Computer Fraud and Security
ISSN (online): 1873-7056

Cross-Cloud Generative AI Framework for AML, KYC, and Claims Fraud Detection

Venkata Raja Ravi Kumar Gelle
Independent Researcher, USA

Abstract

Financial institutions and insurance companies are struggling more and more to spot money laundering cases, identity fraud scams, and fake insurance claims. The existing rule-based systems that detect only know how to react because they have significant shortcomings when adaptive criminal methodologies confront them. Static threshold parameters fail to capture evolving transactional patterns. Conventional machine learning implementations operate within isolated data environments. Semantic reasoning capabilities remain absent from current detection frameworks. A Cross-Cloud Generative AI Framework addresses architectural gaps through integration of large language models, retrieval-augmented generation mechanisms, and multi-agent reasoning systems. The framework establishes unified data ingestion pipelines across distributed cloud storage environments. Domain-tuned language models process heterogeneous data types, including transaction records, identity documents, and claims artifacts. Retrieval-augmented generation grounds analytical outputs in verifiable documentation, reducing hallucination risks. Specialized agents rapidly communicate in different domains of anti-money laundering, identity verification, and claims fraud. Cross-cloud orchestration can distribute inference workloads depending on computational requirements as well as data residency regulations. Zero-trust security principles safeguard sensitive financial information through continuous authentication and micro-segmentation. An explainability layer generates transparent reasoning paths satisfying regulatory examination requirements. The framework establishes a scalable foundation for next-generation financial crime prevention infrastructure.

Keywords: Cross-Cloud Architecture, Generative Artificial Intelligence, Anti-Money Laundering, Know Your Customer, Fraud Detection, Retrieval-Augmented Generation

1. INTRODUCTION

The growing availability of digital financial services has led to various types of crimes that exploit banking systems globally. Money laundering is still one of the main sources of threats to economic stability and institutional integrity. Criminal proceeds are mixed with clean money flowing through legitimate financial channels by using highly sophisticated layering techniques. Schneider et al. documented that money laundering operates through three distinct phases: placement, layering, and integration [1]. The position phase includes introducing illicit budget into the monetary system. Layering creates complex transaction chains to obscure the money trail. Integration returns laundered funds to criminals through seemingly legitimate sources [1]. Financial institutions struggle to detect these activities due to the deliberate complexity of criminal methodologies.

Traditional detection systems rely on rule-based engines with static threshold parameters. Transaction monitoring platforms flag activities exceeding predetermined limits. However, sophisticated actors structure transactions below detection thresholds. Rule-based approaches cannot adapt to evolving criminal techniques. Mat Isa et al. identified significant gaps between regulatory expectations and banking sector capabilities in addressing money laundering risks [2]. The research highlighted that bankers and regulators maintain different perspectives on risk assessment priorities [2]. Regulators emphasize compliance verification and penalty enforcement. Banking institutions focus on operational efficiency and customer service continuity. This misalignment creates vulnerabilities in detection frameworks [2].

Contemporary machine learning implementations offer improved pattern recognition capabilities. Supervised classification algorithms identify suspicious transactions based on historical labeled data. Anomaly detection models flag deviations from established behavioral baselines. However, these systems operate within isolated data environments. Structured transactional records receive analytical attention while unstructured sources remain underutilized. Adverse media reports contain valuable intelligence about criminal networks. Regulatory bulletins provide updated guidance on emerging threats. Claims documentation reveals inconsistencies indicative of

insurance fraud. Current detection frameworks lack semantic reasoning capabilities to process these heterogeneous data types effectively.

The insurance sector faces parallel challenges in claims fraud detection. Fraudulent claims in most cases involve fabricated documentation, exaggeration of the damages, and staging of the incidents. Manual review processes cannot scale with increasing claim volumes. Siloed data systems prevent the correlation of claimant histories across multiple submissions. Cross-domain fraud indicators remain undetected when analytical tools lack contextual understanding.

