PROJECT REPORT

AI-Powered SAR Narrative Generator

A Tech-Agnostic Solution for Auditable and Scalable FinCrime Reporting

Problem Statement: SAR Narrative Generator with Audit Trail
Date of Submission: 17-02-2026

TEAM OVERVIEW

- Team Name: FInTechTacklers
- Team Members:

1. GURU ABIJETH S
2. BHAVITHRAN B K
3. HUDSON DANIEL P
4. GUNALY B
5. ARHI ARHA SUDHAN S

Section 1: Problem Statement

1.1 Limitations of Current Methods

The financial industry currently generates Suspicious Activity Reports (SARs) through two primary, yet flawed, methods:

- Manual Drafting: While nuanced, this process is unscalable, labor-intensive, and prone to subjective inconsistencies that increase regulatory risk.
- Template-Based Automation: Existing legacy systems rely on static templates that are often too rigid to capture the sophisticated and evolving patterns of modern financial crime.

1.2 The "Black Box" and Regulatory Risk

A critical failure in current automation is the "Black Box" nature of narrative generation. These systems often produce final reports without providing a verifiable audit trail.

- Transparency Gap: There is no clear link between specific narrative claims and the underlying raw data.
- Audit Failure: This lack of transparency makes it difficult for institutions to defend the logic of their filings during regulatory audits, leading to significant compliance vulnerability.

1.3 Platform Lock-In

Most current solutions are "locked" into specific, expensive technology stacks. This vendor dependency prevents banks from seamlessly integrating diverse data sources across their infrastructure, creating fragmented intelligence and increasing operational costs.

Section 2: The Solution – Glass-Box Narrative Engine

2.1 Core Mechanism: Retrieval-Augmented Generation (RAG)

The system utilizes a Retrieval-Augmented Generation (RAG) architecture to bridge the gap between automated speed and forensic accuracy. Unlike standard AI, which relies solely on pre-trained knowledge, our engine first retrieves specific, real-time "evidence" from internal bank data—including transaction logs and KYC profiles—before drafting a single word. This ensures every narrative is grounded in factual, current reality rather than AI-generated assumptions.

2.2 Transitioning from "Black Box" to "Glass Box"

To solve the industry’s transparency problem, we replace opaque automation with a "Glass-Box" Audit Trail.

- Data Provenance: Every claim made within the SAR narrative is digitally anchored to its specific data source.
- Traceability: If the report mentions a "series of high-value cash deposits," the system provides a direct link to the specific transaction IDs and timestamps that justify that statement.

- Verification: This creates a transparent map of the AI’s reasoning, allowing human analysts to verify the logic instantly.

2.3 Regulatory Readiness and Defensibility

The ultimate output of the system is a structured, professional narrative that meets strict regulatory standards.

- Standardized Quality: The system eliminates individual writing biases, ensuring a consistent tone and professional vocabulary across all filings.
- Forensic Defensibility: By providing a verifiable audit trail for every sentence, the bank is equipped with a "defensive file" ready for immediate presentation during regulatory examinations or law enforcement inquiries.
- Platform Agnosticism: The solution is designed as a standalone logic layer, allowing it to be integrated into any existing case management system without requiring an overhaul of the bank's core infrastructure.

Section 3: Strategic Impact and Performance Metrics

To measure the success and effectiveness of the system, we focus on three primary impact categories that transform the bank’s compliance posture.

3.1 Regulatory Defensibility: 100% Auditability

The most critical metric for any compliance tool is its ability to withstand regulatory scrutiny.

- The "Glass-Box" Standard: Our system provides a comprehensive audit trail where every sentence in the narrative is digitally mapped to its supporting raw data.
- Metric: We measure "Audit Readiness"—the ability to instantly provide the specific transaction IDs and KYC logs used to justify a filing. This eliminates the risk of "Unsubstantiated Filings" and ensures full forensic transparency for examiners.

3.2 Scalable Consistency: Institutional Uniformity

Manual reporting is often fragmented by the varying skill levels and writing styles of individual analysts.

- Standardized Quality: The engine removes human subjectivity and writing bias, ensuring that every SAR meets a high, professional standard regardless of the investigator assigned.
- Metric: We track "Narrative Variance." By standardizing the logic and vocabulary used in reports, the institution achieves a uniform compliance output that improves the bank’s reputation with law enforcement and regulators.

3.3 Operational Efficiency: Throughput and TAT

The system fundamentally changes the economics of the compliance department by shifting the analyst's role from Writer to Editor.

- TAT Reduction: We measure the "Turnaround Time" (TAT) from the initial alert trigger to the completion of a regulator-ready draft.

- Metric: Success is defined by a significant reduction in manual drafting hours, allowing the existing workforce to handle a higher volume of alerts without a corresponding increase in headcount or operational costs.

