SAR Narrative Generation with Transparent Audit Trail Using LLMS And Rule-Based Detection

Ayush Kurkure[1], Niranjan Kulkarni[2], Bhushan Kurkure[3], Jayaprakash Konduru[4], Jaysh Khobragade[5], Prof. Deepali Deshpande[6]
[1,2,3,4,5,6]*Dept. Information Technology, Vishwakarma Institute Of Technology Pune, India*

***Abstract*—Financial institutions are required to file Suspicious Activity Reports (SARs) to identify potential money laundering, fraud, and other financial crimes. However, SAR drafting is a time-consuming and labor-intensive process that requires high accuracy, consistency, and regulatory compliance. This paper presents an AI-powered SAR Narrative Generator with an integrated audit trail designed to automate regulator-ready SAR narrative creation while maintaining explainability and transparency. The proposed system combines a deterministic Python-based rule engine for suspicious activity detection with Large Language Models (LLMs) for narrative generation. LangChain orchestrates Retrieval-Augmented Generation (RAG), regulatory context retrieval using ChromaDB, prompt construction, and audit logging. Llama 3.1 hosted through vLLM generates human-readable narratives based strictly on structured evidence. PostgreSQL stores customer information, alerts, generated narratives, and full audit logs for traceability. Analysts review and approve generated drafts through a Streamlit dashboard before filing. Experimental observations indicate significant reduction in SAR drafting time from several hours to a few minutes while improving consistency and compliance. The proposed architecture ensures scalability, explainability, and secure deployment within financial institutions.**

***Index Terms*—Suspicious Activity Report, Anti-Money Laundering, Large Language Model, LangChain, RAG, ChromaDB, Audit Trail, Financial Crime Detection.**

I. INTRODUCTION

Financial institutions process millions of financial transactions daily. Detecting suspicious financial activities such as money laundering, fraud, structuring, and illegal fund transfers is one of the most critical responsibilities of banking and compliance departments. Regulatory organizations such as the Financial Crimes Enforcement Network (FinCEN) mandate banks to submit Suspicious Activity Reports (SARs) whenever unusual or potentially criminal financial activity is detected.

1.1. Background

Money laundering and financial fraud have become major threats to the global banking system. Criminal organizations often use complex transaction patterns to disguise illegal funds and move money across multiple accounts and jurisdictions. Banks are legally required to monitor suspicious activities under Anti-Money Laundering (AML) and Know Your Customer (KYC) regulations. Suspicious Activity Reports are essential documents submitted to regulators whenever abnormal financial activity is observed.

Manual preparation of SAR narratives introduces several operational challenges:

- High workload on compliance teams
- Delayed investigation processes
- Inconsistent reporting quality
- Increased operational costs
- Regulatory penalties due to incomplete documentation

Artificial Intelligence technologies such as Natural Language Processing (NLP) and Large Language Models provide opportunities to automate narrative writing tasks. However, AI-generated content must remain transparent, explainable, and auditable to satisfy regulatory standards.

1.2. Problem Statement

SAR preparation systems rely heavily on human analysts for reviewing transaction records and drafting suspicious activity narratives. Analysts often spend 5–6 hours per report due to the complexity of financial investigations and regulatory documentation requirements.

The major problems identified are:
- Large volume of suspicious transaction alerts
- Manual and repetitive report drafting
- Inconsistent narrative quality across analysts
- Operational bottlenecks due to understaffed compliance teams
- Difficulty maintaining audit trails for AI-generated decisions
- Risk of regulatory non-compliance

Therefore, there is a need for an automated, explainable, and secure system capable of generating regulator-ready SAR narratives while maintaining full transparency and human oversight.

1.3. Problem Statement and Research Gap

The proposed project focuses on automating SAR narrative generation for financial institutions using Artificial Intelligence and rule-based systems.

The scope includes:
- Transaction data ingestion from CSV/API sources
- Rule-based suspicious transaction detection
- Retrieval-Augmented Generation (RAG)
- AI-based SAR narrative generation
- Audit logging and traceability
- Analyst review dashboard
- Secure deployment architecture

The system is designed for banks, compliance teams, fraud monitoring departments, and financial investigation units.