This paper introduces a Cross-Cloud Generative AI Framework addressing these architectural limitations. The proposed system synthesizes large language model capabilities with multi-cloud infrastructure optimization. Retrieval-augmented generation mechanisms ground analytical conclusions in verifiable documentation. Multi-agent reasoning architectures coordinate specialized detection functions across anti-money laundering, identity verification, and claims fraud domains. Cross-cloud orchestration protocols optimize computational resource allocation while satisfying jurisdictional data residency requirements. The framework enables a comprehensive analysis of heterogeneous data types. Regulatory compliance, operational resilience, and explainable decision outputs guide the architectural design. This approach establishes a foundation for next-generation financial crime prevention infrastructure.

2. RELATED WORK

Prior contributions in financial crime detection have predominantly focused on rule-based systems and supervised machine learning classifiers. Transaction monitoring platforms rely on static threshold parameters to identify suspicious activities. Random forest and gradient boosting algorithms have demonstrated effectiveness in anomaly detection tasks. However, existing frameworks process structured transactional data in isolation from unstructured intelligence sources. Adverse media reports, regulatory bulletins, and claims documentation remain underutilized in detection workflows.

Recent advances in natural language processing have introduced large language models capable of semantic reasoning across diverse text formats. Domain-specific fine-tuning improves the performance of models on specialized terminology and regulatory language. Retrieval-augmented generation mechanisms provide model outputs that are grounded in verifiable documentation, thereby alleviating concerns about the hallucination of generative systems. Multi-agent architectures grant the possibility of breaking down the tasks into smaller ones and then solving them collaboratively, which are complex analytical domains.

The Cross-Cloud Generative AI Framework synthesizes these technological advances into a unified detection infrastructure. The framework introduces novel integration of domain-tuned language models with multi-cloud orchestration strategies. Specialized agents coordinate across anti-money laundering, identity verification, and claims fraud functions. Cross-cloud workload distribution optimizes computational resource allocation while satisfying jurisdictional data residency requirements. Zero-trust security principles and explainable reasoning paths address regulatory compliance mandates for financial institutions and insurance organizations.

3. ARCHITECTURAL FOUNDATION AND DATA INTEGRATION

3.1 Multi-Cloud Data Ingestion Pipeline

The framework establishes unified data access across distributed cloud storage systems. Financial crime detection requires the consolidation of diverse data sources into coherent analytical environments. Structured transactional databases contain account histories and payment records. Unstructured document repositories store identity verification materials and claims artifacts. Sanction screening lists provide regulatory watchlist information. External adverse media feeds deliver real-time intelligence about criminal networks and suspicious entities.

Traditional data warehouse architectures struggle with heterogeneous data types. Armbrust et al. introduced the lakehouse paradigm to address limitations of conventional data management systems [3]. The lakehouse architecture combines data lake flexibility with data warehouse reliability [3]. This approach enables direct analytical processing on raw data without extensive transformation overhead. The architecture supports machine learning workloads alongside traditional business intelligence queries [3]. Financial institutions benefit from unified access to both structured transaction records and unstructured compliance documents.

Data normalization protocols transform heterogeneous formats into standardized schemas. Ingestion pipelines apply consistent encoding across multi-cloud storage systems. Schema enforcement ensures downstream language

model compatibility. The architecture is equipped with the storage of data lineage for regulatory audit requirements. Cross-cloud replication strategies are there for disaster recovery as well as jurisdictional compliance. Incremental ingestion mechanisms are there to reduce processing latency for time-sensitive fraud detection workflows.

3.2 Vector Database Infrastructure

Semantic search capabilities require specialized storage infrastructure for embedding representations. Vector databases store high-dimensional numerical representations of regulatory documents. Historical case files receive embedding transformations for similarity matching. Compliance guidelines and policy documents populate searchable vector indices. Investigation workflows benefit from rapid retrieval of contextually relevant precedents.

Johnson et al. developed efficient algorithms for billion-scale similarity search operations [4]. The research demonstrated that graphics processing unit acceleration enables substantial performance improvements for approximate nearest neighbor queries [4]. Vector similarity computations scale efficiently across large document collections using optimized indexing structures [4]. Financial crime investigation platforms leverage these capabilities for real-time document retrieval during analyst workflows.

