Suspicious Activity Report (SAR) Generator using Gen-AI

School of Business
THE GEORGE WASHINGTON UNIVERSITY

About us & Meet the Team

Who We Are: We are a team of Master of Science in Business Analytics (MSBA) students from George Washington University collaborating with the Agricultural Bank of China, New York Branch, on a practicum project to enhance compliance efficiency and reduce costs.

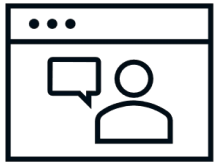
Team Members

1- Chao Hu

2- Jeff Mathew Sam

3- Mohanad AlKhalaf

4-Nour AlZaid



Content

Background and Context

01

Background and Context

What is Money Laundering (ML)?

Money laundering transforms illegally obtained money into seemingly legitimate funds through a three-stage process:

- **Placement**: Introducing funds into the financial system
- **Layering**: Creating complex transaction trails to hide origins
- **Integration**: Merging laundered money into legitimate economy

Transaction Monitoring

Software scans financial transactions to detect unusual patterns.

Red Flags: Large amounts, rapid transfers, or high-risk countries.

Alerts: Generated when thresholds are exceeded; reviewed by analysts.

Alert Investigation:

Initial Review: Analysts assess alerts based on transaction details.

Escalation: Suspicious cases are escalated to SARs.

Outcome: Decide if activity is suspicious or false positive

Background and Context

Suspicious Activity Reports (SARs)

If financial institutions identify activities that seem unusual or suspicious, they are required to file SARs with regulatory authorities like FinCEN.

SARs document:

- Unusual transactions that may indicate financial crimes.
- Patterns suggesting money laundering or fraud.
- Activities potentially linked to terrorism financing.

AML requirements & Current Process

AML Requirements for Banks

- KYC
  - Customer demographic, socio-economic profile data in Database
- Transaction monitoring System
  - Reports suspected violations of law or suspicious activity
  - Required by financial institutions under Bank Secrecy Act (BSA)
    - TD Bank
  - Submitted to FinCEN for suspected money laundering fraud
  - Flags suspicious transaction
- Alert Narrative

Current Process:

- Alert Process
  - Transaction Monitoring System Generates Alerts
  - Analyst Reviews alert, KYC Verification
  - Research & Documentation
  - Writes Alert Narrative
- SAR Filing -
  - 30-Day window, Electronic Filing
  - Five Key Components of which SAR Narrative is the last
- Two Types of Narratives
  - Alert Narrative
  - SAR Narrative
  - Key Differences
    - Purpose
    - Content Level
    - Information Sensitivity
    - Target Audience

Executive Summary

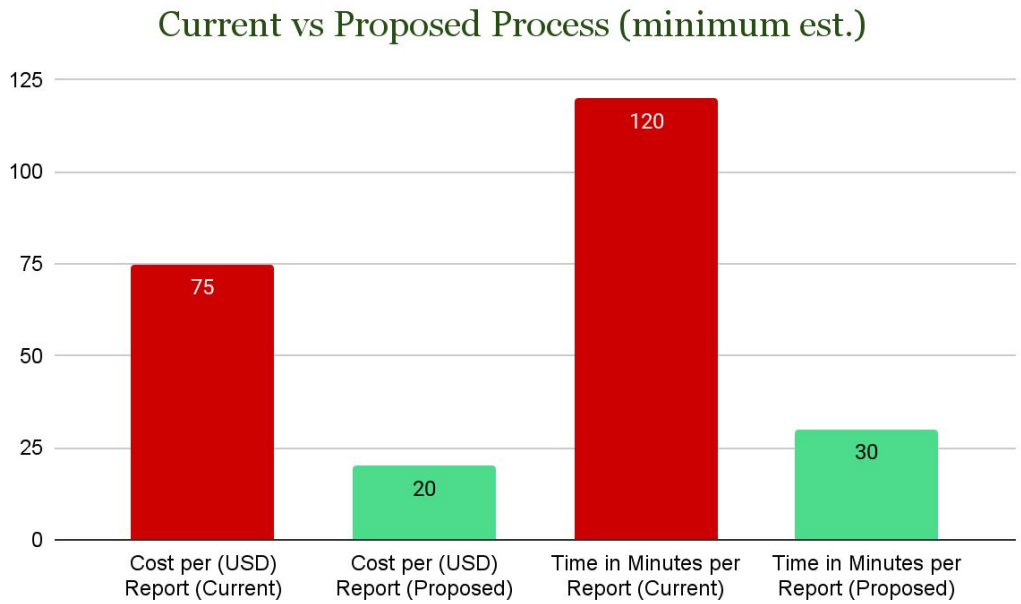
02

Executive Summary

Challenge:

- **Time-Intensive and Costly:** Manual SAR filing demands significant human time and effort, driving up costs and inefficiencies.
- **Data Privacy Risks:** Cloud-based AI poses risks to sensitive customer data, increasing breach potential,

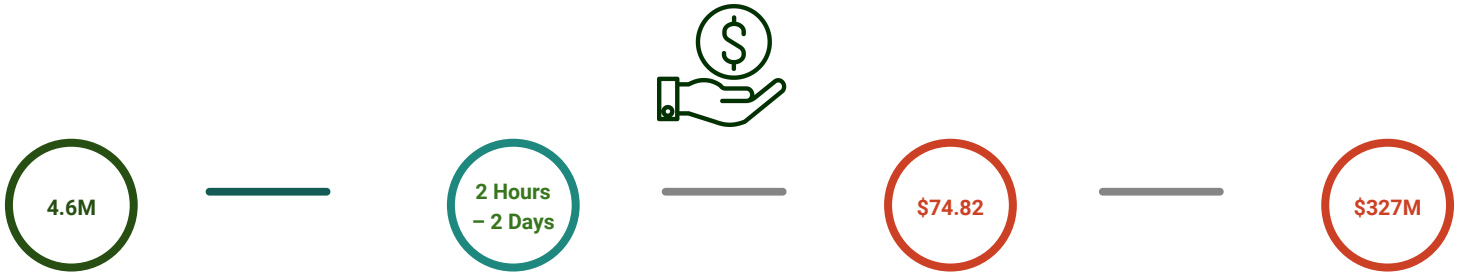
Solution: Deploy a secure, locally-hosted **Generative AI (GenAI)** system to automate SAR generation, reducing costs and ensuring compliance with strict data privacy standards.



