http://www.moderntechno.de/index.php/meit/article/view/meit41-03-055
DOI: 10.30890/2567-5273.2025-41-03-055

УДК 004.89

MCP-ORIENTED GENERATIVE AI ASSISTANTS: TRUSTWORTHY AND POLICY-DRIVEN ARCHITECTURE FOR FINANCIAL SERVICES

MCP-ОРІЄНТОВАНІ ГЕНЕРАТИВНІ AI-АСИСТЕНТИ: НАДІЙНА ТА ПОЛІТИЧНО КЕРОВАНА АРХІТЕКТУРА ДЛЯ ФІНАНСОВИХ СЕРВІСІВ

Tsymbal A.S. / Цимбал А.С.
M.Sc. / магістр наук.
ORCID: 0009-0006-8786-8428
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
Kyiv. 37 Beresteysky ave. 03056,
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського» м. Київ, просп. Берестейський, 37. 03056

Abstract. *Financial institutions have accumulated vast corpora of regulatory policies, procedures, case notes, and operational documents, creating an overwhelming cognitive burden for compliance officers, risk managers, and operations teams. Traditional information retrieval systems fail to ensure factual grounding, auditability, and access control required in regulated environments. This paper introduces an MCP-oriented architecture for Generative AI assistants, integrating Retrieval-Augmented Generation (RAG) with the Model Context Protocol (MCP) as a unified framework for instrumentation, integration, and auditable execution.*

The proposed system enables governed agency through transparent logging, multi-layer guardrails (PII redaction, prompt-injection protection, response policies), and embedded Model Risk Management (MRM) processes. The evaluation methodology encompasses classical information retrieval metrics (nDCG, MRR, Recall@K), factual grounding metrics (Precise-Grounding@K), and operational indicators (Task Success Rate, Time-to-Resolution, First-Contact Resolution), all measured under strict p95/p99 latency SLOs.

Across three role-based scenarios—Compliance Copilot (AML/KYC, sanctions), Operations Copilot (payment incidents, contact center), and Data-Enrichment Copilot (data normalization, entity resolution)—the MCP-based approach demonstrates a substantial increase in factual accuracy, faster query resolution, and complete elimination of PII incidents under properly configured guardrails.

The study delivers a reproducible framework for building GenAI assistants in regulated industries, combining semantic retrieval, MCP-driven instrumentation, and formalized model risk management without compromising accuracy, latency, or auditability.

Keywords: *generative assistants; large language models (LLM); Model Context Protocol (MCP); Retrieval-Augmented Generation (RAG); semantic search; compliance; AML/KYC; operator copilot; guardrails; Model Risk Management (MRM); PII; latency SLO; p95/p99.*

Introduction The financial services sector operates within one of the most complex and tightly regulated environments in the global economy. Every process—from sanctions screening to Know Your Customer (KYC) verification—must satisfy strict transparency and auditability requirements. As data volumes and regulatory frameworks expand, compliance officers and risk managers face rising cognitive and operational burdens. Fragmented documentation and outdated procedures further hinder decision accuracy across institutions.

Recent progress in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) has created new opportunities for intelligent information retrieval in regulated domains [2][4]. However, deploying these models in finance remains limited by systemic risks: unverifiable reasoning, hallucinated outputs, personal data exposure (PII), and strict latency SLOs at p95/p99 thresholds [5][10][12]. Earlier attempts at compliance automation and AML reporting showed that, without clear governance, LLMs become "black boxes" incapable of ensuring reproducibility or traceable provenance [6][11].

The Model Context Protocol (MCP) addresses this challenge by introducing structured communication between models and enterprise tools. MCP formalizes metadata, validation schemas, access policies, and audit logs—transforming ungoverned model actions into transparent, rule-bound operations [1][9].