The vector database infrastructure supports multiple embedding models for different document categories. Regulatory bulletins receive specialized encoding optimized for compliance terminology. Transaction narratives utilize embeddings trained on financial domain corpora. Claims documentation processing employs models attuned to insurance-specific language patterns. Hybrid search mechanisms combine vector similarity with keyword filtering for enhanced precision. The infrastructure enables investigators to identify precedent cases matching current suspicious activity patterns. Contextual retrieval grounds analytical conclusions in documented historical evidence.

Component	Function	Data Types Processed
Data Ingestion Pipeline	Consolidates distributed data sources into a unified analytical environment	Structured databases, unstructured repositories, sanction lists, adverse media feeds
Lakehouse Architecture	Combines data lake flexibility with warehouse reliability	Raw data, transformed schemas, analytical workloads
Schema Normalization	Transforms heterogeneous formats into standardized structures	Multi-format documents, cross-platform records
Vector Database	Stores embedded representations for semantic search	Regulatory documents, historical case files, compliance guidelines
Similarity Search Engine	Enables rapid retrieval during investigation workflows	Embedded document vectors, precedent case indices

Table 1. Architectural Components for Cross-Cloud Data Integration [3, 4].

4. GENERATIVE AI PROCESSING ARCHITECTURE

4.1 Domain-Tuned Language Models

The framework deploys large language models fine-tuned on financial crime investigation corpora. General-purpose language models lack a specialized understanding of financial terminology. Domain adaptation addresses this limitation through targeted training approaches. Wu et al. developed BloombergGPT as a large language model specifically designed for the financial domain [5]. The model training incorporated a mixed dataset construction approach. Financial data sources included news articles, regulatory filings, press releases, and financial instrument documentation [5]. General-purpose text corpora supplemented domain-specific materials to preserve broad language capabilities [5]. This mixed training strategy balanced specialized financial knowledge with foundational language understanding [5].

The tuning process for fraud detection incorporates regulatory lexicons specific to anti-money laundering compliance. Suspicious activity report narratives provide training examples for investigative language patterns. Insurance fraud investigation protocols contribute terminology related to claims assessment workflows. Identity verification documentation offers training material for customer due diligence processes. Domain-adapted models demonstrate enhanced capability in identifying suspicious behavioral indicators. Inconsistencies within customer

identity profiles receive improved detection accuracy. Anomalous patterns in claims submissions trigger appropriate analytical responses.

Fine-tuning strategies must balance domain specialization with general comprehension. Catastrophic forgetting poses risks when models lose broader capabilities during specialized training. The BloombergGPT research demonstrated that mixed dataset approaches mitigate this challenge [5]. Financial task performance improves while general language abilities remain intact. Continuous model updates are there to facilitate ever-changing regulatory requirements as well as newly emerging criminal methodologies.

4.2 Retrieval-Augmented Generation Layer

The Retrieval-Augmented Generation component grounds language model outputs in verifiable documentation. Lewis et al. introduced the RAG paradigm combining parametric and non-parametric memory systems [6]. Parametric memory resides within model parameters learned during training. Non-parametric memory consists of external knowledge sources accessed through retrieval mechanisms [6]. The architecture employs a retrieval component to identify relevant documents from large corpora [6]. A generator component then produces outputs conditioned on both the input query and retrieved passages [6].

During analysis execution, the RAG layer retrieves relevant policy documents from institutional repositories. Historical investigation records provide precedent information for the current case assessment. Regulatory guidance documents offer authoritative interpretations of compliance requirements. The retrieval mechanism queries vector indices to identify contextually relevant materials. Semantic similarity matching ensures retrieved documents align with analytical queries.

The grounding mechanism reduces hallucination risks inherent to generative systems. Lewis et al. demonstrated that retrieved documents provide factual anchoring for generated outputs [6]. Knowledge-intensive tasks benefit substantially from this retrieval augmentation approach [6]. Financial crime investigators verify source materials referenced in analytical outputs. Audit trails document the evidentiary basis for risk assessments. Compliance officers confirm alignment between model conclusions and institutional policies. Generated narratives include citations to supporting documentation. This transparency enables human reviewers to validate analytical reasoning effectively.