The architecture can also be extended for:
- Insurance fraud reporting
- Cybercrime investigation documentation
- Regulatory reporting systems

II. LITERATURE REVIEW

Cardiovascular Several research works and technologies have contributed to the development of AI-assisted compliance systems and financial fraud detection frameworks.

Brown et al. [2] introduced transformer-based Large Language Models capable of performing high-quality text generation using contextual prompts. Their work demonstrated the effectiveness of language models in document automation tasks.

Devlin et al. [7] proposed BERT, a bidirectional transformer architecture that significantly improved Natural Language Processing applications including text classification, semantic understanding, and information retrieval.

Meta AI's Llama 3 model [3] provided an open-weight foundation model capable of generating human-like narratives with strong reasoning capabilities. The model supports enterprise deployment with secure on-premise hosting.

Research on Retrieval-Augmented Generation by Lewis et al. [11] demonstrated how external knowledge retrieval systems can improve factual consistency and reduce hallucinations in LLM-generated outputs.

LangChain [4] introduced orchestration frameworks for managing multi-step AI pipelines involving retrieval, prompt engineering, memory management, and model interactions.

Financial crime detection systems traditionally relied on deterministic rule engines. Studies in Anti-Money Laundering systems emphasize explainability and transparency as key requirements for regulatory acceptance.

Existing solutions often lack:
- Full audit trails
- Human review mechanisms
- Explainable AI workflows
- Secure on-premise deployment
- Regulatory traceability

The proposed system improves upon existing systems by integrating deterministic detection, Retrieval-Augmented Generation, transparent logging, and human oversight.

III. METHODOLOGY

3.1 Materials & Components

Transaction records, customer KYC information, and account details are imported through CSV files or APIs.

The ingestion pipeline performs:
- Schema validation
- Data normalization
- Duplicate removal
- Database insertion

Validated data is stored in PostgreSQL databases.

3.2 Rule Based Detection Engine

The detection engine is implemented using Python and Pandas. Suspicious activities are identified using deterministic rules such as:

- Structuring detection
- Velocity anomalies
- High-risk geography detection
- Flow-through account monitoring
- Counterparty network analysis

Each triggered rule generates a structured alert payload containing:

- Customer details
- Transaction evidence
- Rule metadata
- Risk indicators

3.3 Retrieval Augmented Generation

Regulatory documents and SAR guidelines are stored in ChromaDB vector databases. LangChain retrieves relevant regulatory context during narrative generation.

This ensures:

- Regulatory compliance
- Improved factual consistency
- Reduced hallucination risk

3.4 Retrieval Augmented Generation

LangChain constructs prompts using:

- Customer profile
- Alert metadata
- Evidence transactions
- Regulatory context

Llama 3.1 generates SAR narratives in formal compliance language.

3.5 Audit Logging

All prompts, outputs, edits, timestamps, and approvals are stored in PostgreSQL.

The audit trail ensures:

- Explainability
- Traceability
- Regulatory defensibility

3.6 Human Review

Analysts review generated narratives through a Streamlit dashboard. Human analysts can:

- Edit narratives
- Approve or reject reports
- View transaction evidence
- Export finalized SAR reports

IV. TESTING

4.1. Validation Protocol

Testing of the proposed SAR Narrative Generator with Audit Trail was carried out to ensure that the system works efficiently, securely, and accurately in financial compliance environments. Different testing methods such as unit testing, integration testing, system testing, performance testing, and security testing were performed on the developed framework. Unit testing was used to verify the functionality of individual modules including transaction ingestion, rule-based suspicious activity detection, prompt generation, and API endpoints. Each module produced accurate outputs for both valid and invalid inputs.

Integration testing ensured smooth communication between components such as PostgreSQL, LangChain, ChromaDB, FastAPI, Streamlit, and the Llama 3.1 model. The complete workflow from transaction ingestion to SAR narrative generation and audit logging was tested successfully without data inconsistency. Performance testing was conducted using large transaction datasets to evaluate processing speed, API response time, and database efficiency. The system demonstrated stable performance and generated narratives within a few minutes, significantly faster than manual report preparation.

