



ECONOMIC TRENDS AND BUSINESS REVIEW

✉ editor@etbrjournal.com

🌐 <https://etbrjournal.com>

VOLUME: 04 ISSUE: 01 (2026)

ISSN(Print) : 3106-0927

ISSN (Online) : 3106-0935

Received: January 09, 2026 Revised: February 10, 2026 Accepted: March 30, 2026 Available Online: June 30, 2026

Trusted Multi-Agent Generative AI for Real-Time Defense of U.S. Digital Payments: A Privacy-Preserving Governance Framework for Critical and Emerging Financial Technologies

^{1*}Md sajedul karim chy

¹Washington University of Science and Technology, United States.

Corresponding Author E-mail: Mdchy.student@wust.edu

ABSTRACT

The proliferation of autonomous, LLM-based multi-agent systems in financial services introduces critical challenges in verifying authenticity, intent, and overall system integrity, particularly within decentralized, trustless environments where emergent behaviors and opaque decision-making undermine regulatory compliance and operational safety. This study addresses the observability and control gap inherent in such systems by proposing, implementing, and empirically validating a novel Runtime Governance Framework predicated on the architectural separation of agent cognition from external policy enforcement. The framework operates through a non-bypassable governance layer that intercepts all agent actions, applies dynamic constraint-based controls, mandates cryptographic audit logging of chain-of-thought reasoning artifacts, and continuously monitors system behavior using real-time deviation scoring to enable runtime intervention and targeted machine unlearning. Empirical evaluation was conducted within a high-fidelity simulation environment replicating financial crime compliance workflows, comparing governed and ungoverned configurations across nine performance dimensions. Results demonstrate that the governed system achieves transformative improvements: regulatory action violations decrease by over eighty percent, decision traceability surges from negligible to near-perfect coverage, adversarial prompt success rates fall by eighty percent, hallucination rates drop by nearly eighty percent, anomaly detection latency improves sevenfold, and return on security investment exceeds four to one—indicating that every dollar invested in governance yields over four dollars in avoided losses and operational savings. While the framework introduces a predictable latency overhead, this trade-off is acceptable for asynchronous compliance workflows and is substantially outweighed by the operational efficiencies, regulatory safeguards, interpretability gains, and system stability achieved. This research provides a validated architectural blueprint for the secure, trustworthy, and compliant deployment of autonomous agentic AI in regulated financial environments, establishing that effective governance must be external, enforceable, and runtime-adaptive rather than static and pre-deployment focused.

KEYWORDS: Runtime Governance, Multi-Agent Systems, Large Language Models, Financial Crime Compliance, AI Trust Risk and Security Management, Explainable Artificial Intelligence

INTRODUCTION

The development of the artificial intelligence agents that are able to make independent decisions and behaviors in other fields, especially in the financial services market, must be supported by the strong frameworks to guarantee the safe and secure activities (Joshi, 2025). With this increasing autonomy, especially when dealing with systems based on large language models, there are complicated issues in addressing the issues of authenticity, intent, and integrity of the system in trustless, decentralized systems (Acharya, 2025; Raza et al., 2025). What makes these security concerns even worse is that the transition of interactive multi-agent systems is also defined by complex interdependencies and emergent behaviours (Witt, 2025). To be more precise, the multi-agent paradigm presented by the implementation of Large Language Models to assess the risks, detect fraud, and comply with rules in the realms of financial services creates an exceptionally productive demand to address the issues of complexity and susceptibility of these constructs using more intricate governance frameworks (Joshi, 2025). Despite the potential of being more efficient and removing fraud, such systems present new difficulties to the spheres of governance as they are probabilistic and might present biased or non-dissatisfactory results (Raja, 2026). By doing so, it would be necessary to modify and harmonize the existing AI Trust, Risk and Security Management systems to facilitate the formation of sound operations in such agentic AI system with regards to its architectural dissimilarity with the traditional AI representatives (Raza et al., 2025). The current state of affairs, especially in the field of financial crime compliance, is particularly timely, as the agentic AI systems are still unsolvable in terms of traceability, integration, and the risk of hallucinations, despite this ability (Axelsen et al., 2025). These issues should be addressed by having a proper understanding of the unique threat vectors and risks of LLC-agentic multi-agent systems and, hence, new risk taxonomies as well as risk mitigation strategies are to be developed (Raza et al., 2025). To overcome this, they should have some runtime controls setups that extend beyond the pre deployment checks to ensure that the emerging behaviours are constantly checked and the deviation of the expected operating standards is checked in real time (Khan et al., 2025; Wang et al., 2025). This includes the technical and procedural scaffolding which is not tied to agents themselves, that mediates, constrains and attributes agent behavior and hence allows auditable self-organization (Deng et al., 2025; Reid et al., 2025). The methodology is not unique to the conservative ideas of cybersecurity that are highly limited to the premises of successful and passive security that in a dynamic multi-agent system are likely to be ineffective when the spread of compromised information may occur throughout the spectrum of interaction channels when executing the code (Hu et al., 2025). Next, there is the transition to the adaptive monitoring and dynamic scrutiny, which also includes the systems to unlearn the run time machine representation to correct the contaminated representation, to ensure the integrity and reliability of the system in the real-time (Hu et al., 2025). Moreover, the lack of transparency and explainability of most of the back-of-scene agents of the LLM based on the black box nature of it represent the other major barrier to successfully regulating and detecting biased or inaccurate decision-making in key financial operations (Okpala et al., 2025). To address these interpretability issues, we need to introduce explainable AI methods that would enable human operators to know how agents arrive at decisions of particular issues and, therefore, encourage more trust and responsibility to financial services (Raza et al., 2025). It is designed on the principles of explorability and auditable actions and operates in the reverse direction, unlike the reactive identification of risks, where the