Component	Purpose	Application Domain
Domain-Tuned LLM	Processes specialized financial and regulatory terminology	Anti-money laundering, insurance fraud, and identity verification
Mixed Dataset Training	Balances domain expertise with general language capabilities	Financial news, regulatory filings, and general text corpora
Retrieval Component	Identifies relevant documents from institutional repositories	Policy documents, investigation records, and regulatory guidance
Generator Component	Produces outputs conditioned on queries and retrieved passages	Risk assessments, compliance narratives, evidence summaries
Grounding Mechanism	Reduces hallucination risks through documentary anchoring	Knowledge-intensive analytical tasks

Table 2. Generative AI Components for Financial Crime Detection [5, 6].

5. MULTI-AGENT REASONING AND ORCHESTRATION

5.1 Specialized Agent Coordination

A multi-agent reasoning module coordinates specialized analytical agents addressing distinct fraud detection domains. Complex financial crime investigations require decomposition into manageable subtasks. Dorri et al. defined a multi-agent system as a collection of autonomous agents interacting within a shared environment [7]. Each agent possesses individual capabilities and objectives. Agents exhibit key properties including autonomy, reactivity, proactiveness, and social ability [7]. Autonomy enables independent operation without direct human intervention. Reactivity allows agents to perceive environmental changes and respond appropriately [7]. Proactiveness drives goal-directed behavior beyond simple stimulus response. Social ability facilitates interaction and collaboration with other agents [7].

Transaction pattern agents evaluate financial flow anomalies and network relationships. These agents analyze payment sequences and counterparty connections. Suspicious structuring patterns trigger alert generation for

investigator review. Identity validation agents assess document authenticity and profile consistency. Customer onboarding materials receive scrutiny for forgery indicators. Cross-referencing mechanisms compare submitted information against authoritative sources. Claims analysis agents examine submission irregularities and supporting evidence. Documentation completeness and claimant behavioral patterns inform fraud probability assessment.

Agent coordination requires structured communication protocols. Dorri et al. identified communication as fundamental to multi-agent collaboration [7]. Message passing enables information exchange between specialized components. Shared knowledge repositories provide common reference points for decision-making. Task allocation mechanisms distribute analytical responsibilities based on agent expertise [7]. The synthesis of individual assessments produces comprehensive risk evaluations. Conflicting conclusions undergo arbitration through hierarchical decision mechanisms.

5.2 Cross-Cloud Workload Distribution

Orchestration mechanisms distribute inference workloads across cloud environments. Financial institutions increasingly operate infrastructure spanning multiple providers. Tomarchio et al. defined cloud orchestration as the automated coordination and management of cloud services and resources [8]. Multi-cloud adoption addresses several organizational concerns. Vendor lock-in avoidance reduces dependency on single providers [8]. Geographic distribution supports disaster recovery requirements. Service availability improves through redundant infrastructure deployment.

The research identified significant challenges in multi-cloud orchestration [8]. Heterogeneity across providers complicates unified management approaches. Different interfaces, service models, and operational procedures require abstraction layers [8]. Interoperability standards remain inconsistent across the cloud landscape. Orchestration frameworks must bridge these differences transparently.

Computational requirements influence workload placement decisions. Intensive language model inference benefits from specialized accelerator availability. Data residency regulations constrain processing locations for sensitive financial information. Jurisdictional requirements mandate geographic boundaries for certain data categories. Cost optimization parameters guide resource selection across varying pricing structures.

This distribution strategy ensures operational continuity during infrastructure disruptions. Provider outages do not halt critical fraud detection operations. Failover mechanisms redirect workloads to available resources automatically. Load balancing distributes processing demands across healthy components. The framework maintains compliance with data sovereignty requirements throughout migration events. Encryption protects information during cross-cloud transfers. Audit logging documents all data movements for regulatory examination purposes.