Security testing was also performed to verify JWT authentication, role-based access control, encrypted database communication, and secure local deployment of the LLM through vLLM. User Acceptance Testing was conducted with simulated compliance analysts who reviewed the generated narratives for readability, accuracy, consistency, and regulatory compliance. The testing phase confirmed that the proposed system is reliable, scalable, secure, and suitable for real-world financial crime reporting applications.

V. RESULTS AND DISCUSSIONS

5.1. Results Obtained

The proposed SAR Narrative Generator with Audit Trail successfully automated the process of generating

Suspicious Activity Report narratives for financial institutions. The system was able to process transaction records, detect suspicious patterns using deterministic rules, retrieve regulatory context through the RAG pipeline, and generate regulator-ready SAR narratives using Llama 3.1. Experimental observations showed that the time required for preparing SAR reports was significantly reduced from several hours of manual effort to only a few minutes of automated processing and analyst review. The generated narratives followed a consistent structure, included relevant transaction evidence, and maintained professional compliance language throughout the report.

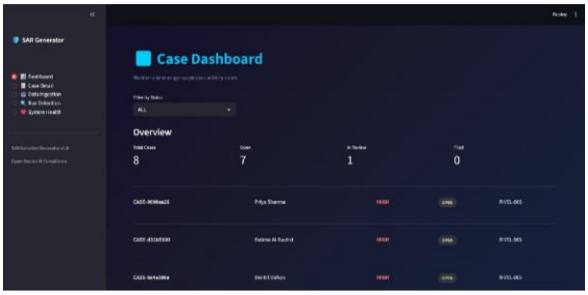

Fig 1: Dashboard

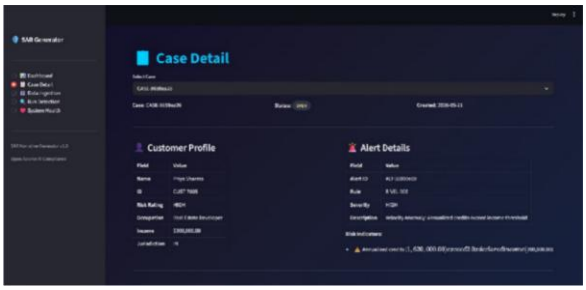

Fig 2 Analyst Dashboard



Fig 3: Report

5.2. Discussion on Rule-Based Detection

The rule-based detection engine performed effectively in identifying suspicious financial activities such as structuring, velocity anomalies, high-risk geographical transfers, and flow-through transactions. Since the detection process was based on predefined deterministic rules implemented using Python and Pandas, the generated alerts were fully explainable and regulator-defensible. Each alert contained detailed metadata including transaction IDs, threshold values, transaction counts, and risk indicators, which improved transparency during investigations. Unlike black-box AI detection systems, the rule engine provided clear reasoning for every flagged activity, making the system more suitable for financial compliance environments.

5.3. Discussion on LLM-Based Narrative Generation

The integration of Llama 3.1 with LangChain enabled high-quality SAR narrative generation using structured transaction evidence and retrieved regulatory guidelines. The generated narratives were written in formal and objective regulatory language, making them suitable for compliance reporting. The Retrieval-Augmented Generation (RAG) pipeline improved factual consistency by injecting FinCEN guidelines and AML regulations into the prompts before generation. The use of low-temperature settings minimized hallucination risks and ensured that the model generated narratives strictly based on the provided evidence. This significantly reduced manual writing effort for analysts while improving consistency across reports.

5.4. Discussion on Scalability/System Performance

The proposed architecture demonstrated strong scalability and stable performance during testing with large transaction datasets. The combination of FastAPI, PostgreSQL, LangChain, ChromaDB, and vLLM enabled efficient handling of multiple cases and concurrent analyst operations. The system processed bulk transaction records efficiently and maintained low response times during narrative generation. Docker and Kubernetes support further improved deployment flexibility and horizontal scalability for enterprise-level adoption. Overall, the system reduced operational bottlenecks, improved analyst productivity, and provided a scalable framework suitable for large financial institutions handling thousands of suspicious activity cases regularly.