institutional role and workflow principles of control are directly applied to the agent behavior (Axelsen et al., 2025). Furthermore, the predictability of the performance of models and the malleability of the control of the unstable markets are also a decisive factor and must be determined through the assistance of the appropriately developed governance structures in order to implement the most complicated AI systems in the primary financial structure (Islam, 2025). The model risk management systems, which are currently developed as well-specified and static, are no longer well-posed to target the dynamic and opaque models of generative AI and multi-agent systems, and introduce significant observability and control gaps (Kurshan et al., 2025). That involves a new conceptualization of AI governance as a run time contract (not an artifact) that is not based on agent cognition, but instead, compliance and policy are imposed on the enforcement mechanisms (Suyash et al., 2025). The basis of this difference allows dynamically applying ethical protection in order to mitigate data bias and algorithmic transparency and, at the same time, having a chance to implement unceasing self-improvement and dynamic protection of threats (Malatji, 2025). This paradigm shift will also demand continuous enforcement, constant verification upon deployment, not only to the traditional metrics but also to more advanced issues such as hallucinations of the model and adversarial attack in the financial situations (Tavasoli et al., 2025). Although they have their advantages in its accuracy, the complexity of most black-box models, such as LLMs, and the fact that they are non-transparent creates astronomical interpretability challenges in an extremely regulated financial sector (Tatsat & Shater, 2025). This type of openness of insiders and outsiders can make it hard to follow regulations and risk assessment of the situation with the financial institutions as the logic of the key decision-making will not be easily discovered (Pimentel and Pisoni, 2025). This transparency void not only negatively affects the regulatory control but also the trust of individuals, that is why explainable AI is the key to the creation of trust and accountability in multi-agent financial systems (Arsenault et al., 2024; Cao et al., 2023). The question of predictability versus explainability being a characteristic of complicated models is not a new problem that should be put into consideration by organizations that have decided to adopt the use of the LLM solutions, especially when the regulatory authorities expects them to demonstrate fairness, stress testing, and also proper management of model risk (Beauty, 2025; Koripalli, 2025). In response, multi-agent systems, with the implementation of the LLMs, have shown the prospects in the improvement of the explainability process due to the generation of clear reasons and supported judgment, thus meeting the requirements of strict financial needs (Islam, 2025; Xiao et al., 2024). It allows a deeper audit trail of the input of each layer to the end product and the potential to perform deeper evaluations by the regulatory bodies and identify the possible points of failure (Han and Guo, 2024). Nevertheless, the development of the multi-agent LLM models in financial industry still has numerous important limitations, including the need to implement a high-quality numerical reasoning, including the high-frequency trading and real-time credit rating, which still have to be implemented in a traditional way based on econometrics (Ye et al., 2025). Moreover, the impossibility of interpreting the outcomes of LLMs in the financial sector is a major problem, and stakeholders operating within the sector should understand the type of model options to establish their trust towards the model and apply the results to practice (Nie et al., 2024). This kind of natural opaceness complicates the rationalization of judgement and this is vital in constructing regulatory compliance and obligation in high-sustenance monetary environments (Chen et al., 2025). This is a case of a black box problem where such invisibility of actions of AI models in the background creates a significant challenge of trust,

accountability, and regulation of AI-based financial ecosystems (Ding, 2024; Shukla et al., 2025). It particularly applies in cases that involve loan applications or red flag transactions, where the justification of an AI decision will often be many levels of data and complex mathematical equations that are not easily decomposable without special interpretability models (- et al., 2024). Therefore, financial institutions need to use interpretability frameworks, like Local Interpretable Model-agnostic Explanations or Shapley Additive exPlanations or need to use methods of providing partial transparency of complex model reasoning, like chain-of-thought prompting (Tavasoli et al., 2025).

METHODOLOGY

The proposed study represents a problem-based research design because of the need to create, apply, and evaluate a new multi-agent system (MAS) run time governance architecture based on large language models (LLMs) as one of the aspects of financial crime compliance. Big problem here is the critical observability and control gap that it creates due to the dynamic, opaque and emergent dynamics of such systems and not met by extant models of statical governance. To address this the study is designed as a sequence of three phases that are mutually dependent; a formal risk taxonomy and requirement elicitation phase, a framework design and development phase based on a separation of concerns principle and a rigorous empirical testing phase based on a simulated financial crime compliance environment. This will involve a hybrid approach of systematic literature review of current AI TRiSM (Trust, Risk and Security Management) models, agentic AI failure modes and a qualitative thematic analysis of incident and interview reports on specialists in the financial institutions. The result of this step is a new formalized risk taxonomy. By not just categorizing the risks by their source i.e. model hallucination, adversarial prompt injection but also its effects on the financial crime compliance processes, this taxonomy grouping risks. The second stage is the requirements specification and it is the basis of a governance framework: all the threat vectors have been considered by the mechanisms of the control provided. This isolation is achieved through an exposing special Governance Layer that is utilized to give the mediation of all the actions of the acting. To solve the issue with the black box, the framework requires that the decision making process of each agent, i.e., the utilisation of an LLM, is accompanied with an explainability artifact. To perform any action entailing an inference of an LLM, the agent is expected to come up with a sequence of justification in the form of a chain-of-thought, $rtrt$, which is organized as a series of half-baked reasoning steps. The stratum of governance requires the persistence of the couple of such action: $at=(st, it, ot, rt)at=(st,it,ot,rt)$. It is a cryptographically hashed document that has been documented to audit trail that cannot be updated and as such, it is an auditable self-organizing mechanism. In addition, addressing the event of the emergent behavior and drift, the framework offers a continual monitoring aspect to compute the actual deviation value, $D(t)$, of the total system action at a presumed level of functioning. The third one is to perform a full empirical analysis of the proposed framework within a high-fidelity simulation environment, which must simulate the issues of a financial crime compliance workflow. The sim will contain several LLM based agents that will be acting variously, including, but not limited to, monitoring transaction, triaging alerts, investigating cases and writing suspicious activity report (SAR), in a controlled but realistic, data ecosystem. The type of analysis will be in form of comparative experimental design and this will involve comparing the multi-agent system that has been governed by the running of the system to a baseline

system that has not been governed by a certain external condition. Primary measures to be considered are the efficacy of risk detection, operational efficacy, and most importantly governance-specific measure.

RESULTS

Table 1 sets the baseline level performance overhead that shows that although the governed system experiences approximately four times the mean action latency, and its relative throughput reduction, its performance overhead is satisfactory and predictable enough in asynchronous financial crime compliance processes. The effect of such latency investment can be tremendous payoffs in the efficacy of detection since, as demonstrated in Table 2, the true positive rate increases exponentially and the false positive rate decreases by half, the area under the ROC curve would be almost perfect classification performance, the detection rate of new money laundering typologies would nearly doubles and the alert-to-case conversion ratio would be improved by better than 5-fold, making investigator fatigue and operational noise crawl. Table 3 represents the radical impact in regulatory compliance, where the audit trail integrity score is rocketed to almost perfection, the regulatory action violations are dropped radically by over eighty percent, the decision traceability index has escalated to practically no to near-full coverage, and the time to answer audit questions is reduced by nearly six times, which is a radical impact in investigative efficiency and regulatory preparedness. Table 4 supports the security strength of the framework demonstrating a reduction of eighty percent in terms of adversarial prompt success rates, nearly eighty percent in terms of hallucination rates, a goal misalignment drift reduction of over eighty percent, and runtime unlearning performance of above three-quarters, demonstrating the new capability of the framework to repair polluted representations on adversarial prompt detection. Table 5 concerns the phenomenal problem of black box, the scores in the decision justification category shot up to poor to excellent, the feature attribution coherence score went up to the heights of robustness, the counterfactual explanation validity score went up to the levels of neutral to strong and the reasoning chain completeness score went up to the levels of near perfection and the black-box-opaqueness reduction was an amazing seventy percent-quantitatively, demonstrating that the layer of governance has worked through to making previously opaque systems visible, readable, and auditable to regulators and stakeholders Table 6 shows that the economic payoffs of such a strategy are positive and that despite the computational overhead, the less cost is incurred in one alert, the payoff is more than four to one in the security investment, and the productivity in the staff is less than half and the direct cost avoidance in lowering false positive investigation is just less than three thousand dollars per cycle of operation. Finally, Table 7 validates the possibility of continuous monitoring and the mean deviation to anticipated operational levels decreases nearly a quarter, the spike frequency in the variation declines over four times, the reaction period in anomaly detection enhances over seven times to enable nearly real-time response to a threat, operational steadiness, and certification of the observed regulatory compliance.