This paper presents an MCP-oriented architecture integrating hybrid RAG retrieval (BM25 + vector embeddings), reranking, and mandatory in-text citations. All tool interactions—such as case management, LOS, IDP, and KMS—are executed through the MCP layer to ensure end-to-end traceability [1][5]. Multi-layer guardrails (PII redaction, prompt-injection protection, and policy-based output filtering) further prevent both unintentional and adversarial model errors [8][9][10].

In addition, Model Risk Management (MRM) processes are embedded into the MCP lifecycle, linking model versions, embeddings, and access policies under unified observability [7]. Prior studies confirm that layered security and PII redaction reduce operational risk without degrading accuracy [8][10].

By combining semantic retrieval, contextual governance, and formalized risk control, the proposed MCP-RAG framework transforms Generative AI from an experimental assistant into a trustworthy, auditable, and regulation-ready system for financial services.

Main text

Problem Context and Real-World Motivation The financial sector operates within one of the most complex and highly regulated environments across all industries. Compliance officers, risk analysts, and operations teams must continuously process and cross-reference thousands of pages of policies, procedures, sanction lists, and case documents. This leads to a well-known phenomenon of *information friction*—when the cost of finding, validating, and applying the correct information exceeds the value of the decision itself [6][11][12]. Even with modern document-management systems, most compliance workflows remain manual, dependent on human interpretation, and burdened by multi-layered approval processes.

Large Language Models (LLMs) offer the potential to remove this bottleneck by providing semantic access to institutional knowledge. However, their deployment in financial environments is constrained by several systemic risks, including the lack of verifiable citations, the danger of hallucinated outputs, and the possible exposure of personally identifiable information (PII) during generation [5][9]. In production systems, implementation is further challenged by strict latency service-level objectives (SLOs) at p95/p99 thresholds, where even a few hundred milliseconds of additional delay can impact customer experience or automated decision pipelines.

In high-risk domains such as banking and compliance, trust depends on traceable reasoning. Every automated recommendation must be grounded in verifiable sources. Yet, most current Retrieval-Augmented Generation (RAG) systems lack unified mechanisms for auditing, access control, and safe tool orchestration. This gap motivates the transition toward architectures built around the Model Context Protocol (MCP), where LLM assistants act not as isolated chatbots but as *auditable operators* within enterprise ecosystems [1][5]. MCP provides a structured layer for describing tools, enforcing access policies, logging requests, and guaranteeing transparency and reproducibility across the AI workflow.

In practice, MCP-based assistants can transform daily financial operations. A compliance officer investigating a potential sanctions violation can instantly retrieve verified policy fragments and related case notes instead of manually searching across multiple repositories. An operations agent handling a chargeback incident can draft a compliant escalation report with automatic PII redaction and in-text citations. These examples illustrate not only higher operational efficiency but also measurable risk reduction and institutional accountability — core priorities for regulators and auditors.

Table 1 — Compliance reality: information overload in financial operations.

Stage	Manual Process (Current)	With MCP-Based Assistant	Expected Improvement	Stage
Regulatory clause search	10–15 min	1–2 sec	−99% time	Regulatory clause search
Policy interpretation errors	Frequent (human ambiguity)	Minimal (with citations)	−85% errors	Policy interpretation errors
Document drafting	30–60 min	5–8 min	−80% time	Document drafting
PII incidents per 1k documents	2.1	0	−100%	PII incidents per 1k documents

Together, these observations demonstrate that the core bottleneck in modern financial operations is not a lack of information but a lack of accessible, verifiable knowledge. Addressing this through governable, reproducible AI systems is essential for scaling compliance and risk management in an environment where regulatory velocity exceeds human capacity. The MCP-oriented architecture provides a foundation that combines accuracy, latency control, and auditability—ensuring that Generative AI systems in financial services remain both reliable and accountable.

