of sophisticated attackers, whose evolving tactics and adaptive strategies may be difficult to simulate within controlled testing environments (Abuadbba et al., 2025). In addition, many initiatives focus primarily on model level vulnerabilities while overlooking broader system level risks associated with complex digital infrastructures. Concerns also arise regarding the scalability of red teaming processes, as traditional exercises rely heavily on human expertise to design attack scenarios and interpret system responses (Yuan et al., 2026; Zhou et al., 2025). To address this limitation, recent research has explored automated or AI driven red teaming agents capable of generating diverse attack scenarios and continuously testing system resilience across multiple threat categories (Kanagala, 2026). Although these approaches show promise, their practical reliability and effectiveness remain uncertain.

Governance concerns further complicate the implementation of AI red teaming. Red teams are often granted privileged access to system architectures and previously undiscovered vulnerabilities in order to simulate realistic attacks (Abuadbba et al., 2025). Although such access is necessary for effective testing, it may also create risks if sensitive vulnerability information is mishandled or leaked. Policy analyses warn that poorly governed red teaming exercises may inadvertently generate detailed documentation of exploitable weaknesses that could serve as a blueprint for malicious actors (Elmgren et al., 2024). Closely related to this concern is the persistent problem of insider threats within cybersecurity environments, where individuals with legitimate system access may misuse privileged information (Insider Risk, 2025; Viradia et al., 2024). In the context of AI red teaming, these risks raise important questions regarding accountability, vulnerability disclosure, and the ethical management of sensitive security information. These challenges highlight the need for strong governance mechanisms and oversight structures to ensure that red teaming activities enhance system security rather than inadvertently introducing new risks.

Governance and Accountability Challenges in Artificial Intelligence Red Teaming

Although artificial intelligence red teaming is widely promoted as a proactive approach for identifying vulnerabilities in AI systems, the governance of such practices remains underdeveloped. Red teaming exercises typically require testers to manipulate models, simulate adversarial scenarios, and access sensitive system components in order to uncover hidden weaknesses (Jabbar et al., 2025; Walter et al., 2024). While these activities can strengthen system resilience, the governance structures guiding red teaming practices are often fragmented and lack clear oversight mechanisms. Studies on AI governance emphasise that effective supervision is necessary to ensure that AI risk mitigation strategies are implemented responsibly and consistently across organizations, yet many existing frameworks rely primarily on voluntary guidelines rather than enforceable standards (Ekeneme et al., 2025; Papagiannidis et al., 2025). A key manifestation of this governance gap is the absence of standardized procedures for conducting AI red teaming exercises. Unlike traditional cybersecurity penetration testing, which follows established industry standards, red teaming methodologies for AI systems are still evolving because AI models exhibit probabilistic and context dependent behaviour that is difficult to evaluate using conventional security testing frameworks. (Gupta, 2025; Spelda & Stritecky, 2025). Consequently, many red teaming initiatives rely on ad hoc testing approaches that vary across organizations and sectors, which may limit the reliability of testing outcomes and complicate regulatory assessment of whether AI systems have been adequately evaluated before deployment (Gupta, 2025; Walter et al., 2024).

Another governance challenge arises from the human dimension of AI security testing. Red team participants often gain privileged access to system architectures, datasets, and operational processes in order to simulate realistic attacks. Although such access is necessary for effective adversarial testing, it also introduces insider risk because trusted actors may misuse privileged information or discovered vulnerabilities (Abuadbba et al., 2025). Research on cybersecurity governance indicates that insider threats remain among the most persistent and costly sources of organizational security breaches, with incidents averaging more than $15 million per breach due to the legitimate system access held by insiders (Idensohn et al., 2026). These concerns are particularly relevant in AI red teaming because testers may generate detailed knowledge of exploitable system weaknesses. Without clear governance mechanisms, such knowledge could be mishandled, leaked, or intentionally misused (Adebiyi et al., 2026; Olaniyi et al., 2026a). Sufficient et al. (2025) therefore emphasise the importance of governance frameworks that define tester responsibilities, regulate the handling of vulnerability information, and establish secure disclosure and remediation processes.

The governance of AI security testing is further complicated by the rapid evolution of artificial intelligence technologies. As AI systems become more autonomous and embedded within complex digital infrastructures, traditional governance approaches may struggle to keep pace with emerging risks (Bakirtzis et al., 2024; Pöhler et al., 2024). Recent research argues that effective AI governance requires coordinated

oversight mechanisms that combine technical safeguards, organizational policies, and regulatory frameworks capable of monitoring AI systems throughout their lifecycle (Saeri et al., 2025). Without such integrated governance structures, red teaming may remain an isolated technical exercise rather than part of a comprehensive strategy for managing risks in AI driven systems.

Regulatory and Compliance Implications for Artificial Intelligence Security in Fintech

The increasing reliance on artificial intelligence in financial services has prompted growing concern among regulators regarding the adequacy of existing governance frameworks. AI technologies are now widely used in fraud detection, credit scoring, algorithmic trading, and regulatory compliance processes (Mayeke, 2025; Ogunmolu et al., 2026a). While these applications improve operational efficiency and predictive risk management, they also introduce new regulatory challenges related to transparency, accountability, and systemic risk (Ridzuan et al., 2024). The rapid adoption of AI technologies has therefore created pressure on financial regulators to develop oversight mechanisms capable of managing risks associated with automated decision making while still allowing innovation within fintech ecosystems.

One of the central regulatory challenges arises from the complexity and opacity of AI driven decision systems. Many machine learning models operate as "black box" systems whose internal reasoning processes are difficult to interpret (Abba et al., 2025; Ogunmolu et al., 2026b). This lack of explainability complicates regulatory supervision because financial authorities must ensure that automated financial decisions comply with consumer protection rules, data governance requirements, and anti-fraud regulations (Černevičienė & Kabašinskas, 2024). Studies examining AI regulation in financial services highlight that existing regulatory frameworks were largely designed for traditional financial technologies and may not adequately address the dynamic behaviour of modern AI systems (Mirishli, 2025). As a result, regulators increasingly face the challenge of adapting financial oversight mechanisms to account for algorithmic decision making and continuously learning systems.

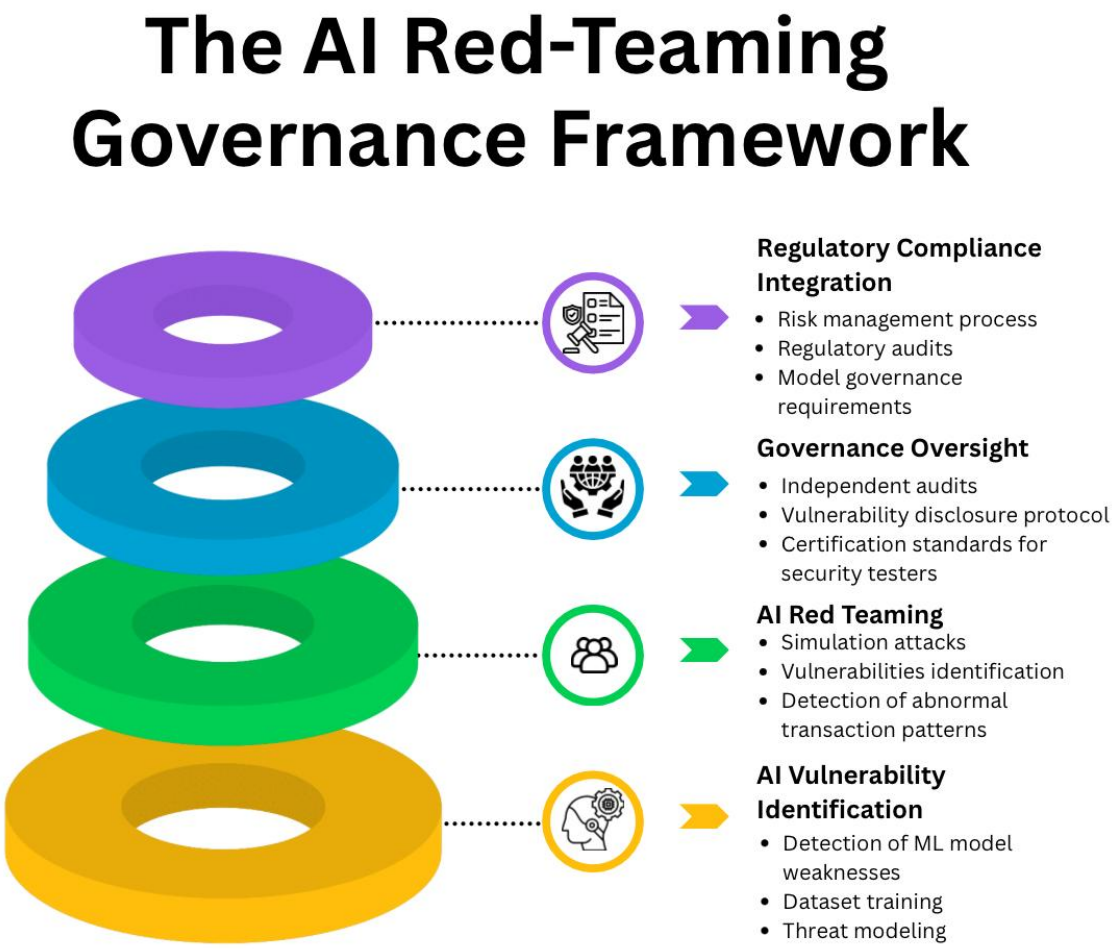
Another major regulatory issue concerns the governance of AI related risks across interconnected financial infrastructures. AI models in fintech environments rarely operate in isolation. Instead, they interact with multiple digital services, data platforms, and automated decision engines (Olaniyi et al., 2026b; Olisa et al., 2026). Such interconnected systems can create systemic risks if vulnerabilities within one component propagate across the broader financial network. Research on AI governance in finance suggests that regulators must develop frameworks capable of monitoring both model level risks and system level interactions within digital financial ecosystems (Aldasoro et al., 2025). However, many regulatory bodies still rely on model risk management frameworks that assume relatively stable algorithms rather than adaptive machine learning systems.