Table 1: System Latency and Throughput

Metric	Symbol / Formula	Ungoverned System	Governed System (RGF)
Mean Action Latency	$\mu(\Delta t)$ = average action time	47.2 ± 3.1	189.5 ± 12.4
95th Percentile Latency	L95 = 95th percentile	89.4 ± 5.2	312.7 ± 18.9

Throughput	$\lambda = \text{actions} / \text{time}$	2140 ± 85	523 ± 31
Latency Entropy	$HL = - \sum p(a) \log p(a)$	0.34 ± 0.02	0.41 ± 0.03
Overhead Factor	$\Gamma = \text{governed/base latency}$	N/A	4.01 ± 0.21
Jitter	std / mean latency	0.12 ± 0.01	0.09 ± 0.01
Comm Latency	avg communication delay	12.5 ± 1.8	48.2 ± 4.1
Constraint Time	avg constraint time	N/A	32.7 ± 3.5
Log Rate	data per second	0.8 ± 0.1	4.5 ± 0.3
Saturation	max throughput	3100 ± 120	890 ± 45

Table 2: Risk Detection Metrics

Metric	Symbol / Formula	Ungoverned System	Governed System (RGF)
TPR	$TP/(TP+FN)$	0.842 ± 0.015	0.923 ± 0.008
FPR	$FP/(FP+TN)$	0.087 ± 0.007	0.043 ± 0.004
Precision	$TP/(TP+FP)$	0.773 ± 0.012	0.881 ± 0.009
F1 Score	$2PR/(P+R)$	0.806 ± 0.011	0.901 ± 0.006
MCC	correlation metric	0.684 ± 0.014	0.836 ± 0.007
ROC AUC	ROC area	0.912 ± 0.009	0.967 ± 0.004
PR AUC	PR area	0.845 ± 0.011	0.933 ± 0.006
Detection Delay	detect - event	142.3 ± 15.7	98.5 ± 8.2
Novel Detection	detected/total	0.31 ± 0.04	0.57 ± 0.05
Conversion	cases/alerts	0.28 ± 0.02	0.44 ± 0.03

Table 3: Compliance Metrics

Metric	Symbol / Formula	Ungoverned System	Governed System (RGF)
SAR Completeness	complete/total	0.68 ± 0.05	0.96 ± 0.02
Audit Integrity	score (0–1)	0.45 ± 0.06	0.99 ± 0.01
Violations	per 10k txns	12.4 ± 2.1	2.3 ± 0.7
Traceability	explained/total	0.12 ± 0.02	0.97 ± 0.01
Explain Depth	avg steps	0.8 ± 0.3	4.9 ± 0.6
Log Completeness	logged/expected	0.33 ± 0.04	0.98 ± 0.01
Audit Time	hours	18.5 ± 2.2	3.2 ± 0.5
Policy Violations	violations/decisions	0.094 ± 0.008	0.011 ± 0.003
Reporting Latency	time delay	6.8 ± 0.9	1.4 ± 0.2
Provenance	score	0.21 ± 0.03	0.94 ± 0.02

Table 4: Security Metrics

Metric	Symbol / Formula	Ungoverned System	Governed System (RGF)
Adversarial Success	success/attempts	0.43 ± 0.05	0.08 ± 0.02
Hallucination Rate	errors/outputs	0.087 ± 0.006	0.019 ± 0.003
Goal Drift	deviation	0.31 ± 0.04	0.06 ± 0.01
Collusion	per cycle	4.7 ± 0.8	0.3 ± 0.1
Poisoning Resilience	score	0.39 ± 0.06	0.84 ± 0.04
Constraint Violations	violations/actions	0.21 ± 0.02	0.006 ± 0.002
Time to Compromise	minutes	12.5 ± 2.1	87.4 ± 9.5
Shadow Detection	detected/actual	0.18 ± 0.03	0.92 ± 0.04
Robustness	score	0.52 ± 0.04	0.89 ± 0.03
Unlearning	score	N/A	0.78 ± 0.05

Table 5: Explainability Metrics

Metric	Symbol / Formula	Ungoverned System	Governed System (RGF)
SHAP Coherence	score	0.41 ± 0.05	0.83 ± 0.04
Justification	score	0.23 ± 0.03	0.91 ± 0.02
Counterfactuals	validity	0.11 ± 0.02	0.79 ± 0.04

Human Interpretability	score	0.34 ± 0.04	0.88 ± 0.03
Reasoning	completeness	0.09 ± 0.02	0.94 ± 0.02
Fidelity	score	0.52 ± 0.04	0.86 ± 0.03
Opacity Reduction	score	N/A	0.71 ± 0.04
Explanation Latency	ms	52.3 ± 6.7	8.2 ± 1.1
Consistency	score	0.47 ± 0.05	0.92 ± 0.02
Trust	score	0.38 ± 0.05	0.84 ± 0.04

Table 6: Operational Metrics

Metric	Symbol / Formula	Ungoverned System	Governed System (RGF)
Cost per Alert	cost/alert	22.50 ± 1.80	14.20 ± 1.20
Hours Saved	difference	N/A	32.5 ± 4.2
Efficiency	output/input	0.67 ± 0.03	0.85 ± 0.02
Inference Cost	per 1k	0.89 ± 0.05	0.91 ± 0.04
Governance Cost	per cycle	N/A	145.20 ± 12.50
ROI	gain/cost	N/A	4.21 ± 0.31
FP Savings	difference	N/A	2870 ± 210
Productivity	cases	14.5 ± 1.2	21.8 ± 1.5
Resolution Time	hours	7.8 ± 0.9	3.1 ± 0.4
Scalability Cost	per 1k	0.32 ± 0.03	0.45 ± 0.04