From RAG to MCP: Towards Trustworthy GenAI in Finance

The evolution of Retrieval-Augmented Generation (RAG) architectures has transformed enterprise information access by combining symbolic retrieval with neural reasoning. In the financial domain, however, this transformation remains incomplete. Traditional RAG implementations successfully enhance factual grounding and reduce hallucination frequency, but they do not guarantee auditability or data governance—two prerequisites for regulatory compliance [1][2][4]. Most current systems still rely on ad-hoc pipelines without formalized access control, latency guarantees, or audit trails. Consequently, while RAG improves what is retrieved, it remains insufficiently transparent in how and why a response is generated.

In high-stakes environments such as Anti-Money Laundering (AML), Know Your Customer (KYC), and sanctions monitoring, this lack of interpretability translates directly into operational and legal risk [6]. A system capable of summarizing or reasoning over sensitive data must also provide traceable provenance of every claim, enforce user-specific permissions, and apply consistent PII protection across multiple stages of retrieval and generation [9][10]. The Model Context Protocol (MCP) addresses these gaps by defining a standardized framework for orchestrating, validating, and monitoring model-driven interactions across distributed financial ecosystems [1][5].

At the architectural level, MCP serves as a meta-layer over RAG pipelines. It introduces structured definitions for tools, schemas, policies, and access conditions. Each model or retriever operates within an auditable context that enforces timeouts, retries, and validation, ensuring deterministic behavior under latency SLOs (p95/p99). By coupling RAG with MCP, institutions can maintain semantic accuracy and compliance integrity simultaneously—two objectives previously at odds in AI adoption for regulated domains.

Formally, the hybrid retrieval function within an MCP-oriented system can be expressed as a linear combination of lexical and semantic relevance components:

$$Score_{hybrid}(q, d) = \lambda \cdot BM25(q, d) + (1 - \lambda) \cdot Sim_{vector}\big(E(q), E(d)\big) \qquad (1)$$

where λ controls the trade-off between keyword precision and semantic recall;$E(q)$ and $E(d)$ denote embedding functions for the query and document, respectively; and Sim_{vector} measures cosine or inner-product similarity in the latent space [2][4].

In practical deployment, MCP enforces policy-driven selection of tools and model versions, ensuring consistent audit logging and repeatable outputs. Every invocation—retrieval, rerank, generation, or tool use—is captured with contextual metadata: tenant ID, model hash, latency, and result status. This allows organizations to trace a single recommendation back through its retrieval chain, validation policies, and source documents—fulfilling the requirements of model risk management (MRM) [7][12].

The transition from standalone RAG systems to MCP-based architectures marks a fundamental step toward trustworthy Generative AI in financial services. It operationalizes the principle of explainable autonomy—where AI assistants remain capable of semantic reasoning while still obeying strict institutional and regulatory guardrails. By embedding auditability, latency control, and role-based orchestration into the AI stack, MCP enables the next generation of GenAI systems to function not merely as conversational engines but as compliant, traceable, and accountable components of digital finance.

Table 2 — Transition from RAG to MCP in financial AI systems.

Dimension	Traditional RAG	MCP-Oriented GenAI	Expected Impact	Dimension
Access control	Ad-hoc or static	Policy-driven, role-aware	+Governance	Access control
Citation and grounding	Optional or heuristic	Mandatory in-text citation	+Factual accuracy	Citation and grounding
Latency SLOs	Uncontrolled	p95/p99 monitored per stage	+Predictability	Latency SLOs
Auditability	Partial logging	Full context traceability	+Compliance	Auditability

Empirical Evaluation and Results

This section presents a quantitative evaluation of the proposed MCP-RAG architecture across retrieval, grounding, and operational performance metrics. The proposed MCP-oriented Generative AI system follows a layered architecture that integrates retrieval, generation, and governance into a unified workflow. At the core lies the Model Context Protocol (MCP), which orchestrates interactions between retrieval modules, language models, and enterprise tools through structured, auditable interfaces. Each module—retriever, reranker, generator, and policy engine—operates within a controlled execution context, enforcing standardized inputs, timeouts, and logging.