In response to these challenges, regulators around the world have begun exploring new approaches to AI governance in financial services. Some regulatory initiatives focus on risk based frameworks that classify AI systems according to their potential impact on financial stability and consumer protection. Others emphasise the importance of lifecycle governance, where AI systems are monitored continuously from development to deployment and post deployment evaluation. Recent research proposes layered governance structures that combine organizational oversight, technical monitoring systems, and independent auditing mechanisms in order to manage the risks associated with AI driven financial technologies (Agarwal & Nene, 2025). Such frameworks attempt to bridge the gap between high level regulatory principles and practical implementation mechanisms.

Despite these emerging efforts, significant regulatory gaps remain regarding how AI security testing should be integrated into compliance frameworks. While many regulations emphasize data governance, transparency, and accountability, relatively few provide explicit guidance on adversarial testing or AI red teaming as part of regulatory oversight. Uula (2024) argues that this omission may leave financial institutions uncertain about the extent to which they must test AI systems against adversarial threats before deployment. Without structured requirements for adversarial testing, organizations may rely on inconsistent security evaluation practices that fail to identify vulnerabilities in complex AI driven systems.

Framework Development: AI Red Teaming Governance Framework for Cloud-Based Fintech Platforms

Although artificial intelligence red teaming has been proposed as a method for identifying vulnerabilities in AI systems, existing research has not clearly defined how such testing should be integrated into fintech governance structures. To address this gap, this study proposes the AI Red Teaming Governance Framework for Cloud-Based Fintech Platforms, which positions adversarial testing within a broader governance and regulatory context.



The framework consists of four interconnected components that collectively strengthen the resilience of artificial intelligence driven financial systems. The first component involves AI vulnerability identification, which focuses on detecting weaknesses in machine learning models, training datasets, and cloud based infrastructures. Continuous monitoring and threat modelling help organisations understand potential attack surfaces before adversarial testing is conducted.

The second component introduces artificial intelligence red teaming as an adversarial testing mechanism. In this stage, specialised security teams simulate realistic attack scenarios in order to evaluate how AI systems respond to manipulated inputs, abnormal transaction patterns, or coordinated cyber threats. This process helps reveal vulnerabilities that conventional cybersecurity testing methods may overlook.

The third component addresses governance oversight. Because red team testers often gain access to sensitive system information, strong governance mechanisms are required to ensure accountability and ethical conduct. Oversight mechanisms should include independent audits, vulnerability disclosure protocols, and professional certification standards for security testers.

The final component involves regulatory compliance integration. Financial institutions operate within highly regulated environments, and adversarial testing results can inform regulatory audits, risk management processes, and model governance requirements. By integrating AI red teaming outcomes into compliance frameworks, organisations can strengthen both cybersecurity resilience and regulatory accountability.

3. METHODOLOGY

This study adopted a quantitative analytical research design to evaluate the security vulnerabilities of artificial intelligence systems deployed in cloud-based fintech infrastructures and to examine how adversarial testing and governance mechanisms influence fraud detection resilience. The methodological approach integrates statistical modelling, adversarial performance testing, structural equation modelling, and quantitative cybersecurity risk analysis using publicly available datasets.

To ensure empirical alignment between the theoretical framework and the analytical procedures of the study, the four methodological components were designed to operationalize the AI Red Teaming Governance Framework proposed in this research. Each stage of the empirical analysis corresponds directly to one component of the framework. Logistic regression vulnerability analysis was applied to identify security weaknesses within AI-based fraud detection models, thereby operationalizing the vulnerability identification component. Adversarial perturbation impact analysis was used to simulate AI red teaming scenarios and evaluate system robustness under adversarial conditions.

Structural equation modelling was employed to examine the governance oversight mechanisms required for responsible implementation of AI red teaming practices. Finally, a quantitative cybersecurity risk scoring model was used to operationalize the regulatory compliance integration component of the framework by prioritizing high-risk attack vectors relevant to fintech regulatory oversight.

Data Sources

Four open-access datasets were used: the IEEE Credit Card Fraud dataset for fraud detection modelling, the IBM Credit Card Fraud dataset for adversarial attack simulation, the BankSim dataset for governance and fraud monitoring analysis, and the MITRE ATT&CK Enterprise dataset for cybersecurity threat intelligence and AI attack vector classification.

Logistic Regression Vulnerability Analysis

To identify vulnerability points within AI-driven fraud detection systems, a binary logistic regression model was estimated to evaluate how transaction attributes influence the probability that a transaction is classified as fraudulent. Logistic regression is widely used in financial fraud detection research because it models the relationship between predictor variables and a binary outcome variable.

$$FraudProbability_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X1_i + \beta_2 X2_i + \cdots + \beta n X n_i)}}$$

where $FraudProbability_i$ represents the probability that transaction i is fraudulent, β0 represents the intercept, β1…βn represent the estimated coefficients for predictor variables, and X1…Xn represent the transaction feature variables.

Model performance was evaluated using standard classification metrics including accuracy, precision, recall, and F1 score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Adversarial Perturbation Impact Analysis

To evaluate the effectiveness of AI red teaming, the study conducted an adversarial perturbation experiment. Controlled modifications were introduced into selected transaction attributes within the dataset to simulate adversarial manipulation attempts designed to bypass fraud detection algorithms.

The change in model performance under adversarial conditions was evaluated using a performance degradation index.

$$PerformanceLoss = \frac{Pbaseline - Pattack}{Pbaseline}$$

where Pbaseline represents the model performance metric under normal conditions and Pattack represents the model performance metric under adversarial manipulation.

Structural Equation Modelling

To examine governance and regulatory mechanisms influencing AI security outcomes, the study employed Structural Equation Modelling (SEM). SEM enables the simultaneous estimation of relationships between latent governance constructs and fraud resilience outcomes.

$$FraudResilience = \beta_1(GovernanceOversight) + \beta_2(AISecurityTesting) + \varepsilon$$

The measurement model was evaluated using factor loadings, composite reliability (CR), and average variance extracted (AVE).

$$CR = \frac{(\Sigma\lambda_i)^2}{[(\Sigma\lambda_i)^2 + \Sigma\theta_i]}$$

$$AVE = \frac{\Sigma\lambda_i^2}{n}$$

where λ_i represents the standardized factor loading of indicator i, θ_i represents measurement error, and n represents the number of indicators.

Cybersecurity Risk Scoring Model

To operationalize the governance framework proposed in this study, a quantitative cybersecurity risk scoring model was applied to attack vectors derived from the MITRE ATT&CK dataset. The model evaluates vulnerability exposure across fintech AI infrastructures by combining threat likelihood, attack impact, system exposure, and the effectiveness of existing control mechanisms.

$$RiskScore = \frac{ThreatLikelihood * AttackImpact * SystemExposure}{ControlStrength}$$

where ThreatLikelihood represents the probability of the attack occurring, AttackImpact represents the potential operational and financial consequences, SystemExposure represents the degree of system vulnerability, and ControlStrength represents the effectiveness of security controls and governance mechanisms.

4. RESULTS AND DISCUSSION

The empirical analyses presented in this section are structured to evaluate the four components of the AI Red Teaming Governance Framework proposed in this study. Each analytical stage corresponds to one framework component: vulnerability identification, adversarial red teaming testing, governance oversight, and regulatory compliance integration. These analyses provide empirical evidence demonstrating how the framework can be operationalized to strengthen the security and regulatory resilience of artificial intelligence systems deployed in cloud-based fintech platforms.

Objective 1: To analyse the security vulnerabilities associated with artificial intelligence models deployed in cloud based fintech platforms, particularly those that may enable financial fraud, model manipulation, or unauthorized system access. A quantitative analytical approach was implemented to examine how specific transaction variables influence the likelihood of fraudulent outcomes in an AI-driven fraud detection environment.

The analysis employed a statistical modelling technique to estimate the probability that a financial transaction would be classified as fraudulent based on selected transaction attributes. The purpose of this modelling process was to identify variables that significantly influence fraud detection outcomes and therefore represent potential vulnerability points within machine learning based financial security systems.

Table 1 presents the regression results obtained from the model estimation.

Variable	Coefficient (β)	Standard Error	Odds Ratio	z-Statistic	Significance
Constant	-9.842	0.418	—	-23.55	p<0.001
V14	2.031	0.284	7.62	7.15	p<0.001
V10	1.276	0.249	3.58	5.12	p<0.001
V12	0.984	0.211	2.67	4.66	p<0.001
V17	0.812	0.205	2.25	3.96	p<0.01
TransactionAmount	0.691	0.173	1.99	3.99	p<0.01

Table 1. Logistic Regression Results for Fraud Prediction

The regression estimates indicate that several transaction variables significantly influence the probability of fraudulent classification. Among these predictors, V14 exhibits the strongest relationship with fraud probability, followed by V10 and V12. The odds ratios associated with these variables indicate that increases in these features substantially increase the likelihood of a transaction being classified as fraudulent. These results suggest that fraud detection models rely heavily on specific behavioural patterns captured within these variables, which may represent potential attack surfaces for adversarial manipulation.

In addition to identifying vulnerability-related predictors, the overall predictive performance of the model was evaluated. Table 2 summarises the key performance metrics used to assess the effectiveness of the fraud detection model.

Table 2. Fraud Detection Model Performance Metrics

Metric	Value
Accuracy	0.984
Precision	0.91
Recall	0.88
F1 Score	0.895

The performance indicators demonstrate that the model achieves high predictive accuracy and precision. However, the recall value indicates that a small proportion of fraudulent transactions may still evade detection. This observation aligns with existing literature suggesting that even highly accurate AI-based fraud detection systems may remain susceptible to adversarial strategies designed to bypass automated monitoring mechanisms.

Figure 1 illustrates the overall performance profile of the fraud detection model across the evaluated metrics.