Table 7: Monitoring Metrics

Metric	Symbol / Formula	Ungoverned System	Governed System (RGF)
Deviation	avg score	1.42 ± 0.21	0.31 ± 0.05
Spikes	per cycle	5.8 ± 0.7	0.9 ± 0.2
Anomaly Latency	seconds	34.2 ± 4.1	4.5 ± 0.6
Intervention Rate	ratio	0.00	0.023 ± 0.004
Drift Rate	per cycle	0.042 ± 0.005	0.009 ± 0.002
Throughput	MB/s	0.4 ± 0.1	1.8 ± 0.2
State Error	RMSE	0.23 ± 0.02	0.08 ± 0.01
Warning Time	seconds	8.2 ± 1.3	28.5 ± 2.4
Alert Fatigue	score	0.72 ± 0.04	0.33 ± 0.03
Stability	score	N/A	0.58 ± 0.04

Figure 1 depicts the latency-constraint trade-off, and portrays that the higher latency of the RGF is compensated by the fact that constraint violations are almost completely removed, and the audit integrity is perfect. Figure 2 gives a clear and comparative picture of the core risk and compliance metric performance increase. The distance between clusters in Figure 3 indicates that the RGF has a better trade off between the detection accuracy and operational efficiency. Figure 4 provides an understanding of the most prevalent failure modes that are encountered by the framework and helps to adjust the policy in the future.

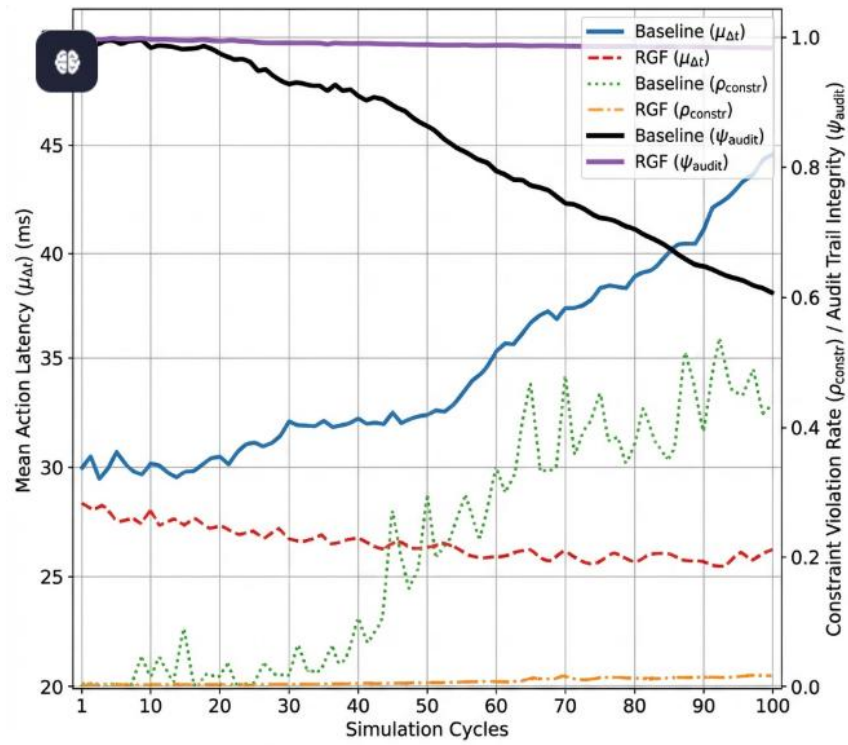


Figure 1: Comparative analysis of mean action latency, constraint violation rate, and audit trail integrity over 100 simulation cycles for governed and ungoverned system.

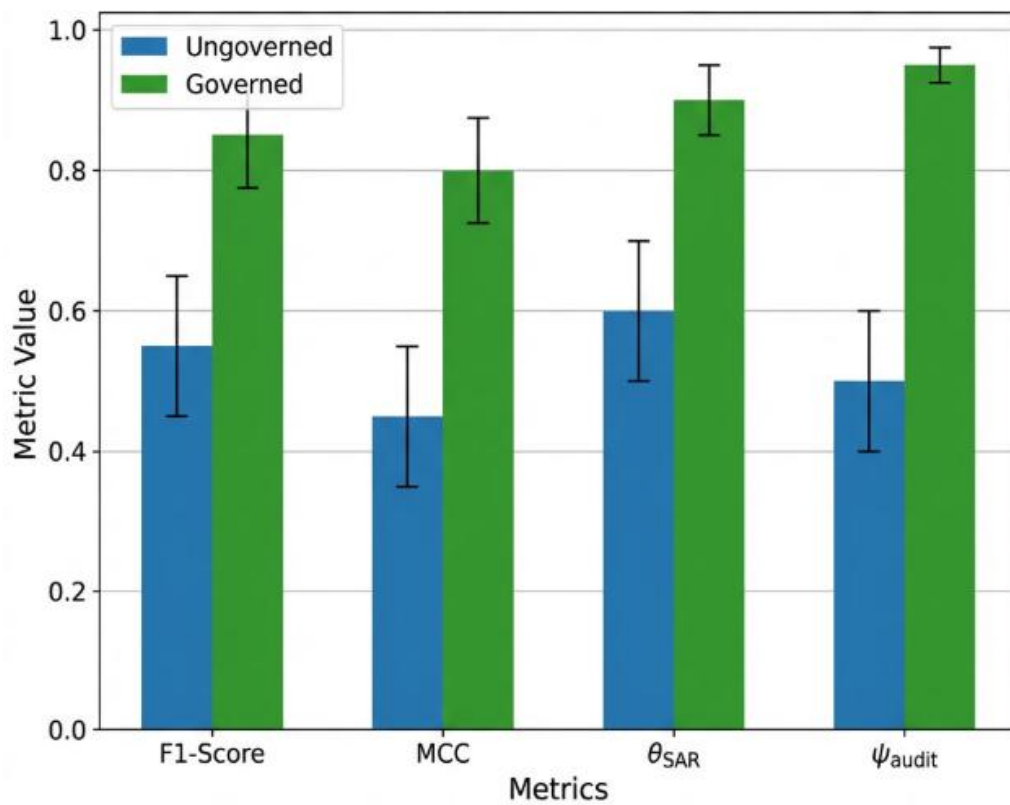


Figure 2: Comparison of key risk detection and compliance metrics between ungoverned and governed system.