The evaluation was conducted on a dataset of 1,200 anonymized compliance queries derived from production AML/KYC and operations case logs (2022–2024) at a regulated financial institution. Due to regulatory constraints, the dataset contains sensitive policy information and cannot be publicly released. Each query was linked to

5–10 labeled document passages, reviewed by two domain experts with Cohen's κ = 0.82 inter-annotator agreement.

Table 3 — Comparative Results on Compliance Query Dataset.

Metric	BM25	Vanilla RAG	MCP-RAG (Proposed)	Δ vs RAG	OpenAI Assistants	MCP-RAG (ours)
nDCG@5	0.67	0.74	0.86	+16 %	—	—
nDCG@10	0.71	0.78	0.88	+13 %	0.81	0.88
MRR	0.69	0.76	0.89	+17 %	—	—
PG@5	0.62	0.70	0.91	+30 %	0.85	0.91
TSR	0.68	0.79	0.92	+16 %	—	—
FCR	0.63	0.74	0.89	+20 %	—	—
p95 Latency (ms)	820	910	960	+5.5 % overhead	1240	960

Retrieval and grounding performance were measured via nDCG@K, MRR, Recall@K, and Precise-Grounding@K; operational effectiveness was captured through Task Success Rate (TSR) and First-Contact Resolution (FCR) obtained from simulated workflows under the MCP orchestration layer.

The workflow begins when a user or agent issues a compliance-related query. The retrieval layer performs a hybrid search across vector and lexical indexes, followed by a lightweight reranking step to refine contextual relevance. The generation layer then synthesizes a grounded response, embedding in-text citations derived from the retrieved sources. Finally, the MCP governance layer validates outputs against response policies, applies PII redaction if required, and records full trace metadata, including source identifiers and latency metrics.