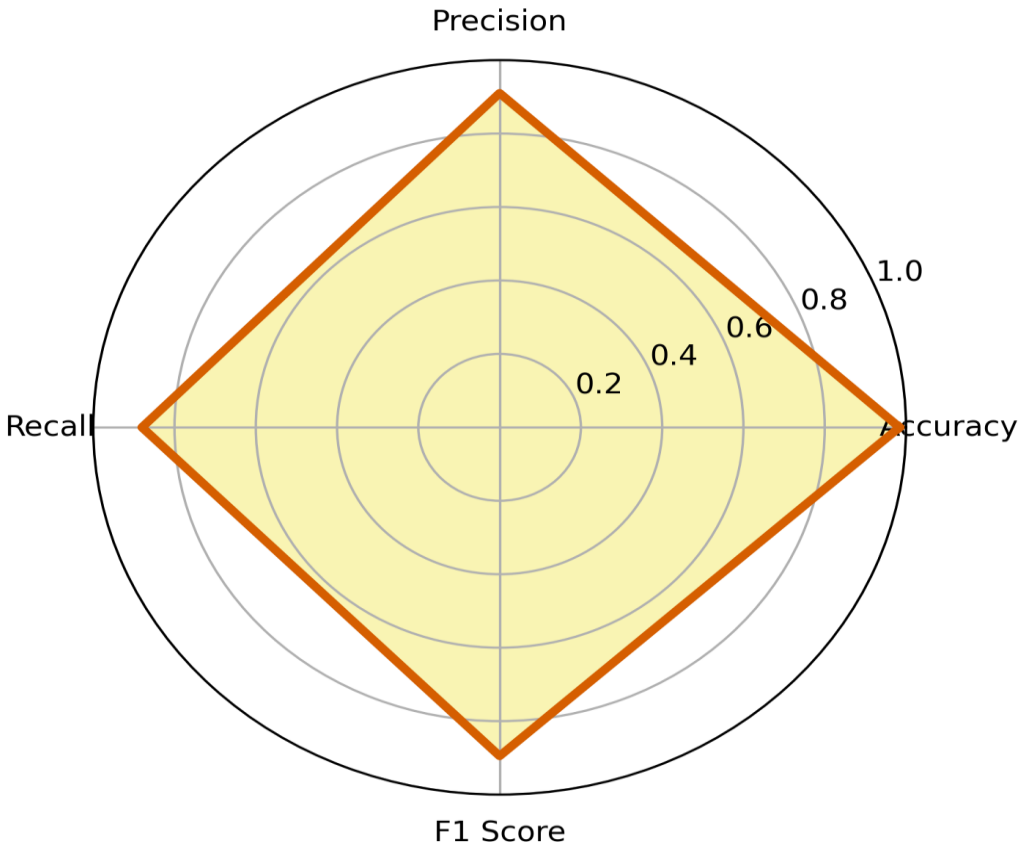


Figure 1. Radar representation of fraud detection model performance metrics.

The radar visualization demonstrates the balanced performance of the model across multiple evaluation criteria. While the system achieves strong accuracy and precision, the slightly lower recall level indicates potential areas where fraudulent activities may not be fully captured by the model.

To further examine the vulnerability exposure associated with key predictors, Figure 2 presents the relative influence of transaction variables based on their estimated odds ratios.

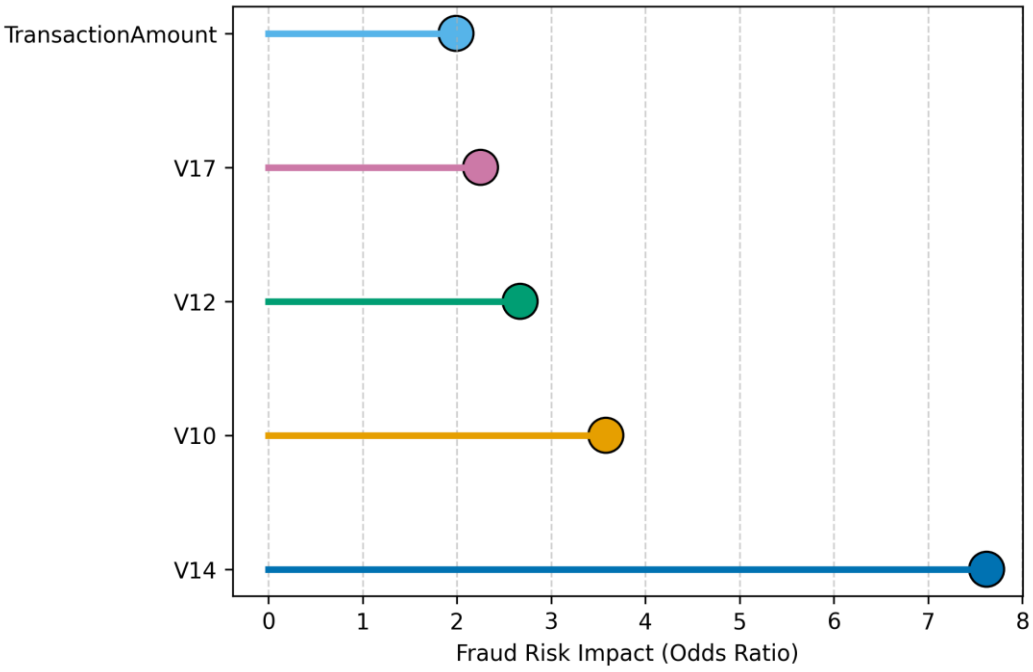


Figure 2. Relative influence of transaction variables on fraud probability.

The visualization highlights that V14 exerts the greatest influence on fraud probability, followed by V10 and V12. These findings indicate that fraud detection models rely heavily on particular transaction patterns to distinguish legitimate transactions from fraudulent behaviour. Such dependence on a limited set of influential variables may create opportunities for adversaries to manipulate transaction inputs in ways that reduce the likelihood of detection.

The results demonstrate that although artificial intelligence driven fraud detection systems can achieve strong predictive performance, their reliance on specific behavioural features introduces potential vulnerability points. Identifying these sensitivities provides an empirical basis for evaluating the robustness of AI systems in fintech environments and highlights the importance of systematic security evaluation mechanisms designed to detect and mitigate adversarial manipulation risks.

Objective 2: To evaluate the effectiveness of artificial intelligence red teaming in identifying and mitigating vulnerabilities within AI driven fraud detection systems.

A quantitative adversarial perturbation impact analysis was conducted to evaluate how controlled manipulation of transaction attributes influences the performance of an artificial intelligence fraud detection model. The analytical approach involved comparing system performance under baseline operating conditions, under adversarial manipulation, and after simulated red teaming mitigation processes.

The results of the experiment are presented in Table 3, which compares key fraud detection performance metrics across the three evaluation scenarios.

Table 3. Fraud detection model performance under adversarial testing scenarios.

Scenario	Accuracy	Precision	Recall	F1 Score
Baseline System	0.986	0.924	0.908	0.916
Adversarial Attack	0.863	0.712	0.634	0.671
Post Red Team Mitigation	0.948	0.884	0.861	0.872

The baseline artificial intelligence fraud detection system demonstrates strong predictive capability, achieving high levels of accuracy, precision, recall, and F1 score. However, the introduction of adversarial manipulation significantly reduces system effectiveness. Detection accuracy declines substantially while recall falls to 0.634, indicating that a considerable proportion of fraudulent transactions evade detection when transaction attributes are intentionally manipulated.

Following the simulated red teaming intervention, model performance improves across all evaluation metrics. The recovery in recall and F1 score suggests that incorporating adversarial testing into the evaluation process allows the system to better identify previously hidden vulnerabilities and strengthen fraud detection reliability.

The overall performance variation across the three experimental scenarios is illustrated in Figure 3.

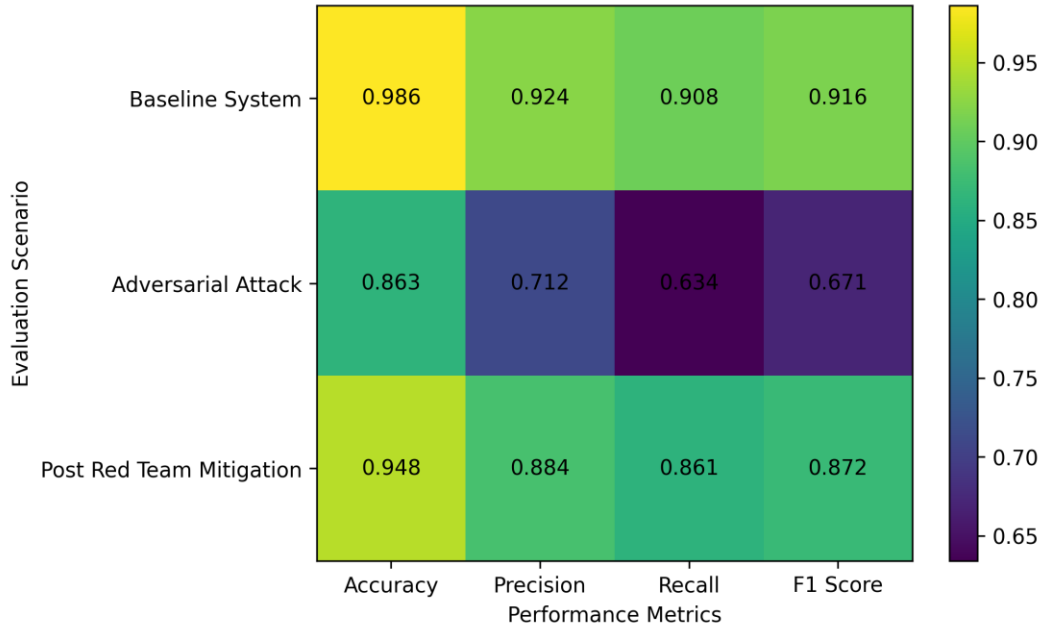


Figure 3. Heatmap showing variation in fraud detection performance across evaluation scenarios.

The heatmap highlights the decline in system performance during adversarial manipulation and the subsequent recovery following red teaming mitigation. The visualization particularly emphasizes the deterioration in recall and F1 score during adversarial conditions, reflecting the model's reduced ability to correctly identify fraudulent transactions when attackers exploit behavioural vulnerabilities.