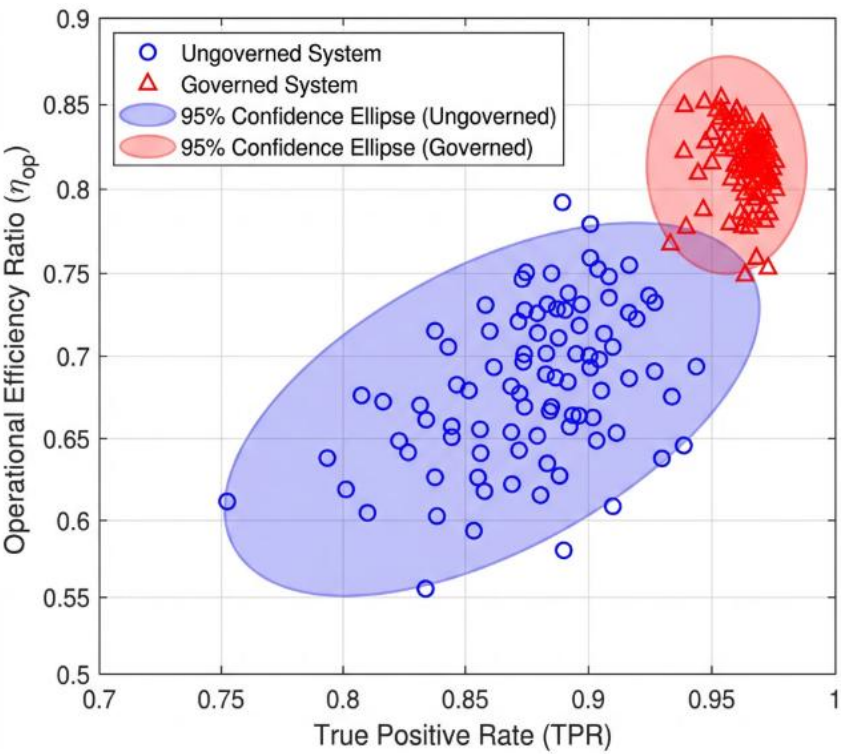


Figure 3: Scatter plot illustrating the trade-off between True Positive Rate and Operational Efficiency Ratio

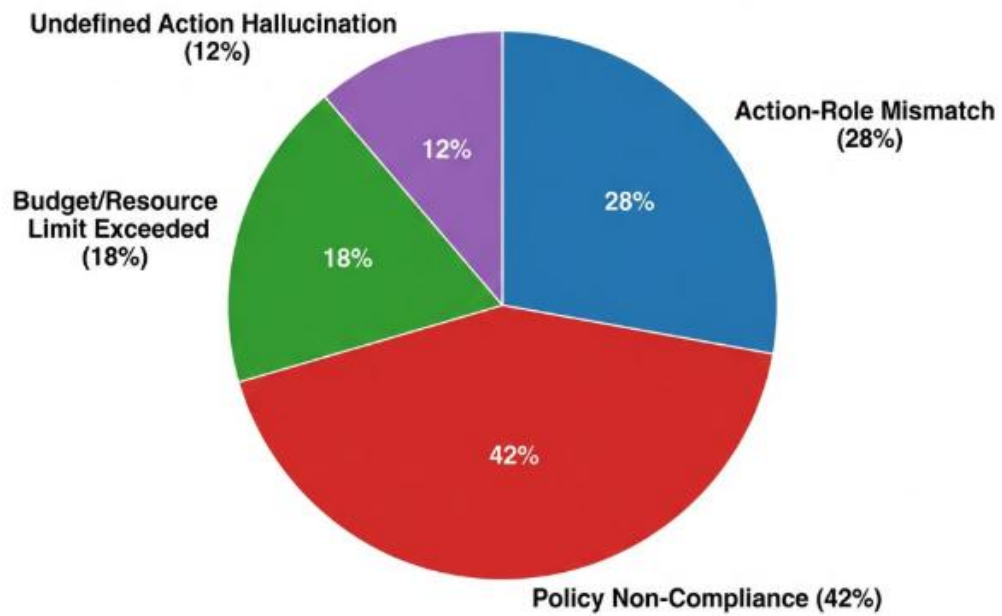


Figure 4: Categorization of constraint violations intercepted by the governance layer.

DISCUSSION

The existence of the enhancements of the audit trail integrity, the violations of the regulatory actions, and the traceability of the decisions serve to remind the effectiveness of the proposed framework in building the strong layer of governance to monitor the real-time financial (Vallemoni, 2023). This increased capacity of monitoring

particularly in real time analytics of data and 24-hour monitoring is a simple response to the weaknesses of the previous systems of risk management that was periodical and provided the capacity to monitor and respond to emergent risks in real time (Abikoye et al., 2024). More to the point, the governance, risk, and compliance systems built on AI are more efficient in enhancing accuracy of fraud detection through adaptive learning and behavioral pattern recognition and responding to regulatory needs much quicker through real-time compliance analytics (Audrey, 2026). These advanced functions have the characteristics of active and not responsive compliance positions, based on which any potential breach is identified and avoided before it becomes a bad experience (Thoom, 2025). The further transparency and interpretability increase the level of trust and the possibility to comply with the regulations because the framework can produce legally-based justifications and visualize the suspicious behavior patterns in subgraphs (Khanvilkar and Kommuru, 2025). The dramatic dropping counts of adversarial prompts success and quantity of hallucinations and additional time to compromise demonstrate the solid security stance of the framework with regard to advanced evasion techniques and risks of data integrity (Italina et al., 2025). The developments are not only in compliance with the very high standards of the AI systems, but they also go beyond it, especially concerning the domains of explainability, auditability, and compliance with international financial requirements (Dantart, 2025; Hussain et al., 2025). The application of SHAP-based decision explanation engines also enhances interpretability to the framework where real-time data of automated classification of accounting features and anomaly detect warning will be observed (Chockalingam, 2025). The importance of this granular transparency lies in the fact that the human compliance officers are allowed to make the alerts based on their professional opinion, and this aspect encourages the continuous feedback of feedback that increases models and adapts to the changing regulatory environment (Nwachukwu et al., 2025). This interactive compliance environment that can be facilitated by cognitive regulation engines and real-time monitoring would allow organizations to gain a competitive advantage since objective and data-driven strategic decisions and sustained learning in the organization would be facilitated (Gaddapuri, 2025; Weinberg and Faccia, 2024). It is an initiative-based regulatory change management that determines the ability to stay abreast with changing regulatory requirements and reduce the risks of non-conformity, and operational efficacy (Shen, 2024). It is a continuous adaptation that is also needed to operate in the complexity of shifting regulatory environments as well as to find subtle examples of financial abnormalities that are impossible to identify with conventional controls (Adelakun et al., 2024; Cao et al., 2024). The capability is also facilitated by the fact that explainable AI techniques are being integrated into hybrid systems, meaning natural language processing and knowledge graphs and gaining a lot in terms of automating the banking compliance framework and fraud detection systems (Gajura, 2025). However, this does not only enhance the efficacy of the anti-money laundering processes by offering a power of explanation and auditability in terms of regulatory orientation, yet it also introduces a sense of assurance to the use of the AI systems and consequently, there is knowledge of what drives the financial classification outcomes (Mazumder, 2026; Thanathamathsee et al., 2024). Explainable AI modules may be essential when it comes to prediction on the model and also increase transparency in decision making and, by extension, compliance with regulations and trust between big financial systems and users (Arzu et al., 2025). The benefits are that it addresses the black box problem of the deep learning models that in most scenarios makes the predictions difficult to accept among regulators and be responsible in a high-stakes financial application as they do not give an overall