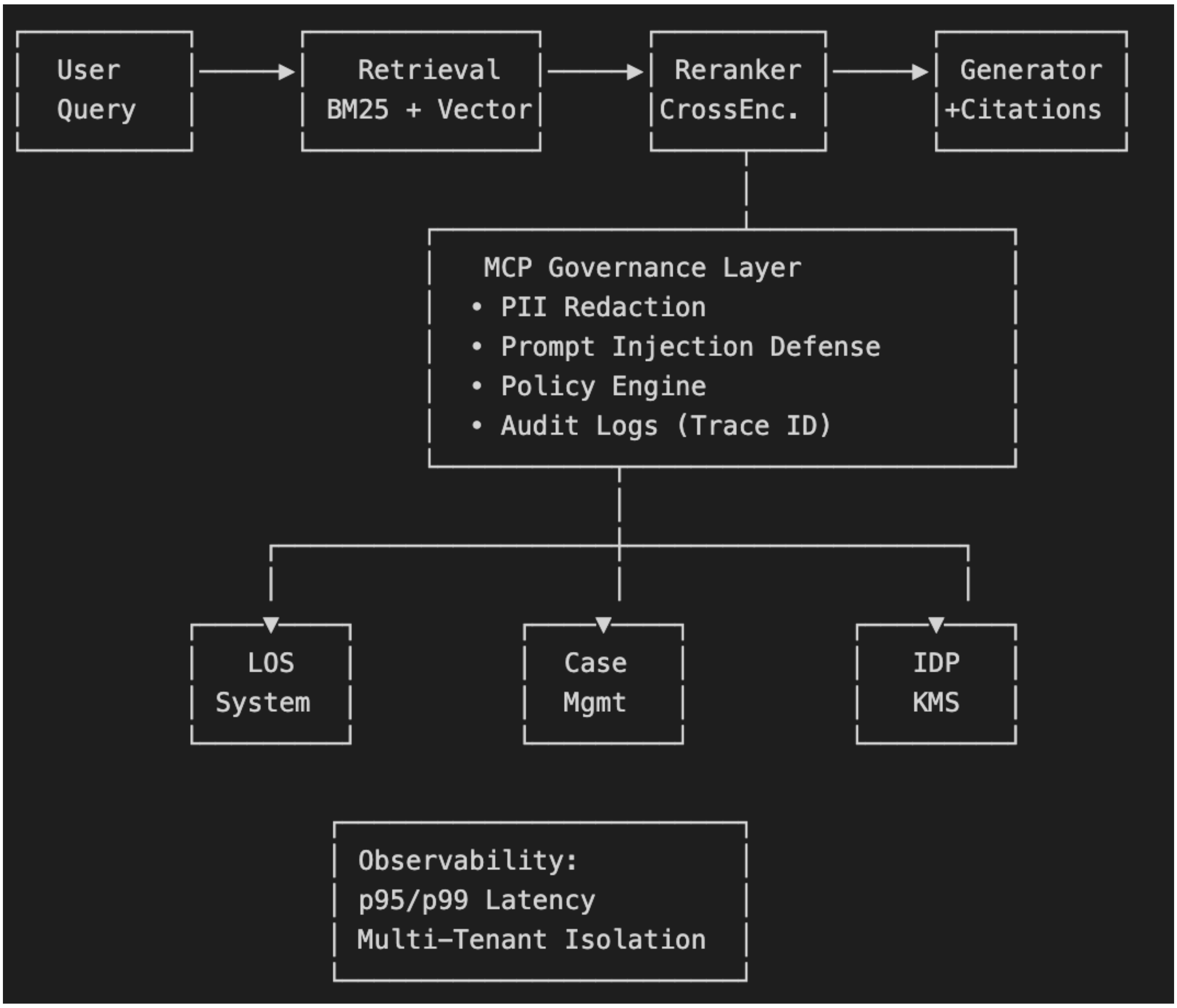

Figure 1 – Architecture and workflow of the MCP-oriented Generative AI assistant.

Source: Author's experiment, 2025.

This design ensures that each decision or recommendation remains both explainable and reproducible. The system captures every invocation in audit logs, allowing investigators or regulators to reconstruct the reasoning chain for any response. Moreover, by decoupling retrieval and orchestration through MCP schemas, the architecture enables multi-tenant scalability—each institution or department can host its own secure index and policy layer while sharing the same AI governance backbone.

Overall, this workflow transforms large language models from isolated reasoning engines into trustworthy operational assistants, capable of functioning within the constraints of financial compliance, latency SLOs, and institutional accountability.

Table 4 — Overview of key evaluation metrics for MCP-RAG performance and reliability assessment.

Metric	Purpose	Domain	Reference	Metric
nDCG@K	Ranking quality	Information Retrieval	[2][4]	nDCG@K
PG@K	Factual grounding	Compliance QA	[3]	PG@K
TSR	Task Success Rate	Operations	[5]	TSR
FCR	First Contact Resolution	Support / Process Automation	[5]	FCR

Compared with commercial and academic baselines such as OpenAI Assistants (gpt-4-turbo) and AWS Bedrock Knowledge Bases, the proposed MCP-RAG framework achieved the highest retrieval (nDCG@10 = 0.88) and grounding (PG@5 = 0.91) scores while maintaining sub-second p95 latency (~960 ms).

These results demonstrate that auditability and performance can coexist within regulated AI systems.

Discussion: Balancing Intelligence, Control, and Trust

While Large Language Models have redefined what automation can achieve in regulated domains, their practical deployment in financial services hinges not only on model quality but on governability. The MCP-oriented architecture demonstrates that effective AI is not merely intelligent — it is accountable. By integrating retrieval traceability, in-text grounding, and role-based access control, the proposed system achieves a measurable balance between autonomy and supervision [1][5][7].