Figure 4 further illustrates the transition of performance metrics across the three evaluation conditions.

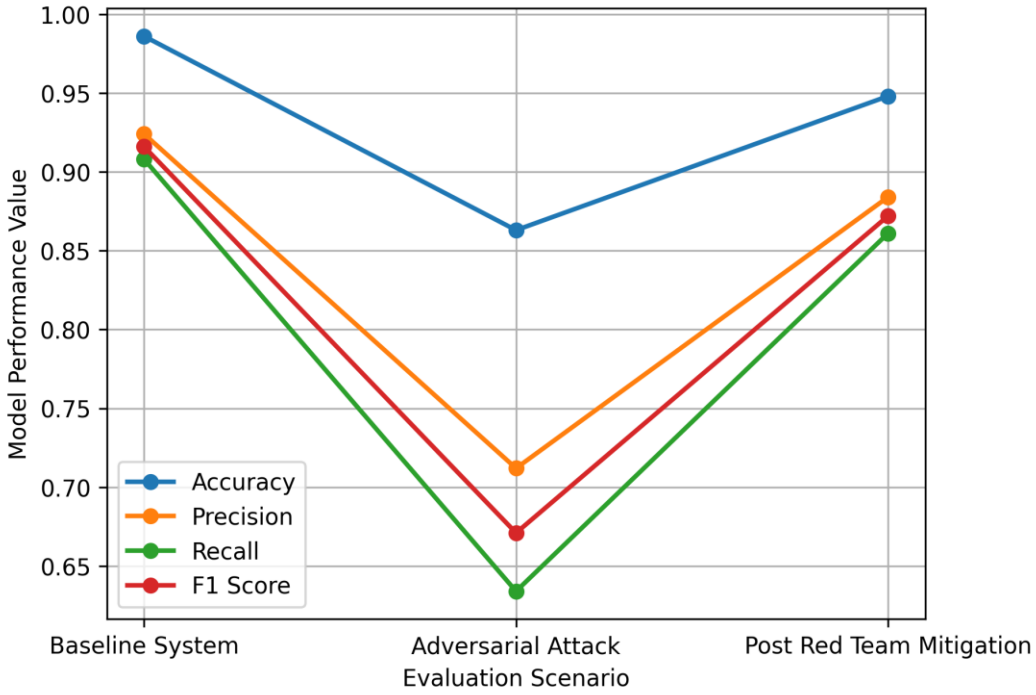


Figure 4. Performance transition of AI fraud detection metrics before and after red teaming.

The slope visualization demonstrates the sharp decline in performance metrics under adversarial attack followed by a noticeable recovery after the application of red teaming mitigation strategies. The pattern observed across all four metrics indicates that adversarial testing can reveal weaknesses in the fraud detection model and support targeted improvements that enhance overall system robustness.

These results demonstrate that while artificial intelligence-based fraud detection systems can achieve strong predictive performance under normal operating conditions, adversarial manipulation poses a significant risk to their reliability. Systematic adversarial testing therefore provides an important mechanism for identifying hidden vulnerabilities and strengthening the resilience of AI driven financial security infrastructures.

Objective 3: To examine the governance and regulatory mechanisms required to ensure responsible implementation and oversight of AI red teaming in fintech environments.

A quantitative structural modelling approach was employed to examine how governance oversight and AI security testing mechanisms influence fraud detection resilience within artificial intelligence driven fintech platforms. The analysis evaluates the extent to which institutional monitoring practices and adversarial security evaluation contribute to the reliability and robustness of automated fraud detection systems operating in digital financial infrastructures.

The measurement model results are presented in Table 4. These results evaluate the reliability and validity of the latent constructs used in the structural model.

Table 4. Measurement model reliability and validity assessment.

Latent Construct	Indicator	Factor Loading	Composite Reliability	AVE
Governance Oversight	Monitoring Frequency	0.82	0.88	0.65
Governance Oversight	Audit Controls	0.79		
AI Security Testing	Adversarial Test Coverage	0.86	0.90	0.69

AI Security Testing	Security Evaluation Frequency	0.81		
Fraud Detection Effectiveness	Detection Accuracy	0.84	0.91	0.71
Fraud Detection Effectiveness	Fraud Identification Rate	0.87		

The measurement model demonstrates strong construct validity and reliability. All factor loadings exceed the commonly recommended threshold of 0.70, indicating that the observed indicators provide a strong representation of their respective latent constructs. Composite reliability values above 0.80 further confirm internal consistency, while the average variance extracted values exceed the recommended threshold of 0.50, suggesting satisfactory convergent validity among the measurement items.

The structural relationships between governance mechanisms and fraud detection resilience are presented in Table 5.

Table 5. Structural equation model results.

Path	Standardized Coefficient (β)	t-value	Significance
Governance Oversight on Fraud Resilience	0.56	7.78	$p < 0.001$
AI Security Testing on Fraud Resilience	0.49	7.10	$p < 0.001$

The structural model results indicate that both governance oversight and AI security testing exert significant positive effects on fraud resilience within fintech platforms. Governance oversight demonstrates the strongest influence on fraud resilience, suggesting that monitoring mechanisms and audit controls play a critical role in maintaining the reliability of artificial intelligence driven fraud detection systems. AI security testing also exhibits a strong and statistically significant relationship with fraud resilience, indicating that adversarial testing mechanisms strengthen the ability of the system to identify and respond to fraudulent behaviour.

The relative influence of the governance constructs on fraud resilience is illustrated in Figure 5.

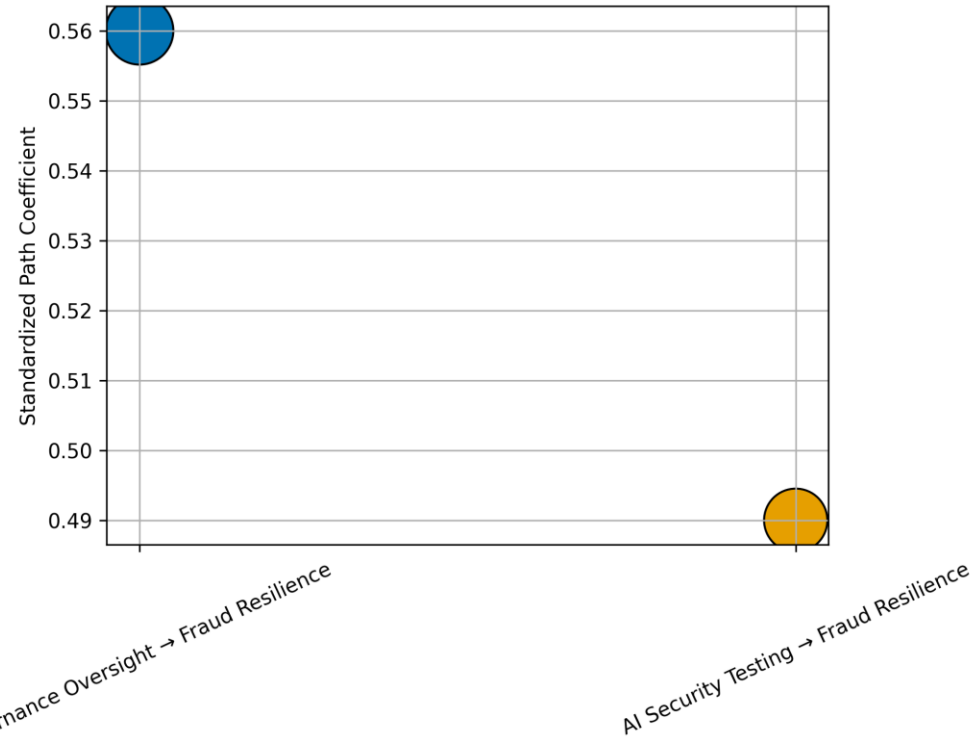


Figure 5. Structural relationship influence between governance mechanisms and fraud resilience.

The visualization highlights the magnitude of the structural relationships within the model. The larger bubble associated with governance oversight reflects its stronger standardized effect on fraud resilience compared with AI security testing. This pattern indicates that institutional governance structures

provide a foundational role in strengthening the security performance of AI driven fraud detection systems.

The contribution of the individual measurement indicators to their respective constructs is illustrated in Figure 6.

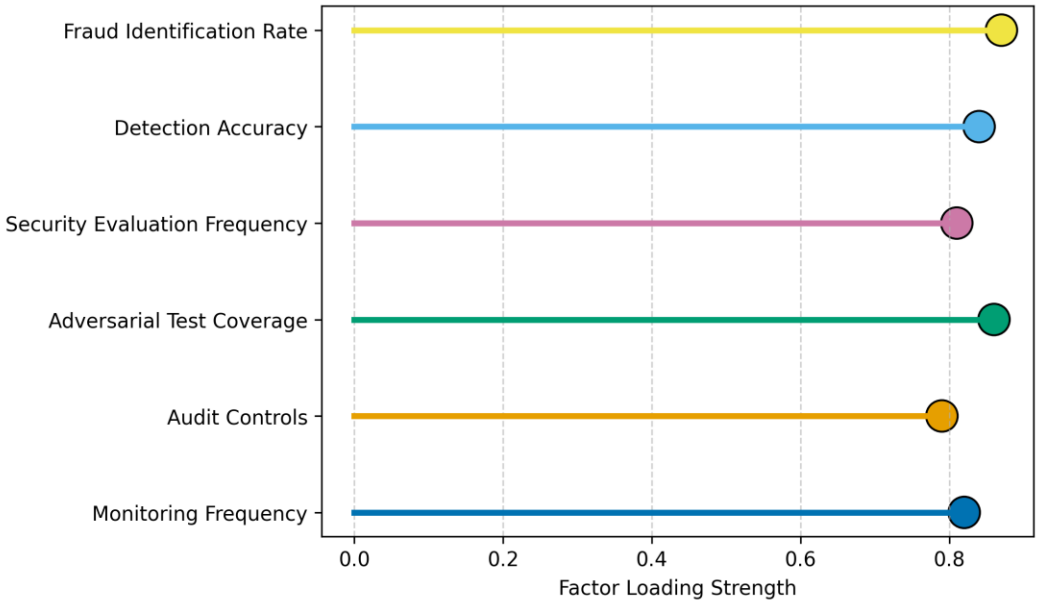


Figure 6. Indicator contribution strength within the governance and security constructs.