Agent Type	Analytical Function	Key Capabilities
Transaction Pattern Agent	Evaluates financial flow anomalies	Payment sequence analysis, counterparty connection mapping, and structuring detection
Identity Validation Agent	Assesses document authenticity	Forgery indicator identification, cross-reference verification, and profile consistency checks
Claims Analysis Agent	Examines submission irregularities	Documentation completeness review, behavioral pattern assessment, evidence coherence evaluation
Orchestration Module	Distributes workloads across cloud environments	Computational optimization, data residency compliance, and failover management
Coordination Protocol	Enables inter-agent collaboration	Message passing, task allocation, conflict arbitration

Table 1. Multi-Agent Reasoning Components for Fraud Detection [7, 8].

6. SECURITY FRAMEWORK AND EXPLAINABILITY

6.1 Zero-Trust Security Architecture

The architecture reflects a change in security strategy by the adoption of the most stringent security measures of the zero-trust type to all the components of the framework. Traditional security models are based on perimeter-based defenses. Network boundaries define trust zones in conventional approaches. He et al. surveyed zero-trust architecture and identified fundamental shifts from perimeter-centric security paradigms [9]. The zero-trust model

operates on the principle of never trust and always verify [9]. Every access request requires authentication regardless of network location. Internal network position does not confer implicit trust status.

The framework enforces continuous verification throughout user sessions. He et al. identified identity verification as a core component of zero-trust implementations [9]. Multi-factor authentication strengthens credential validation processes. Session behaviors undergo monitoring for anomalous patterns. Micro-segmentation divides network resources into isolated security zones [9]. Lateral movement between segments requires explicit authorization. This containment strategy limits potential damage from compromised credentials.

Least-privilege access control restricts user permissions to the minimum necessary levels [9]. Role-based mechanisms map organizational responsibilities to system capabilities. Privilege escalation requires explicit approval through defined workflows. Encrypted data transmission protects information traversing network segments. Transport layer security protocols secure inter-component communications. Data at rest receives cryptographic protection using strong algorithms. Federated identity management enables secure access across multi-cloud deployments. Security tokens convey authentication status to distributed service providers [11].

6.2 Explainable AI Layer

The explainability layer generates transparent reasoning paths for AI-generated outputs. Financial crime detection systems must provide interpretable results. Adadi and Berrada surveyed explainable artificial intelligence and identified a growing demand for transparency in machine learning systems [10]. Black-box models achieve high predictive accuracy without revealing internal logic [10]. This opacity creates significant challenges for regulated industries. Decision subjects and regulators require an understanding of automated reasoning processes.

The research identified the right to explanation as an emerging legal and ethical requirement [10]. Automated decisions affecting individuals demand justification capabilities. Model verification requires transparency to confirm intended system behavior. Bias detection depends on interpretable outputs that reveal discriminatory patterns [10]. User trust develops through comprehensible system operations. Adadi and Berrada noted inherent tension between model complexity and interpretability [10]. Highly accurate models often sacrifice transparency for predictive performance.

Risk indicators accompany each analytical assessment generated by the framework. Investigators review factor contributions influencing overall risk determinations. Evidence summaries compile relevant documentation supporting conclusions. Source citations enable verification against original materials. Reasoning paths document logical steps connecting inputs to outputs.

The explainability layer satisfies regulatory examination requirements. Financial regulators expect an explanation of the automated decision rationale. Model risk management frameworks mandate documentation of AI system behavior [10]. Audit trails preserve reasoning records for retrospective examination. Compliance officers validate alignment between outputs and institutional policies. Human reviewers retain authority to override AI-generated recommendations. This oversight ensures accountability remains with qualified personnel rather than autonomous systems [12,13].

Component	Function	Implementation Mechanism
Continuous Authentication	Verifies user identity throughout sessions	Multi-factor authentication, session behavior monitoring
Micro-Segmentation	Isolates network resources into security zones	Lateral movement restriction, containment strategies
Least-Privilege Access	Restricts permissions to the minimum necessary levels	Role-based access control, privilege escalation workflows
Reasoning Path Generator	Documents logical steps connecting inputs to outputs	Factor contribution analysis, evidence compilation
Audit Trail System	Preserves records for regulatory examination	Source citations, decision rationale documentation

Table 4. Security Architecture and AI Transparency Mechanisms [9, 10].