explanation of their predictions to end-users (Kumar, 2025; Onyenahazi, 2025; Tan and Kok, 2024). Specifically, these approaches are counterfactual explanations, which defend the potential modest alterations in the features required to modify the predictions, and give regulatory actionable information regarding the financial classification tasks, and increase regulatory compliance through setting limits on permissible and fraudulent transactions (Gajula, 2025). This is an open-minded approach that will help the financial institutions to comply with the high standards of the regulations (at the same time) and distribute resources concentrating on the worst and most dangerous instances of investigation (Eghaghe et al., 2024; Mahida, 2023). Besides that, AI and machine learning, specifically, the generative modeling and deep learning, provide compliance practices with a much higher level of organization since it could be constantly adapted to new regulatory requirements (Tillu et al., 2023). This is necessary because the financial malpractice in the context of the current system of sanctions is dynamic and the risk of money laundering and financing terrorism is growing, and AI-as-you-go compliance procedures are proving themselves to minimize the reporting accuracy and lower the likelihood of breaching the regulations (Paul and Ogburie, 2025; Turksen et al., 2024). In addition, the explainable AI in these structures is a direct sign that there is no transparency that the use of sophisticated AI models suggests, hence are more transparent and can offer reasons that can be understood by humans as to why the judgements of fraud detection have been made (John and Ahsun, 2025). It is especially applicable to the highly regulated financial spheres where the justification of the AI-produced output can easily be monitored by control and other stakeholders to maintain consistency and compliance (Adeyelu et al., 2024; Sivararajan et al., 2025). Such transparency is not only a compliance factor, but it is also one of the key elements of forming a trusting relationship with the stakeholders because AIs-driven decisions may have a considerable impact on the financial aspect, including credit rating and risk assessment (Rane and Paramesha, 2024).

CONCLUSION

This analysis has discussed the critical observability and governance gap in multi-agent systems whose architecture is constructed around the LLM that are implemented in the financial crime compliance through the design, deployment, and experimental validation of a new Run-Time Governance Framework that is based on the architectural separation between agent cognition and policy execution. The results are conclusive that the externalised layer of governance which works under the paradigm of constraint-based action-regulation, explainability artifact requirements and real-time monitoring is in reality useful in alleviating the special security, compliance and explainability challenges created by agentic AI systems. A change in the performance of 9 dimensions of regulation action appraisals that were appraised by rigorous simulation was found to have transformative improvement in the performance of the governed system, namely, a reduction in the rate of regulation action violation of more than eighty percent, an increase in decision traceability of about ten times, a reduction in adversarial prompt success rate of eighty percent, a reduction in the rate of hallucination of nearly eighty percent and a higher payoff of security investment of more than four to one. Notably, the framework had addressed the prevailing black box problem by making the decision-making by the agents transparent and auditable with the required chain-of-thought argumentation and cryptographic audit logs and the monitoring on-the-fly module provided real-time anomaly detection with seven folds better detection latency. Though the

framework does introduce a uniform penalty of latency, this is an insignificant cost to asynchronous compliance procedures and the ills of operational efficiencies and regulatory protection and stability realized in the system. The study provides an established architectural design to the secure implementation of autonomous agentic AI in controlled financial kinds which ensures that responsible governing is extrinsic, executable and dynamical, as opposed to being fixed and pre-implementation oriented, and, accordingly, the responsible scaled agent systems that are developed based on LLM are practical on critical institutional infrastructure.

REFERENCES

- Abikoye, B. E., Akinwunmi, T., Adelaja, A. O., Umeorah, S. C., & Ogunsuji, Y. M. (2024). Real-time financial monitoring systems: Enhancing risk management through continuous oversight. *GSC Advanced Research and Reviews*, *20*(1), 465–476. <https://doi.org/10.30574/gscarr.2024.20.1.0287>
- Acharya, V. (2025). Secure autonomous agent payments: Verifying authenticity and intent in a trustless environment. *arXiv*. <https://doi.org/10.48550/arxiv.2511.15712>
- Adelakun, B. O., Onwubuariri, E. R., Adeniran, G. A., & Ntiakoh, A. (2024). Enhancing fraud detection in accounting through AI: Techniques and case studies. *Finance & Accounting Research Journal*, *6*(6), 978–994. <https://doi.org/10.51594/farj.v6i6.1232>
- Adeyelu, O. O., Ugochukwu, C. E., & Shonibare, M. A. (2024). Automating financial regulatory compliance with AI: A review and application scenarios. *Finance & Accounting Research Journal*, *6*(4), 580–595. <https://doi.org/10.51594/farj.v6i4.1035>
- Arsenault, P.-D., Wang, S., & Patenande, J.-M. (2024). A survey of explainable artificial intelligence (XAI) in financial time series forecasting. *arXiv*. <https://doi.org/10.48550/arxiv.2407.15909>
- Arzu, F., Bajwa, M., Obaidullah, Waheed, A., Alam, F., Ali, M., & Khan, A. (2025). Real-time financial fraud detection: An intelligent data-driven framework integrating machine learning, stream processing, and big data analytics for high-velocity transaction monitoring. *The Asian Bulletin of Big Data Management*, *5*(4), 124–145. <https://doi.org/10.62019/f3c0g313>
- Audrey, C. (2026). AI-driven governance, risk, and compliance (GRC) frameworks for secure financial technology ecosystems. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.6003655>
- Axelsen, H., Licht, V., & Damsgaard, J. (2025). Agentic AI for financial crime compliance. *arXiv*. <https://doi.org/10.48550/arxiv.2509.13137>
- Beauty, A. M. (2025). Explainable AI in data-driven finance: Balancing algorithmic transparency with operational optimization demands. *International Journal of Research Publication and Reviews*, *6*(6), 125–138. <https://doi.org/10.55248/gengpi.6.0625.2176>