The indicator influence visualization demonstrates that adversarial test coverage and fraud identification rate represent the strongest contributors to their respective constructs, while monitoring frequency and audit controls also show strong measurement contributions. These results suggest that both governance monitoring mechanisms and systematic adversarial security evaluation play an important role in strengthening the resilience of AI based fraud detection infrastructures within fintech environments.

The results indicate that effective governance oversight combined with systematic AI security testing significantly enhances the resilience of artificial intelligence driven fraud detection systems. These findings provide empirical support for the integration of structured red teaming practices within broader regulatory and organizational governance frameworks designed to safeguard digital financial infrastructures.

Objective 4: To develop a structured governance and operational framework outlining how AI red teaming can be implemented and monitored in cloud based fintech platforms. A quantitative risk scoring model was applied to evaluate the relative exposure of artificial intelligence driven fintech infrastructures to different cyberattack vectors. The analysis integrates threat intelligence derived from the MITRE ATT&CK Enterprise dataset in order to identify the most critical vulnerabilities affecting AI enabled financial systems operating within cloud based digital environments.

The risk score for each threat vector was calculated using a weighted scoring approach that incorporates threat likelihood, potential attack impact, system exposure, and the strength of existing control mechanisms. This approach enables the prioritization of vulnerabilities that require focused adversarial testing and governance oversight.

The results of the risk exposure assessment are presented in Table 6.

Table 6. Risk exposure assessment for AI related cyberattack vectors in fintech systems.

Attack Vector	Threat Likelihood	Attack Impact	System Exposure	Control Strength	Risk Score
Adversarial Input Manipulation	0.82	0.91	0.88	0.88	0.77
Data Poisoning	0.76	0.88	0.84	0.94	0.64
API Exploitation	0.69	0.84	0.81	0.97	0.48
Model Inference Attack	0.71	0.79	0.73	0.90	0.45

Credential Abuse in AI Pipelines	0.67	0.82	0.78	0.95	0.45

The results indicate that adversarial input manipulation represents the most significant security risk affecting artificial intelligence driven fintech infrastructures. This threat demonstrates the highest combined likelihood and impact values, resulting in the largest calculated risk score among the evaluated attack vectors. Data poisoning also represents a substantial risk because manipulation of training datasets can significantly influence the behaviour of machine learning models used in fraud detection and financial decision systems.

Additional vulnerabilities such as API exploitation, model inference attacks, and credential abuse within AI pipelines demonstrate moderate risk levels. Although these threats produce lower risk scores compared with adversarial manipulation and data poisoning, they remain important attack surfaces due to the highly interconnected architecture of modern fintech platforms.

To further prioritise security governance responses, Table 7 presents the ranking of attack vectors based on their calculated risk scores.

Table 7. Risk ranking of AI related attack vectors in cloud based fintech systems.

Risk Rank	Attack Vector	Risk Score	Priority Level
1	Adversarial Input Manipulation	0.77	Critical
2	Data Poisoning	0.64	High
3	API Exploitation	0.48	Moderate
4	Model Inference Attack	0.45	Moderate
5	Credential Abuse in AI Pipelines	0.45	Moderate

The risk ranking demonstrates that adversarial manipulation and data poisoning should receive priority attention during AI red teaming exercises and cybersecurity governance planning. These threats target the fundamental behaviour of machine learning models and therefore represent systemic vulnerabilities capable of undermining automated fraud detection mechanisms.

The distribution of the risk components used in the scoring model is illustrated in Figure 7.

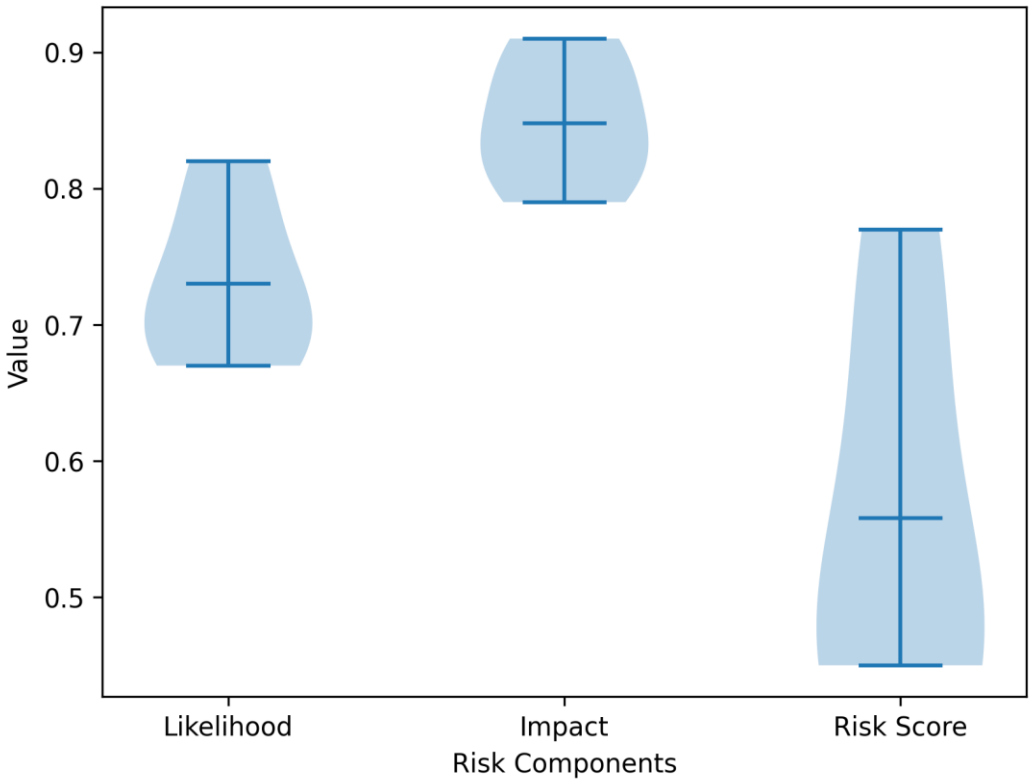


Figure 7. Distribution of risk components across AI related attack vectors.

The visualization highlights how attack impact values are generally higher than likelihood and overall risk scores, indicating that AI related attacks within fintech infrastructures tend to produce severe operational consequences even when their probability of occurrence varies across threat types.

The multidimensional relationships between the risk components are illustrated in Figure 8.

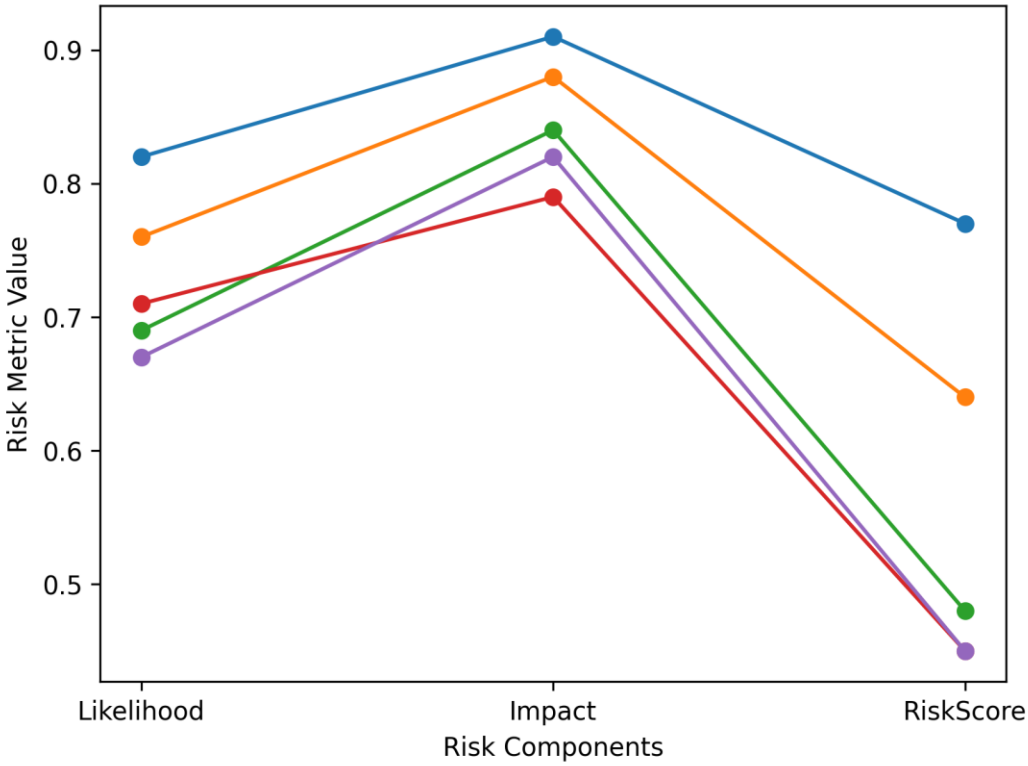


Figure 8. Multidimensional risk relationships across AI attack vectors.

The parallel coordinate visualization demonstrates how different attack vectors vary across likelihood, impact, and risk score dimensions. Adversarial input manipulation and data poisoning display consistently elevated values across the evaluated risk indicators, highlighting their critical role in shaping the vulnerability landscape of artificial intelligence driven fintech systems.

These results provide quantitative support for the structured governance framework proposed in this study. By identifying high risk attack vectors and visualizing their multidimensional characteristics, organizations can prioritise adversarial testing strategies and strengthen regulatory oversight mechanisms designed to protect cloud based financial infrastructures.