Section 4: Key Assumptions and Operational Constraints

4.1 Functional Assumptions

To ensure the successful deployment and accuracy of the narrative engine, the following operational conditions are assumed:

- Data Availability and Access: The system assumes secure connectivity to the institution's internal data environment (via REST APIs or SQL databases). The input data is assumed to be in a structured or semi-structured format (e.g., JSON, CSV, or relational tables) containing transaction logs and customer profiles.
- Human-in-the-Loop (HITL) Oversight: The system is designed as a sophisticated drafting tool, not a fully autonomous filing agent. It is assumed that a qualified human analyst will perform a final review, edit the AI-generated draft where necessary, and take legal responsibility for the final submission to regulators.

4.2 Critical Constraints

The solution operates under specific technical and regulatory boundaries:

- Data Privacy and Security: Due to the sensitivity of Personally Identifiable Information (PII) in banking, the system must operate within the institution’s secure perimeter. No data can be used to train public models, and all processing must comply with local data protection laws (e.g., GDPR, CCPA).
- Zero-Hallucination Mandate: Unlike creative AI applications, this system is constrained to a "Data-Only" logic. If a required data point is missing from the input, the system is programmed to flag the missing information rather than "hallucinating" or inferring details.
- Regulatory Alignment: The generated output is strictly constrained to follow the specific terminology, "Red Flag" categories, and formatting standards required by financial intelligence units (such as FinCEN or the FCA).

Section 5: Implementation Strategy and Scalability

5.1 Seamless Implementation: The "Plug-and-Play" Approach

The system is designed for rapid deployment with minimal disruption to existing banking operations.

- API-First Integration: Built as a modular engine, the solution functions as a "Zero-Displacement" plugin. It integrates directly into legacy Case Management Systems (CMS) via RESTful APIs, allowing institutions to modernize their reporting without replacing multi-million dollar infrastructure.

- Data-Agnostic Architecture: The engine is decoupled from specific database vendors. It can ingest data from any source—SQL, Cloud Storage, or legacy CSV exports—ensuring compatibility with any bank's current technology stack.
- Human-Centric UX: The interface is designed as an intuitive "Editor" tool. By mimicking familiar word-processing workflows, it ensures that analysts can adopt the system immediately with minimal technical training.

5.2 Engineered for Scalability

The architecture is built to grow alongside the institution's alert volume and geographic footprint.

- Elastic Parallel Processing: Utilizing a cloud-native, containerized architecture (Docker/Celery), the system can process 10 or 10,000 narratives simultaneously. This horizontal scaling ensures that performance remains consistent even during sudden surges in suspicious activity alerts.
- Logic-Layer Adaptability: Unlike hard-coded templates, our system uses a decoupled Logic Layer. When AML laws or "Red Flag" guidelines change, the system prompt/logic is updated once, and the change is instantly applied to all future reports across the entire institution.
- Global Regulatory Readiness: The system is built for international deployment. It can be localized for different languages and reconfigured for specific international regulatory formats (e.g., FinCEN in the US vs. FCA in the UK) with minimal architectural adjustment.

Section 6: Technology Stack and Rationale

The selection of our technology stack was driven by three core requirements: Data Privacy, Processing Speed, and System Traceability.

6.1 AI Reasoning and Backend

- Llama 3 (via Ollama): Chosen as the core Large Language Model for its high reasoning capabilities and the ability to be deployed locally. This ensures that sensitive bank data never leaves the institution's secure perimeter, maintaining strict privacy compliance.
- LangChain: Functions as the orchestration layer to manage the Retrieval-Augmented Generation (RAG) logic. It chains the data retrieval process with the LLM, ensuring the model only writes based on provided "evidence."
- ChromaDB: A high-performance vector database used to store and retrieve transaction context. It is the key to our "Glass-Box" audit trail, allowing the system to pinpoint and cite the exact data points used in the narrative.
- FastAPI: A modern, high-performance web framework used for the backend API. Its asynchronous capabilities allow for fast, non-blocking communication between the data sources and the AI engine.

- Celery & Redis: Implemented as a distributed task queue and message broker. Since AI narrative generation is a heavy process, these tools handle the workload in the background, ensuring the user interface remains responsive.
- PostgreSQL: The primary relational database for storing persistent metadata, user logs, and finalized reports, ensuring long-term data integrity and reliability.

6.2 Frontend and Infrastructure

- Tailwind CSS: A utility-first CSS framework used to build a clean, professional, and responsive analyst dashboard. It allows for rapid development of an intuitive "Editor" interface.
- Docker & Docker Compose: Used for containerization, ensuring the entire application is platform-agnostic. This allows the solution to be deployed consistently across any environment, from local developer machines to massive bank servers.
- Nginx: Serves as a high-performance reverse proxy and load balancer. It manages secure traffic flow and improves the overall security and stability of the web service.