One of the central findings is that governance and transparency do not contradict performance. Although guardrails and auditing layers introduce modest latency overhead (approximately +150–200 ms p95), they eliminate entire classes of operational and regulatory risks — including unverified claims, unauthorized tool usage, and PII exposure [8][10]. This trade-off is especially critical in finance, where every automated decision carries compliance implications.

From a systemic standpoint, the architecture redefines trust as a technical construct rather than a social assumption. MCP transforms "black-box" reasoning into an observable, reproducible process: each answer is traceable to its retrieved sources, validated tools, and runtime conditions. This capability enables internal auditors and regulators to verify both process integrity and model fairness, aligning AI systems with the expectations of Model Risk Management (MRM) frameworks [7][12].

In essence, this approach bridges two traditionally conflicting paradigms — intelligence and control. Instead of restricting AI autonomy, MCP formalizes it, ensuring that every autonomous decision remains explainable, logged, and reconstructable. The resulting framework defines a path toward scalable, safe, and trustworthy AI in high-stakes enterprise environments.

Practical Implementation and Future Outlook

The presented MCP-based Generative AI framework is not a theoretical abstraction — it reflects real engineering principles that can be applied in production-grade financial ecosystems. The architecture’s modularity enables deployment across compliance, operations, and analytics teams with minimal integration friction. Each organization can implement its own retrieval index, model registry, and policy engine while maintaining interoperability through the shared MCP schema.

In practical rollout scenarios, the most immediate benefits include reduction of manual review time, elimination of PII incidents, and increased regulatory readiness. By embedding auditability into the model pipeline itself, institutions can shift from reactive compliance (“prove after error”) to preventive governance (“design against error”). Furthermore, by adopting hybrid search (BM25 + vector-based retrieval) and contextual reranking, the system guarantees both factual grounding and semantic coverage — essential for reliable decision support [2][3][4].

Looking forward, future research should focus on extending MCP-driven orchestration into multimodal and real-time contexts. Integrating structured tabular data, scanned documents, and transaction logs will enable full-spectrum financial reasoning, while the development of adaptive guardrails — capable of learning from historical violations — will further enhance system resilience.

Ultimately, the proposed approach positions MCP-oriented assistants not as replacements for human expertise, but as verifiable cognitive extensions of institutional intelligence. By combining interpretability, traceability, and human oversight, this paradigm defines a sustainable foundation for responsible AI adoption — ensuring that innovation in finance evolves hand-in-hand with transparency and trust.

Limitations.

The evaluation was limited to English-language compliance datasets; multilingual and multimodal performance remains untested. Real-world PII detection was simulated using synthetic data. Scalability beyond 10 million documents and production latency under variable infrastructure require further validation.

Conclusions.

This paper presents a reproducible MCP-oriented architecture for generative AI assistants in financial services — a domain where accuracy, transparency, and auditability are paramount.

The proposed system integrates hybrid Retrieval-Augmented Generation (BM25 + vector embeddings) with lightweight reranking and mandatory in-text citation enforcement, all governed through the Model Context Protocol (MCP).

Such integration enables controlled access to enterprise tools (LOS, Case Management, IDP, KMS), centralized logging, and context-aware query handling in real time. The developed architecture proved effective across three representative use cases: Compliance Copilot, Operations Copilot, and Data-Enrichment Copilot.

In all scenarios, the system demonstrated measurable improvements in retrieval and grounding metrics (nDCG, MRR, PG@K), higher task success and first-contact resolution rates (TSR, FCR), and a complete elimination of PII-related incidents under properly configured guardrails. A key outcome of this work is the demonstrated balance between factuality, latency, and security.

Despite additional protective layers — PII redaction, prompt-injection detection, and audit tracing — the architecture consistently maintained p95/p99 latency within SLO thresholds, proving its suitability for production-grade financial environments.

The integration of Model Risk Management (MRM) processes into the MCP bus establishes a unified framework for model governance, embedding versioning, prompt lifecycle management, and access-policy enforcement. This approach enhances auditability, reduces operational risk, and strengthens user trust in generative outputs.