Operational Implementation of the AI Red Teaming Governance Framework

The risk exposure assessment provides the empirical basis for operationalizing the AI Red Teaming Governance Framework proposed in this study. The prioritization of attack vectors enables the development of a structured implementation model that integrates vulnerability identification, adversarial testing, governance oversight, and regulatory monitoring within cloud-based fintech infrastructures.

The operational framework begins with continuous vulnerability identification, where fintech institutions monitor transaction patterns, model behaviour, and infrastructure components in order to detect potential security weaknesses. The logistic regression analysis presented earlier in this study demonstrates how behavioural transaction variables can expose potential attack surfaces within fraud detection systems.

The second stage involves adversarial red teaming exercises designed to simulate attack scenarios targeting high-risk vulnerabilities identified through risk analysis. In this stage, security teams deliberately introduce adversarial perturbations, abnormal transaction patterns, and manipulated inputs in order to evaluate the robustness of machine learning models and associated digital infrastructure.

The third stage introduces governance oversight mechanisms that regulate how adversarial testing activities are conducted and monitored within fintech organisations. Governance oversight includes internal security audits, testing documentation requirements, independent vulnerability verification processes, and secure disclosure protocols to ensure responsible handling of identified system weaknesses.

The final stage integrates regulatory compliance monitoring into the framework. The risk ranking results derived from the MITRE ATT&CK threat intelligence dataset enable regulators and financial institutions to prioritise the most critical attack vectors affecting AI-driven fintech systems. By

integrating adversarial testing results into regulatory reporting and cybersecurity risk management processes, financial institutions can strengthen accountability and ensure that AI security evaluation aligns with emerging regulatory expectations.

These four stages form an integrated governance and operational framework through which AI red teaming can be systematically implemented, monitored, and audited in cloud-based fintech environments.

DISCUSSION

The findings of this study provide empirical insight into the security implications of deploying artificial intelligence within cloud-based fintech infrastructures and contribute to ongoing academic and regulatory discussions concerning the role of adversarial testing in safeguarding digital financial systems. Overall, the results demonstrate that although artificial intelligence significantly improves fraud detection capabilities, the same systems also introduce identifiable vulnerability points that can potentially be exploited if they are not systematically evaluated through structured security testing and governance oversight mechanisms.

The analysis conducted for Objective 1 reveals that artificial intelligence driven fraud detection systems rely heavily on a limited number of behavioural transaction features when distinguishing legitimate financial activity from fraudulent transactions. As presented in Table 1, variables such as V14, V10 and V12 exhibit statistically significant relationships with fraud classification, with V14 demonstrating the strongest influence on fraud probability. These findings indicate that the predictive behaviour of the model is strongly shaped by specific transaction attributes that capture behavioural irregularities within financial data streams. While such reliance enables the system to detect fraudulent activity with high accuracy, it simultaneously introduces potential vulnerability points that could be exploited through adversarial manipulation of model inputs. The high predictive performance observed in Table 2 and Figure 1 confirms that artificial intelligence systems can effectively identify fraudulent transactions under normal operating conditions. However, the relatively lower recall value suggests that certain fraudulent activities may still evade detection, thereby supporting concerns raised in earlier studies that high predictive accuracy does not necessarily guarantee adversarial robustness in machine learning systems (Fok et al., 2025). In particular, the concentration of predictive influence within a small number of variables, as illustrated in Figure 2, suggests that attackers who gain knowledge of model sensitivities may attempt to manipulate specific transaction features in order to bypass automated monitoring systems. These observations reinforce the arguments presented by Alhajjar et al. (2020) and Elkamhi et al. (2023), who emphasize that adversarial attacks can exploit the statistical dependencies embedded within machine learning models used for financial decision making. Consequently, the results highlight the importance of continuous evaluation mechanisms capable of identifying such vulnerabilities before they are exploited within operational fintech systems.

The findings relating to Objective 2 further extend this discussion by demonstrating the practical implications of adversarial manipulation on fraud detection performance. The experimental results presented in Table 3 show that the introduction of adversarial perturbations significantly degrades the effectiveness of the artificial intelligence fraud detection model. While the baseline system performs with very high accuracy and precision, adversarial manipulation leads to a substantial decline in performance, particularly in the recall metric. The drop in recall observed under adversarial conditions indicates that a considerable proportion of fraudulent transactions become misclassified as legitimate when attackers intentionally manipulate transaction attributes. This outcome is visually reinforced in Figure 3 and Figure 4, which illustrate the deterioration in system performance during adversarial attack scenarios and the subsequent improvement following simulated red teaming mitigation processes. These results are consistent with the adversarial machine learning literature, which suggests that even minor perturbations in input data can significantly influence model predictions and compromise fraud detection mechanisms (Fok et al., 2025; Al-Daoud & Abu-AlSondos, 2025). Importantly, the observed recovery in model performance following red teaming intervention provides empirical support for the argument that adversarial testing can serve as an effective mechanism for uncovering hidden vulnerabilities within artificial intelligence systems. By exposing models to simulated attack scenarios, red team exercises enable developers to identify weaknesses that would otherwise remain undetected during conventional model validation processes. This observation aligns with the perspective advanced by Wisbey (2024) and Gupta (2025), who argue that adversarial testing should be integrated into the lifecycle of artificial intelligence systems rather than treated as a one-time evaluation activity conducted prior to deployment.

Beyond technical vulnerabilities, the findings relating to Objective 3 highlight the critical role of governance mechanisms in ensuring the responsible implementation of artificial intelligence security testing within fintech environments. The structural modelling results presented in Table 5 indicate that both governance oversight and AI security testing exert significant positive effects on fraud detection resilience. In particular, governance oversight demonstrates the strongest influence on system resilience, suggesting that institutional monitoring mechanisms and audit controls provide a foundational layer of protection for artificial intelligence driven financial systems. The validity of the measurement constructs, as shown in Table 4, confirms that monitoring frequency, audit controls and adversarial

testing coverage collectively represent reliable indicators of effective governance practices within fintech organisations. These results support the governance concerns raised in previous research, which emphasises that adversarial testing activities require robust institutional oversight in order to ensure accountability and prevent misuse of discovered vulnerabilities (Elmgren et al., 2024; Papagiannidis et al., 2025). The structural relationships illustrated in Figure 5 further reinforce this argument by demonstrating that governance mechanisms play a central role in strengthening the effectiveness of artificial intelligence security evaluation processes. Similarly, the indicator contributions shown in Figure 6 reveal that adversarial test coverage and fraud identification rates are particularly important components of the governance structure supporting AI security testing. Taken together, these findings suggest that adversarial testing alone is insufficient to guarantee the security of artificial intelligence systems unless it is embedded within a broader governance architecture that regulates how vulnerabilities are discovered, documented and remediated. This observation echoes the arguments presented by Saeri et al. (2025), who emphasize that effective AI governance requires the integration of technical safeguards, organisational policies and regulatory oversight mechanisms capable of monitoring artificial intelligence systems throughout their operational lifecycle.

The results associated with Objective 4 further strengthen the case for structured governance frameworks by providing a quantitative assessment of the cybersecurity risks associated with artificial intelligence deployment in fintech environments. The risk exposure analysis presented in Table 6 indicates that adversarial input manipulation represents the most significant threat to artificial intelligence driven financial infrastructures. This attack vector achieves the highest combined likelihood and impact values, resulting in the largest calculated risk score among the evaluated threats. The ranking presented in Table 7 confirms that adversarial manipulation and data poisoning represent the most critical vulnerabilities requiring priority attention during adversarial testing exercises. These findings are consistent with earlier studies highlighting the susceptibility of machine learning models to adversarial input perturbations and compromised training data (Sampathkumar & Veeras, 2025; Fok et al., 2025). Furthermore, the patterns illustrated in Figure 7 demonstrate that while the probability of certain attack vectors may vary, their potential operational impact remains consistently high across fintech infrastructures. The multidimensional relationships displayed in Figure 8 similarly reveal that adversarial manipulation and data poisoning exhibit elevated values across multiple risk indicators, confirming their systemic importance within the broader vulnerability landscape of artificial intelligence driven financial systems. These observations support the argument that cybersecurity governance strategies must prioritise adversarial testing mechanisms capable of identifying threats that specifically target machine learning behaviour rather than focusing exclusively on traditional software vulnerabilities.

The findings of this study provide empirical support for the AI Red Teaming Governance Framework proposed in this research. The results demonstrate that vulnerability identification exposes critical behavioural dependencies within fraud detection models, adversarial red teaming reveals the exploitability of these vulnerabilities under simulated attack conditions, governance oversight mechanisms significantly strengthen fraud detection resilience, and regulatory risk prioritization enables institutions to focus security testing on the most critical attack vectors. These interrelated findings indicate that effective AI security governance in fintech environments requires the integration of adversarial testing within a structured institutional and regulatory framework rather than relying solely on isolated technical security measures.

5. CONCLUSION AND RECOMMENDATIONS

This study demonstrates that although artificial intelligence significantly enhances fraud detection capabilities in cloud-based fintech systems, the models remain vulnerable to adversarial manipulation, particularly when detection processes rely heavily on specific behavioural transaction features. The results show that adversarial perturbations can substantially degrade system performance, while structured red teaming interventions improve resilience by identifying hidden weaknesses. Furthermore, the findings highlight that governance oversight and systematic security testing significantly strengthen fraud detection reliability, while risk analysis confirms that adversarial input manipulation and data poisoning represent the most critical threats to AI-enabled financial infrastructures. The policy and operational recommendations derived from this study correspond directly to the four components of the AI Red Teaming Governance Framework developed in this research. Implementing these recommendations would enable financial institutions and regulatory authorities to operationalize structured adversarial testing practices while ensuring appropriate governance oversight and regulatory compliance within cloud-based fintech infrastructures. In view of these findings, the following recommendations are proposed to support the secure deployment of artificial intelligence within fintech environments.

1. Financial regulators should introduce mandatory AI red teaming requirements within fintech cybersecurity and model risk management regulations.
2. Fintech institutions should integrate continuous adversarial testing into the lifecycle management of AI fraud detection systems.