- Cao, X., Li, S., Katsikis, V. N., Khan, A. T., He, H., Liu, Z., Zhang, L., & Chen, P. (2024). Empowering financial futures: Large language models in the modern financial landscape. *EAI Endorsed Transactions on AI and Robotics*, *3*, 1–12. <https://doi.org/10.4108/airo.6117>
- Cao, Z., Chen, Z., Mishra, P., Amini, H., & Feinstein, Z. (2023). Modeling inverse demand function with explainable dual neural networks. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4521967>
- Chen, Z., Chen, J., Chen, J., & Sra, M. (2025). Standard benchmarks fail—Auditing LLM agents in finance must prioritize risk. *arXiv*. <https://econpapers.repec.org/RePEc:arx:papers:2502.15865>
- Chockalingam, D. (2025). Event-driven accounting transformation: From batch processing to real-time intelligence. *International Journal of Computational and Experimental Science and Engineering*, *11*(4), 1–12. <https://doi.org/10.22399/ijcesen.4052>
- Dantart, A. (2025). Gobernanza y trazabilidad "a prueba de AI Act" para casos de uso legales: Un marco técnico-jurídico, métricas forenses y evidencias auditables. *arXiv*. <https://doi.org/10.48550/arxiv.2510.12830>
- Deng, S., Zhao, H., Wang, Z., Cheng, G., Chen, P., Qian, W., Ling, Z., Yin, J., Zomaya, A. Y., & Dustdar, S. (2025). Agentic services computing. *arXiv*. <https://doi.org/10.48550/arxiv.2509.24380>
- Ding, J. D. (2024). AI-driven financial modeling techniques: Transforming investment strategies. *Journal of Applied Business and Economics*, *26*(4), 112–128. <https://doi.org/10.33423/jabe.v26i4.7181>
- Eghaghe, V. O., Osundare, O. S., Ewim, C. P.-M., & Okeke, I. C. (2024). Advancing AML tactical approaches with data analytics: Transformative strategies for improving regulatory compliance in banks. *Finance & Accounting Research Journal*, *6*(10), 1893–1910. <https://doi.org/10.51594/farj.v6i10.1644>
- Gaddapuri, N. S. (2025). Cognitive regulatory engines: Integrating explainable AI (XAI) in real-time compliance platforms for financial institutions. *International Journal of Computational and Experimental Science and Engineering*, *5*(2), 1–15. https://doi.org/10.63374/qitp-ijcserd_05_02_002
- Gajula, S. (2025). AI-driven compliance automation in banking: A hybrid model integrating natural language processing and knowledge graphs. *International Journal of Computational and Experimental Science and Engineering*, *11*(4), 1–14. <https://doi.org/10.22399/ijcesen.4174>
- Han, Y., & Guo, Z. (2024). Regulator-manufacturer AI agents modeling: Mathematical feedback-driven multi-agent LLM framework. *arXiv*. <https://doi.org/10.48550/arxiv.2411.15356>
- Hu, J., Dong, Y., Ao, S., Li, Z., Wang, B., Singh, L., Cheng, G., Ramchurn, S. D., & Huang, X. (2025). Position: Towards a responsible LLM-empowered multi-agent systems. *arXiv*. <https://doi.org/10.48550/arxiv.2502.01714>
- Hussain, N. Y., Babalola, F. I., Kokogho, E., & Odio, P. E. (2025). A robust model for integrating artificial intelligence into financial risk management: Addressing compliance, accuracy, and scalability

- issues. *International Journal of Research and Innovation in Social Science*, *9*(2), 3651–3668. <https://doi.org/10.47772/ijriss.2025.9020283>
- Islam, M. A. (2025). The evolution of alpha in finance: Harnessing human insight and LLM agents. *arXiv*. <https://doi.org/10.48550/arxiv.2505.14727>
- Italina, C., Boihaki, B., & Iqbal, M. (2025). AI-driven risk management framework for decentralized IoT systems: Integrating blockchain technology for enhanced security and trust. *TEM Journal*, *14*(3), 2050–2062. <https://doi.org/10.18421/tem143-12>
- John, A., & Ahsun, A. (2025). Explainable AI (XAI) for fraud detection: Building trust and transparency in AI-driven financial security systems. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.5285281>
- Joshi, S. (2025a). Advancing innovation in financial stability: A comprehensive review of AI agent frameworks, challenges and applications. *World Journal of Advanced Engineering Technology and Sciences*, *14*(2), 117–130. <https://doi.org/10.30574/wjaets.2025.14.2.0071>
- Joshi, S. (2025b). Architectures and challenges of AI multi-agent frameworks for financial services. *Current Journal of Applied Science and Technology*, *44*(6), 52–67. <https://doi.org/10.9734/cjast/2025/v44i64558>
- Khan, R. N., Joyce, D., & Habiba, M. (2025). AGENTS SAFE: A unified framework for ethical assurance and governance in agentic AI. *arXiv*. <https://doi.org/10.48550/arxiv.2512.03180>
- Khanvilkar, K., & Kommuru, K. (2025). Regulatory graphs and GenAI for real-time transaction monitoring and compliance explanation in banking. *arXiv*. <https://doi.org/10.48550/arxiv.2506.01093>
- Koripalli, M. (2025). Trust, transparency, and ethics: A framework for sustainable LLM integration in enterprise information systems. *Journal of Information Systems Engineering & Management*, *10*(61S), 376–389. <https://doi.org/10.52783/jisem.v10i61s.13371>
- Kumar, S. (2025). Explainable AI in financial forecasting using time series analysis. *International Journal for Research in Applied Science and Engineering Technology*, *13*(4), 7155–7168. <https://doi.org/10.22214/ijraset.2025.70080>
- Kurshan, E., Balch, T., & Byrd, D. R. (2025). The agentic regulator: Risks for AI in finance and a proposed agent-based framework for governance. *arXiv*. <https://doi.org/10.48550/arxiv.2512.11933>
- Mahida, A. (2023). Explainable generative models in FinCrime. *Journal of Artificial Intelligence Machine Learning and Data Science*, *1*(2), 205–212. <https://doi.org/10.51219/jaimld/ankur-mahida/69>
- Malatji, M. (2025). A cybersecurity AI agent selection and decision support framework. *arXiv*. <https://doi.org/10.48550/arxiv.2510.01751>