VI. FUTURE SCOPE

Proposed SAR Narrative Generator with Audit Trail provides a strong foundation for AI-assisted financial compliance systems and can be enhanced further in multiple areas. In the future, the system can be extended to support real-time transaction monitoring instead of batch-based processing, allowing financial institutions to detect suspicious activities instantly. Advanced machine learning and graph-based fraud detection techniques can also be integrated to identify complex money laundering networks and hidden relationships between accounts.

The system can further be improved by adding multilingual SAR generation capabilities so that reports can be generated in different regulatory languages based on institutional requirements. Integration with blockchain technology may provide tamper-proof audit trails and stronger security for sensitive financial records. Future versions may also include advanced explainable AI dashboards with visual analytics to help compliance analysts better understand suspicious transaction patterns.

In addition, the framework can be expanded beyond banking applications to domains such as insurance fraud detection, cybersecurity incident reporting, tax fraud analysis, and government regulatory compliance systems. Cloud-native deployment and federated AI models can also improve scalability and collaboration between financial institutions while preserving data privacy and security. Overall, the proposed system has significant potential for future research and enterprise-level adoption in intelligent compliance automation.

VII. CONCLUSION

SAR Narrative Generator with Audit Trail successfully demonstrates how Artificial Intelligence can be integrated into financial compliance systems to automate Suspicious Activity Report generation while maintaining transparency, explainability, and regulatory compliance. The system combines deterministic rule-based suspicious activity detection with Large Language Models to generate professional and regulator-ready SAR narratives from structured transaction evidence.

The integration of technologies such as LangChain, ChromaDB, PostgreSQL, FastAPI, Streamlit, and Llama 3.1 enabled the development of a scalable, secure, and efficient compliance automation framework. The system significantly reduced SAR drafting time, improved reporting consistency, minimized analyst workload, and maintained a complete audit trail for every generated report.

The Retrieval-Augmented Generation (RAG) pipeline improved factual accuracy by providing relevant regulatory context during narrative generation, while the audit logging mechanism ensured full traceability and regulatory defensibility. Human analyst review through the Streamlit dashboard further enhanced reliability and maintained necessary human oversight in the reporting process.

Overall, the proposed framework provides an effective solution for modern financial institutions handling large volumes of suspicious transaction cases and demonstrates the potential of responsible AI adoption in high-risk regulatory environments.

ACKNOWLEDGMENT

The authors wish to express their sincere gratitude to our project guide and faculty members for their invaluable guidance, insightful feedback, and continuous support throughout the development of this research work. Their encouragement and technical expertise greatly contributed to the successful completion of this project.

Special appreciation is extended to the organizations and researchers working in the fields of Artificial Intelligence, Anti-Money Laundering (AML), and financial compliance for their valuable contributions and publicly available research resources.

Finally, gratitude is extended to all team members, colleagues, and reviewers whose suggestions and support helped improve the quality and implementation of this system.

REFERENCES

[1] Financial Crimes Enforcement Network (FinCEN), “Suspicious Activity Report (SAR) Electronic Filing Requirements,” U.S. Department of the Treasury, 2023.

[2] T. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[3] H. Touvron et al., “Llama 3: Open foundation and fine-tuned chat models,” Meta AI, 2024.
[4] LangChain, “LangChain documentation,” 2024. [Online]. Available: https://www.langchain.com/
[5] ChromaDB, “Chroma vector database documentation,” 2024. [Online]. Available: https://www.trychroma.com/
[6] PostgreSQL Global Development Group, “PostgreSQL documentation,” 2024. [Online]. Available: https://www.postgresql.org/
[7] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019.
[8] OpenAI, “GPT-4 technical report,” arXiv preprint arXiv:2303.08774, 2023.
[9] Pandas Development Team, “Panda’s documentation,” 2024. [Online]. Available: https://pandas.pydata.org/
[10] FastAPI, “FastAPI framework documentation,” 2024. [Online]. Available: https://fastapi.tiangolo.com/
[11] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Advances in Neural Information Processing Systems*, 2020.