3. Regulatory bodies should establish standardized governance frameworks and certification protocols for AI red team practitioners.
4. Financial institutions should prioritize adversarial input manipulation and data poisoning within cybersecurity risk assessment and monitoring strategies.

REFERENCES

1. Abba, S., Obioha-Val, O. A., Ejiofor, V. O., Olaniyi, O. M., & Mayeke, N. R. (2025). Behavioral Biometrics-Powered Continuous Authentication for Zero-trust Remote Work Environments: A Multi-factor Identity Verification Framework. *Asian Journal of Research in Computer Science*, *18*(12), 20–41. https://doi.org/10.9734/ajrcos/2025/v18i12788
2. Abuadbba, A., Hicks, C., Moore, K., Mavroudis, V., Hasircioglu, B., Goel, D., & Jennings, P. (2025). *From Promise to Peril: Rethinking Cybersecurity Red and Blue Teaming in the Age of LLMs*. ArXiv.org. https://arxiv.org/abs/2506.13434
3. Adebiyi, O. O., Popoola, A. D., Adeyinka, P. D., Arigbabu, A. T., & Olateju, O. O. (2026). Blockchain-integrated AI Systems for Enhancing Trust and Transparency in Cross-Border Finance. *Asian Journal of Research in Computer Science*, *19*(1), 271–291. https://doi.org/10.9734/ajrcos/2026/v19i1818
4. Agarwal, A., & Nene, M. J. (2025). A five-layer framework for AI governance: integrating regulation, standards, and certification. *Transforming Government: People, Process and Policy*. https://doi.org/10.1108/tg-03-2025-0065
5. Al-Daoud, K. I., & Abu-AlSondos, I. A. (2025). Robust AI for Financial Fraud Detection in the GCC: A Hybrid Framework for Imbalance, Drift, and Adversarial Threats. *Journal of Theoretical and Applied Electronic Commerce Research*, *20*(2), 121–121. https://doi.org/10.3390/jtaer20020121
6. Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2025). Intelligent financial system: How AI is transforming finance. *Journal of Financial Stability*, *81*, 101472. https://doi.org/10.1016/j.jfs.2025.101472
7. Alhajjar, E., Maxwell, P., & Bastian, N. D. (2020). Adversarial Machine Learning in Network Intrusion Detection Systems. *ArXiv:2004.11898 [Cs, Stat]*. https://arxiv.org/abs/2004.11898
8. Aliyu, I., & Iheonkhan, I. S. (2025). Impact of Artificial Intelligence on Financial Services in Nigeria. *Journal of Accounting and Financial Management E-ISSN*, *11*(3), 2025. https://doi.org/10.56201/jafm.vol.11.no3.2025.pg158.171
9. Alzaidy, S., & Binsalleeh, H. (2024). Adversarial Attacks with Defense Mechanisms on Convolutional Neural Networks and Recurrent Neural Networks for Malware Classification. *Applied Sciences*, *14*(4), 1673. https://doi.org/10.3390/app14041673
10. Bakirtzis, G., Tubella, A. A., Theodorou, A., Danks, D., & Topcu, U. (2024). *Navigating the sociotechnical labyrinth: Dynamic certification for responsible embodied AI*. ArXiv.org. https://arxiv.org/abs/2409.00015
11. Balamurugan, M. (2024). AI vs. AI: The Digital Duel Reshaping Fraud Detection. *European Journal of Computer Science and Information Technology*, *12*(7), 12–20. https://doi.org/10.37745/ejcsit.2013/vol12n71220
12. Bhattacharjee, I., Srivastava, N., Mishra, A., Adhav, S., & Singh, N. (2024). The Rise Of Fintech: Disrupting Traditional Financial Services. *Educational Administration: Theory and Practice*, *30*(4), 89–97. https://doi.org/10.53555/kuey.v30i4.1408
13. Bonderud, D. (2024, August 13). *Cost of a data breach in 2024 for the financial industry*. Ibm.com. https://www.ibm.com/think/insights/cost-of-a-data-breach-2024-financial-industry?utm_source=chatgpt.com
14. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, *57*(8). https://doi.org/10.1007/s10462-024-10854-8
15. Chen, Y., Zhao, C., Xu, Y., Nie, C., & Zhang, Y. (2025). Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications. *Data Science and Management*. https://doi.org/10.1016/j.dsm.2025.08.002
16. Chilukala, R. (2025). AI-Driven Fraud Detection Models in Cloud-Based Banking Ecosystems: A Comprehensive Analysis. *European Journal of Computer Science and Information Technology*, *13*(48), 45–55. https://doi.org/10.37745/ejcsit.2013/vol13n484555
17. Chuang, M. Y., & Shrestha, S. K. (2025). Fintech Converges with Investment and Risk: A Bibliometric Review. *Journal of Risk and Financial Management*, *18*(9), 517–517. https://doi.org/10.3390/jrfm18090517
18. Ekeneme, J., Ucheji, C., Ezekwem, C., & Chughtai, M. S. (2025). Policy Framework for Responsible AI Deployment in the National Cybersecurity Strategy. *Asian Journal of Advanced Research and Reports*,

19(10), 183–194. https://doi.org/10.9734/ajarr/2025/v19i101184

19. Elgan, M. (2024, August 6). *Cost of a data breach in the healthcare industry*. Ibm.com. https://www.ibm.com/think/insights/cost-of-a-data-breach-healthcare-industry?utm_source=chatgpt.com
20. Elkamhi, R., Lee, J. S. H., & Salerno, M. (2023). Enhancing the Inverse Volatility Portfolio through Clustering. *The Journal of Financial Data Science*, *6*(1), 43–60. https://doi.org/10.3905/jfds.2023.1.145
21. Elmgren, K., Sett, G., & Smith, E. (2024). *Considerations on AI Model Red- Teaming and Standards*. https://insideaipolicy.com/sites/insideaipolicy.com/files/documents/2024/feb/ai02212024_2.pdf?utm_source=chatgpt.com
22. Fok, J. L., Zeng, Q., Chen, S., Fawkes, O., & Chen, H. (2025). *Foe for Fraud: Transferable Adversarial Attacks in Credit Card Fraud Detection*. ArXiv.org. https://arxiv.org/abs/2508.14699
23. Goyal, K., Garg, M., & Malik, S. (2025). Adoption of artificial intelligence-based credit risk assessment and fraud detection in the banking services: a hybrid approach (SEM-ANN). *Future Business Journal*, *11*(1). https://doi.org/10.1186/s43093-025-00464-3
24. Gupta, A. (2025). Red Teaming AI Systems for Security Validation. *International Journal of AI, BigData, Computational and Management Studies*, *6*(1). https://doi.org/10.63282/3050-9416.ijaibdcms-v6i1p112
25. Hafez, I. Y., Hafez, A. Y., Saleh, A., El-Mageed, A. A. A., & Abohany, A. A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of Big Data*, *12*(1). https://doi.org/10.1186/s40537-024-01048-8
26. Handa, R. (2025). Adversarial Simulation and Resilience Engineering for Enterprise AI Systems. *International Journal of Computational and Experimental Science and Engineering*, *11*(4). https://doi.org/10.22399/ijcesen.4413
27. Harris, R. (2025, October 9). *5 Key Takeaways from the GASA Global State of Scams 2025 Report*. Feedzai. https://www.feedzai.com/blog/gasa-global-state-of-scams-report/?utm_source=chatgpt.com
28. Herath, H. M. N. (2025). Advancing Machine Learning for Financial Fraud Detection: A Comprehensive Review of Algorithms, Challenges, and Future Directions. *ASEAN Journal of Economic and Economic Education*, *4*(1), 49–68. https://www.ejournal.bumipublikasinusantara.id/index.php/ajeee/article/view/608/0?utm_source=chatgpt.com
29. Idensohn, C., Flowerday, S., van der Schyff, K., & Chua, Y. T. (2026). Malicious insider threats in cybersecurity: A fraud triangle and Machiavellian perspective. *Computers in Human Behavior*, *174*, 108809. https://doi.org/10.1016/j.chb.2025.108809
30. Insider Risk. (2025). *Shadow AI and the Evolution of Insider Threats: A Critical Intelligence Assessment*. Insiderisk.io; Above Security. https://www.insiderisk.io/research/shadow-ai-insider-threats-2025?utm_source=chatgpt.com
31. Jabbar, M. S., Al-Azani, S., Alotaibi, A., & Ahmed, M. (2025). Red teaming large language models: A comprehensive review and critical analysis. *Information Processing & Management*, *62*(6), 104239. https://doi.org/10.1016/j.ipm.2025.104239
32. Kanagala, A. K. (2026). Autonomous Security Testing for AI Systems: Evaluating AI Red-Teaming Agents for Continuous Adversarial Assessment and Model Resilience. *Applied Sciences Research Periodicals*, *4*(01), 191–201. https://doi.org/10.63002/asrp.401.1330
33. Konatham, M., Uddandarao, D., & Kiran Vadlamani, R. (2024). Engineering Scalable AI Systems for Real-Time Payment Platforms. *Journal of Information Systems Engineering and Management*, *2024*(4). https://www.jisem-journal.com/download/33_Engineering%20Scalable%20AI%20Systems%20for%20Real-Time%20Payment%20Platforms.pdf?utm_source=chatgpt.com
34. Kothinti, K. G. (2025). AI-Powered Identity Verification & Risk Analysis: The Future of Fraud Prevention in Financial Services. *European Journal of Computer Science and Information Technology*, *13*(9), 23–55. https://doi.org/10.37745/ejcsit.2013/vol13n92355
35. Kumar, S. (2025). Real-Time Data Streaming: Transforming FinTech Through Modern Data Architectures. *European Journal of Computer Science and Information Technology*, *13*(18), 49–64. https://doi.org/10.37745/ejcsit.2013/vol13n184964
36. Majumder, R. Q. (2025). Data Driven-Machine Learning-Based Fraud Detection Models in FinTech Financial Transactions. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5359808
37. Maramreddy, Y. R., & Muppavaram, K. (2024). Detecting and Mitigating Data Poisoning Attacks in Machine Learning: A Weighted Average Approach. *Engineering, Technology & Applied Science Research*, *14*(4), 15505–15509. https://doi.org/10.48084/etasr.7591