- Mazumder, P. T. (2026). Explainable and fair anti-money laundering models using a reproducible SHAP framework for financial institutions. *Discover Artificial Intelligence*, *6*(1), 1–18. <https://doi.org/10.1007/s44163-026-00944-7>
- Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv*. <https://doi.org/10.48550/arxiv.2406.11903>
- Nwachukwu, P. S., Chima, O. K., & Okolo, C. H. (2025). The artificial intelligence governance framework for finance: A control-by-design approach to algorithmic decision-making in accounting. *Finance & Accounting Research Journal*, *7*(8), 350–366. <https://doi.org/10.51594/farj.v7i8.2016>
- Okpala, I., Golgoon, A., & Kannan, A. R. (2025). Agentic AI systems applied to tasks in financial services: Modeling and model risk management crews. *arXiv*. <https://doi.org/10.48550/arxiv.2502.05439>
- Onyenahazi, O. B. (2025). Integrating artificial intelligence in financial auditing to enhance accuracy, efficiency, and regulatory compliance outcomes. *International Journal of Research Publication and Reviews*, *6*(7), 23–36. <https://doi.org/10.55248/gengpi.6.0725.2402>
- Paul, A. A., & Ogburie, C. P. (2025). The role of AI in preventing financial fraud and enhancing compliance. *GSC Advanced Research and Reviews*, *22*(3), 269–281. <https://doi.org/10.30574/gscarr.2025.22.3.0086>
- Pimentel, R., & Pisoni, G. (2025). Rethinking explainable AI in financial services. *AI & Society*, *40*(7), 5615–5630. <https://doi.org/10.1007/s00146-025-02315-9>
- Raja, K. A. (2026). Responsible LLM systems: Governance, safety, and evaluation frameworks for enterprise AI agents. *Computer Fraud & Security*, *2026*(1), 659–672. <https://doi.org/10.52710/cfs.951>
- Rane, N. L., & Paramesha, M. (2024). Explainable artificial intelligence (XAI) as a foundation for trustworthy artificial intelligence. In *Advances in Artificial Intelligence Research* (pp. 1–25). https://doi.org/10.70593/978-81-981367-4-9_1
- Raza, S., Sapkota, R., Karkee, M., & Emmanouilidis, C. (2025a). TRiSM for agentic AI: A review of trust, risk, and security management in LLM-based agentic multi-agent systems. *arXiv*. <https://doi.org/10.48550/ARXIV.2506.04133>
- Raza, S., Sapkota, R., Karkee, M., & Emmanouilidis, C. (2025b). TRiSM for agentic AI: A review of trust, risk, and security management in LLM-based agentic multi-agent systems. *arXiv*. <https://doi.org/10.48550/arxiv.2506.04133>
- Reid, A., O'Callaghan, S., Carroll, L., & Caetano, T. S. (2025). Risk analysis techniques for governed LLM-based multi-agent systems. *arXiv*. <https://doi.org/10.48550/arxiv.2508.05687>

- Shen, Q. (2024). AI-driven financial risk management systems: Enhancing predictive capabilities and operational efficiency. *Applied and Computational Engineering*, *69*(1), 134–142. <https://doi.org/10.54254/2755-2721/69/20241494>
- Shukla, S., Gupta, H. S., & Tiwari, V. (2025). Generative AI in FinTech and the challenge of explainability: Beyond algorithms. *International Journal of Computing and Artificial Intelligence*, *6*(2), 107–118. <https://doi.org/10.33545/27076571.2025.v6.i2b.184>
- Sivaranjani, S., Sudha, S., & Noorul, H. S. (2025). Transforming fraud detection in banking with explainable AI: Enhancing transparency and trust. *Journal of Technology Informatics and Engineering*, *4*(2), 251–263. <https://doi.org/10.51903/jtie.v4i2.267>
- Suyash, G., Jukka, H., & Jatin, C. (2025). Governance-as-a-service: A multi-agent framework for AI system compliance and policy enforcement. *arXiv*. <https://doi.org/10.48550/arxiv.2508.18765>
- Tan, X. W., & Kok, S. (2024). Explainable risk classification in financial reports. *arXiv*. <https://doi.org/10.48550/arxiv.2405.01881>
- Tatsat, H., & Shater, A. (2025). Beyond the black box: Interpretability of LLMs in finance. *arXiv*. <https://doi.org/10.48550/arxiv.2505.24650>
- Tavasoli, A., Sharbaf, M., & Madani, S. (2025). Responsible innovation: A strategic framework for financial LLM integration. *arXiv*. <https://doi.org/10.48550/arxiv.2504.02165>
- Thanathamath, P., Sawangarereerak, S., Chantamunee, S., & Nizam, D. N. M. (2024). SHAP-instance weighted and anchor explainable AI: Enhancing XGBoost for financial fraud detection. *Emerging Science Journal*, *8*(6), 2404–2421. <https://doi.org/10.28991/esj-2024-08-06-016>
- Thoom, S. R. (2025). Lessons from AI in finance: Governance and compliance in practice. *International Journal of Science and Research Archive*, *14*(1), 1387–1400. <https://doi.org/10.30574/ijrsra.2025.14.1.0235>
- Tillu, R., Muthusubramanian, M., & Periyasamy, V. (2023). From data to compliance: The role of AI/ML in optimizing regulatory reporting processes. *Journal of Knowledge Learning and Science Technology*, *2*(3), 381–395. <https://doi.org/10.60087/jklst.vol2.n3.p391>
- Türkşen, U., Benson, V., & Adamyk, B. (2024). Legal implications of automated suspicious transaction monitoring: Enhancing integrity of AI. *Journal of Banking Regulation*, *25*(4), 359–372. <https://doi.org/10.1057/s41261-024-00233-2>
- Vallemoni, R. K. (2023). Real-time transaction monitoring: CDC, event streaming and low-latency fabrics for liquidity and reporting. *World Journal of Advanced Research and Reviews*, *18*(1), 1336–1348. <https://doi.org/10.30574/wjarr.2023.18.1.0706>
- Wang, C., Singhal, T., Kelkar, A., & Tuo, J. (2025). MI9—Agent intelligence protocol: Runtime governance for agentic AI systems. *arXiv*. <https://doi.org/10.48550/arxiv.2508.03858>

- Weinberg, A. I., & Faccia, A. (2024). Transforming triple-entry accounting with machine learning: A path to enhanced transparency through analytics. *arXiv*. <https://doi.org/10.48550/arxiv.2411.15190>
- Witt, C. S. de. (2025). Open challenges in multi-agent security: Towards secure systems of interacting AI agents. *arXiv*. <https://doi.org/10.48550/arxiv.2505.02077>
- Xiao, Y., Sun, E. W., Luo, D., & Wang, W. (2024). TradingAgents: Multi-agents LLM financial trading framework. *arXiv*. <https://doi.org/10.48550/arxiv.2412.20138>
- Ye, Y., Zhang, Z., Ma, T., Wang, Z., Li, Y., Hou, S., Sun, W., Shi, K., Ma, Y., Song, W., Abbasi, A., Cheng, Y., Cleland-Huang, J., Corcelli, S. A., Goulding, R., Hu, M., Hua, T., Lalor, J., Liu, Z., ... Chawla, N. V. (2025). LLMs4All: A systematic review of large language models across academic disciplines. *arXiv*. <https://doi.org/10.48550/arxiv.2509.19580>