CONCLUSION

The Cross-Cloud Generative AI Framework represents a significant advancement in financial crime detection and insurance fraud prevention capabilities. Integration of large language models with retrieval-augmented generation

enables semantic understanding of diverse data sources previously inaccessible to traditional detection systems. Domain-tuned models demonstrate enhanced capability in processing regulatory terminology, suspicious activity narratives, and claims documentation. Multi-agent architectures decompose complex investigative tasks into specialized analytical functions. Transaction pattern agents, identity validation agents, and claims analysis agents collaborate through structured communication protocols. Coordination mechanisms are individual assessments in comprehensive risk evaluations. Cross-cloud orchestration guarantees computational efficiency at the same time as it maintaining compliance with the requirements of data sovereignty of the jurisdiction. Failover mechanisms deliver operational resilience during a break in the infrastructure. The implementation of zero-trust security is aimed at protecting sensitive financial and personal data by continuous verification and least-privilege access controls. Micro-segmentation limits the possible breaches to the isolated network zones. The explainability layer addresses regulatory transparency mandates through interpretable reasoning paths and evidence documentation. Human reviewers retain override authority over automated recommendations. Continuous refinement of the model through the feedback loops of the analysts is a way to ensure the model's effectiveness against the ever-changing criminal techniques. The architectural design provides the infrastructure with the capability of being extended to be able to include new artificial intelligence capabilities and, at the same time, maintain regulatory compliance and operational stability in the distributed cloud environments.

References

[1] Schneider et al., "Money Laundering: Some Facts," Economics of Security Working Paper Series, 2010. [Online]. Available: https://www.econstor.eu/bitstream/10419/119350/1/diw_econsec0025.pdf

[2] Chris Cole, "Next Generation CFP Modules," 2012. [Online]. Available: https://web.archive.org/web/20160825160812id_/https://www.finisar.com/sites/default/files/resources/next_generation_cfp_modules_ofc_nfoec_2012.pdf

[3] Athira Nambiar and Divyansh Mundra, "An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management," MDPI, 2022. [Online]. Available: https://www.mdpi.com/2504-2289/6/4/132

[4] Jeff Johnson et al., "Billion-scale similarity search with GPUs," arXiv, 2017. [Online]. Available: https://arxiv.org/pdf/1702.08734

[5] Shijie Wu et al., "BloombergGPT: A Large Language Model for Finance," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2303.17564

[6] Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 34th Conference on Neural Information Processing Systems, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

[7] ALI DORRI et al., "Multi-Agent Systems: A Survey," IEEE Access, 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8352646

[8] Orazio Tomarchio, "Cloud resource orchestration in the multi-cloud landscape: a systematic review of existing frameworks," Journal of Cloud Computing, 2020. [Online]. Available: https://link.springer.com/content/pdf/10.1186/s13677-020-00194-7.pdf

[9] Yuanhang He et al., "A Survey on Zero Trust Architecture: Challenges and Future Trends," Wiley, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/6476274

[10] AMINA ADADI AND MOHAMMED BERRADA, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, 2018. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8466590

[11] R. Kaur, "Operational challenges and strategic adaptation of non-profit animal shelters: A case study of the Winnipeg Pet Rescue Shelter," International Journal of Environmental Sciences, vol. 12, no. 1, pp. 1–8, 2021. [Online]. Available: https://doi.org/10.64252/mktp4w61

[12] V. Sahoo, "Visual analytics and machine learning for scalable growth-oriented product management," Journal of Economics Intelligence and Technology, vol. 1, no. 2, pp. 24–31, 2025.

[13] A. Y. L. Guarin, "Holistic fitness as a competitive advantage: Expanding market share through female-oriented movement practices," Journal of Economics Intelligence and Technology, vol. 1, no. 2, pp. 16–23, 2025.