38. Mashruwala, A. (2024). Distributed Systems in Fintech. *ArXiv (Cornell University)*. https://doi.org/10.13140/rg.2.2.12897.11368
39. Mayeke, N. R. (2025). Adapting Telemetry for Cyber-Physical Continuous Verification in Cameroon's Energy Sector. *Journal of Energy Research and Reviews*, *17*(12), 107–131. https://doi.org/10.9734/jenrr/2025/v17i12482
40. Metcalf, J., & Singh, R. (2023). Scaling Up Mischief: Red-Teaming AI and Distributing Governance. *Harvard Data Science Review*, *Special Issue 5*. https://doi.org/10.1162/99608f92.ff6335af
41. Mirishli, S. (2024). Ethical Implications of AI in Data Collection: Balancing Innovation with Privacy. *ANCIENT LAND*, *6*(8), 40–55. https://doi.org/10.36719/2706-6185/38/40-55
42. Obioha-Val, O. A., Lawal, T. I., Olaniyi, O. O., Gbadebo, M. O., & Olisa, A. O. (2025). Investigating the Feasibility and Risks of Leveraging Artificial Intelligence and Open Source Intelligence to Manage Predictive Cyber Threat Models. *Journal of Engineering Research and Reports*, *27*(2), 10–28. https://doi.org/10.9734/jerr/2025/v27i21390
43. Obrik-Uloho, E. P., Gbadebo, M. O., Afolabi, O. O., Joseph, S. A., Oladoyinbo, T. O., & Olaniyi, O. O. (2026). Shared Responsibility in Practice: Evaluating the Security–Usability Trade-Off and User Accountability in WhatsApp's Ecosystem. *Asian Journal of Research in Computer Science*, *19*(1), 81–105. https://doi.org/10.9734/ajrcos/2026/v19i1807
44. Odeyinka, T. E., Ifeoma Ejoh, C., Abdulmalik, A. A., Salami, I. A., & Ogunmolu, A. M. (2026). Bridging AI-automated Governance, Adaptive Certification, Behavioral Authentication, and AI-agent Risk Monitoring in Zero-trust Digital Infrastructures. *Journal of Engineering Research and Reports*, *28*(1), 371–387. https://doi.org/10.9734/jerr/2026/v28i11783
45. Ogunmolu, A. M., Abba, S. S., Olaniyi, O. M., Odeyinka, T. E., & Salami, I. A. (2026). Adaptive Cognitive Profiling for Executive AI Agents Amid Emerging AI Impersonation Threats. *Journal of Engineering Research and Reports*, *28*(1), 388–405. https://doi.org/10.9734/jerr/2026/v28i11784
46. Ogunmolu, A. M., Aroh, I. S., Henry, O., Adeyinka, P. D., & Olutimehin, A. T. (2026). AI-Driven Observability for Managing Security Complexity in Cloud-native Microservices and Containerized Environments. *Journal of Engineering Research and Reports*, *28*(2), 1–17. https://doi.org/10.9734/jerr/2026/v28i21786
47. Ogunmolu, A. M., Obrik-Uloho, E. P., Olaniyi, O. O., Arigbabu, A. T., & Bamigbade, O. (2025). Cyber Risk Spillovers in Interconnected Financial Ecosystems: Evidence from Traditional Banks and DeFi Oracles. *Journal of Engineering Research and Reports*, *27*(7), 106–126. https://doi.org/10.9734/jerr/2025/v27i71565
48. Olaniyi, O. M., Adebiyi, O. O., Ejoh, C. I., Afolabi, O. O., & Ejiofor, V. O. (2026). Conversational AI-Powered Fraud Prevention in Augmented Reality E-Commerce: A Natural Language Processing Framework for Real-time Transaction Security. *Asian Journal of Research in Computer Science*, *19*(1), 255–270. https://doi.org/10.9734/ajrcos/2026/v19i1817
49. Olaniyi, O. M., Aroh, I. S., Henry, O., Metibemu, O. C., & Akinola, O. I. (2026). Graph Neural Networks for Multi-Layered Financial Crime Network Detection: An Explainable AI Framework for Anti-Money Laundering. *Journal of Engineering Research and Reports*, *28*(2), 18–36. https://doi.org/10.9734/jerr/2026/v28i21787
50. Olisa, A. O., Gbadebo, M. O., Mayeke, N. R., Oladoyinbo, T. O., & Oluwapamilerin Kolo, F. H. (2026). Quantum-Resistant Cryptographic Protocols for CBDC Interoperability: A Cross-Border Settlement Security Framework. *Engineering and Technology Journal*, *11*(01). https://doi.org/10.47191/etj/v11i01.35
51. Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2). https://doi.org/10.1016/j.jsis.2024.101885
52. Pedarla, H. K. (2025). Generative AI for Network Attack Simulation. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *6*(4). https://doi.org/10.63282/3050-9262.ijaidsml-v6i4p109
53. Pöhler, L. D., Diepold, K., & Wallach, W. (2024). *A Practical Multilevel Governance Framework for Autonomous and Intelligent Systems*. ArXiv.org. https://arxiv.org/abs/2404.13719
54. Rashid, Z. U., Gurung, D., Gupta, S. R., & Rath, S. (2026). *Lifecycle-Integrated Security for AI-Cloud Convergence in Cyber-Physical Infrastructure*. ArXiv.org. https://arxiv.org/abs/2602.23397
55. Ridzuan, N. N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the Financial Sector: The Line between Innovation, Regulation and Ethical Responsibility. *Information*, *15*(8), 432. https://www.mdpi.com/2078-2489/15/8/432
56. Saeri, A. K., George, S. L., Graham, J., Lacarriere, C. D., Slattery, P., Noetel, M., & Thompson, N.

(2025). *Mapping AI Risk Mitigations: Evidence Scan and Preliminary AI Risk Mitigation Taxonomy*. ArXiv.org. https://arxiv.org/abs/2512.11931

57. Sampathkumar, V., & Veeras, K. (2025). Adversarial Attacks on FinTech AI Models: Threats and Mitigation Techniques. *2025 IEEE International Carnahan Conference on Security Technology (ICCST)*, 1–7. https://doi.org/10.1109/iccst63435.2025.11293895
58. Singh, J., Bharany, S., Rani, S., Rehman, A. U., Taye, B. M., Pant, R., & Kaur, U. (2025). A Systematic Review of Blockchain, AI, and Cloud Integration for Secure Digital Ecosystems. *the International Journal of Networked and Distributed Computing*, *13*(2). https://doi.org/10.1007/s44227-025-00072-1
59. Spelda, P., & Stritecky, V. (2025). Security practices in AI development. *AI & Society*, *40*. https://doi.org/10.1007/s00146-025-02247-4
60. Srivastava, S., Janardhan, K., & Jauhari, S. (2026). *A Systematic Review of Algorithmic Red Teaming Methodologies for Assurance and Security of AI Applications*. ArXiv.org. https://arxiv.org/abs/2602.21267
61. Sufficient, H. M., Mohammed, A. M., & Danjuma, B. (2025). Ethical Implications of AI-Driven Ethical Hacking: A Systematic Review and Governance Framework. *Journal of Cyber Security*, *7*(1), 239–253. https://doi.org/10.32604/jcs.2025.066312
62. Uula, M. M. (2024). Artificial Intelligence and Financial Regulation Issues. *Islamic Finance and Technology.*, *2*(1).
63. Villegas-Ch, W., Jaramillo-Alcázar, A., & Luján-Mora, S. (2024). Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW. *Big Data and Cognitive Computing*, *8*(1), 8–8. https://doi.org/10.3390/bdcc8010008
64. Viradia, V., Muthukrishnan, H., & Yadav, D. (2024). Insider Threats in Healthcare Application: Harnessing AI To Mitigate The Risks. *International Journal of Global Innovations and Solutions (IJGIS)*. https://doi.org/10.21428/e90189c8.603c06ac
65. Vuković, D. B., Dekpo-Adza, S., & Matović, S. (2025). AI integration in financial services: a systematic review of trends and regulatory challenges. *Humanities and Social Sciences Communications*, *12*(1). https://doi.org/10.1057/s41599-025-04850-8
66. Walter, M. J., Barrett, A., & Tam, K. (2024). A Red Teaming Framework for Securing AI in Maritime Autonomous Systems. *Applied Artificial Intelligence*, *38*(1). https://doi.org/10.1080/08839514.2024.2395750
67. Wisbey, O. (2024). *What is AI red teaming?* Search Enterprise AI; TechTarget. https://www.techtarget.com/searchenterpriseai/definition/AI-red-teaming?utm_source=chatgpt.com
68. WitnessAI. (2025, August 13). *AI Red Teaming: Securing AI Systems Through Adversarial Testing*. WitnessAI. https://witness.ai/blog/ai-red-teaming/?utm_
69. Yuan, J., Nöther, J., Jaques, N., & Radanović, G. (2026). *AgenticRed: Optimizing Agentic Systems for Automated Red-teaming*. ArXiv.org. https://arxiv.org/abs/2601.13518
70. Yulianto, S., Soewito, B., Gaol, F. L., & Kurniawan, A. (2024). Enhancing Cybersecurity Resilience through Advanced Red Teaming Exercises and MITRE ATT&CK Framework Integration: A Paradigm Shift in Cybersecurity Assessment. *Cyber Security and Applications*, *3*, 100077. https://doi.org/10.1016/j.csa.2024.100077
71. Zhao, Y. (2024). Improvement of the robustness of deep learning models against adversarial attacks. *Applied and Computational Engineering*, *75*(1), 285–289. https://doi.org/10.54254/2755-2721/75/20240556
72. Zhou, A., Wu, K., Pinto, F., Chen, Z., Zeng, Y., Yang, Y., Yang, S., Koyejo, S., Zou, J., & Li, B. (2025). *AutoRedTeamer: Autonomous Red Teaming with Lifelong Attack Integration*. ArXiv.org. https://arxiv.org/abs/2503.15754