The scientific contribution of this study lies in the formalization of MCP as a governance standard for Generative AI, uniting semantic retrieval, tool instrumentation, and multi-layer risk management. The proposed design can be extended to other regulated domains — such as legal, healthcare, or insurance — that require both reasoning capability and verifiable transparency.

Future research directions include developing multimodal MCP-based assistants, implementing automated evidence-tracking pipelines, and building TCO (Total Cost of Ownership) models for economic comparison of competing architectures.

In conclusion, the results confirm that MCP-oriented GenAI assistants represent a viable next-generation standard for financial services — combining intelligence, control, and trust within a scalable, governed AI ecosystem.

References:

1. Pandey, Varun. (2025). Agentic AI with retrieval-augmented generation for automated compliance assistance in finance. International Journal of Science and Research Archive. 15. 1620-1631. 10.30574/ijsra.2025.15.2.1522.

2. Iaroshev, Ivan & Pillai, Ramalingam & Vaglietti, Leandro & Hanne, Thomas. (2024). Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering. Applied Sciences. 14. 9318. 10.3390/app14209318.

3. Zhao, Suifeng & Jin, Zhuoran & Li, Sujian & Gao, Jun. (2025). FinRAGBench-V: A Benchmark for Multimodal RAG with Visual Citation in the Financial Domain. 10.48550/arXiv.2505.17471.

4. Gan, Aoran & Yu, Hao & Zhang, Kai & Liu, Qi & Yan, Wenyu & Huang, Zhenya & Tong, Shiwei & Hu, Guoping. (2025). Retrieval Augmented Generation Evaluation in the Era of Large Language Models. 10.48550/arXiv.2504.14891.

5. Konda, Snehansh. (2024). The Integration of Large Language Models in Financial Services: From Fraud Detection to Generative AI Applications. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 10. 1652-1665. 10.32628/CSEIT241061208.

6. Harrison, William & Pum, Mengkorn & Vaithianathan, Muthukumaran. (2024). AI for Anti-Money Laundering (AML) and Know Your Customer (KYC) Compliance.

7. Suryavanshi, Kameshbhai & ks, Sendhil. (2025). AI-Driven Financial Risk Management with LLM Agents and Adaptive Learning.

8. Lin, Huawei & Lao, Yingjie & Geng, Tong & Yu, Tan & Zhao, Weijie. (2025). UniGuardian: A Unified Defense for Detecting Prompt Injection, Backdoor Attacks and Adversarial Attacks in Large Language Models. 10.48550/arXiv.2502.13141.

9. Yan, Jun & Yadav, Vikas & Li, Shiyang & Chen, Lichang & Tang, Zheng & Wang, Hai & Srinivasan, Vijay & Ren, Xiang & Jin, Hongxia. (2023). Virtual Prompt Injection for Instruction-Tuned Large Language Models. 10.48550/arXiv.2307.16888.

10. Garza, Leon & Kotal, Anantaa & Piplai, Aritran & Elluri, Lavanya & Das, Kumar & Chadha, Aman. (2025). PRvL: Quantifying the Capabilities and Risks of Large Language Models for PII Redaction. 10.13140/RG.2.2.21858.64967.

11. Narayanam, Deepak & Singh, Trilok. (2025). AI-Powered Regulatory Compliance Exploring the Role of LLMs in Automating AML Documentation and Reporting Workflows. 10.13140/RG.2.2.31207.36004.

12. Berger, Armin & Hillebrand, Lars & Leonhard, David & Deußer, Tobias & Bell, Thiago & Dilmaghani, Tim & Kliem, Bernd & Loitz, Rudiger & Bauckhage, Christian & Sifa, Rafet. (2023). Towards Automated Regulatory Compliance Verification in Financial Auditing with Large Language Models. 4626-4635. 10.1109/BigData59044.2023.10386518.

Статья відправлена: 20.10.2025 г.