Problem Understanding And Project Objectives

03

Problem Understanding



SAR Filing Volume

Represents the total number of SARs filed across the industry.

Time to File Each SAR

Represents the average time burden for preparing and filing each SAR.

Cost per SAR

Highlights the cost associated with each SAR based on the time burden and hourly rate.

Industry Annual Cost

Represents the total annual cost for filing SARs based on FinCEN's estimates.

Critical Data Risks and Local Need

SARs contain sensitive customer information (e.g., names, tax IDs), making cloud-based solutions risky. To address these risks, the client requires a secure, locally deployed system to protect data and ensure compliance.

Project Objectives

1. Identify and choose the optimal LLM model for local deployment.
2. Performance Enhancement through integration with LlamaIndex and using RAG to search relevant document
3. Set up a local MySQL or PostgreSQL server database with Jupyter Notebook or Google Colab
4. The LLM should extract data from this database to create detailed SAR narratives.
5. Format and Design a Jupyter Notebook or Google Colab format for clear SAR narrative presentation.
6. Analyze hallucination occurrences in narratives and how to mitigate them.
7. Fine-tuning the LLM model for optimal narrative generation.

AI Technologies Overview

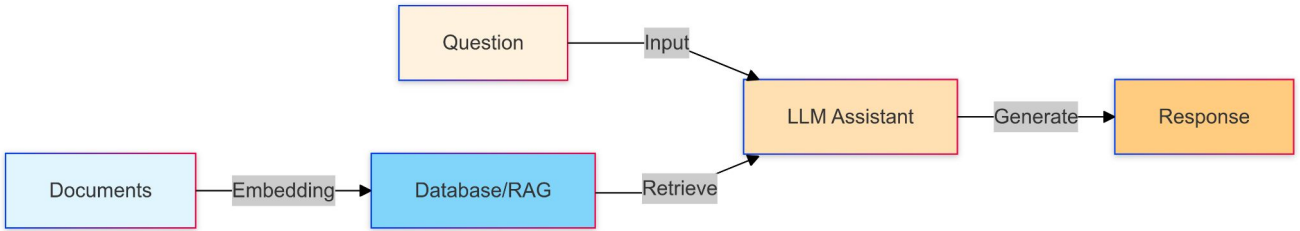
04

AI Technologies Overview

Large Language Models (LLMs): It is a computer program that can understand and write text to mimic humans. They are trained on much information and can help create reports like SARs by writing clear and accurate details. Some models, like LLaMA and Mistral, are better suited for different tasks.

RAG (Retrieval-Augmented Generation): This is a way to make the LLM's smarter by letting it pull extra information from files or databases while writing. It ensures the reports include all the right details and follow the rules, making the process faster and more accurate.

Embedding Models: These models help the computer find the right information by turning words into numbers that it can quickly compare. This makes it easier to find the best data for writing reports.



AI Technologies Overview

Problem

Cost

Manual

Cloud

Security

Solution

LLm

RAG

Embedding

Secure

Output

Automation

Save

Time

Comply

Solution overview process

05

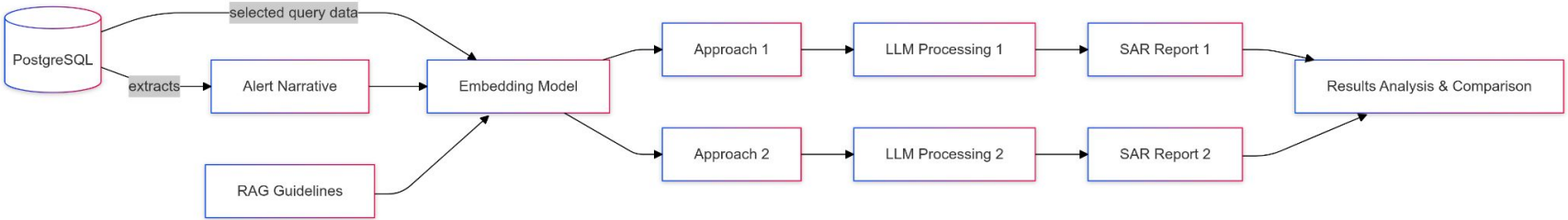
Solution Overview Process

Approach 1 (Summary)

This approach employs Mistral Nemo LLM and an embedding model to generate SAR narratives. The process begins by connecting to a PostgreSQL database to gather relevant information from queries. The system then processes an Alert Narrative PDF, splitting it into logical sections using a parser tool in Llama index. The SAR generation follows a five-paragraph structure: an introduction (static), transaction details (dynamic), customer KYC and research findings (dynamic), SAR recommendations (dynamic), and a conclusion (static), which are generated using Zero-shot learning. To minimize hallucinations and optimize performance, only the dynamic sections utilize Retrieval-Augmented Generation (RAG) through separate vector indices.

Approach 2 (Summary)

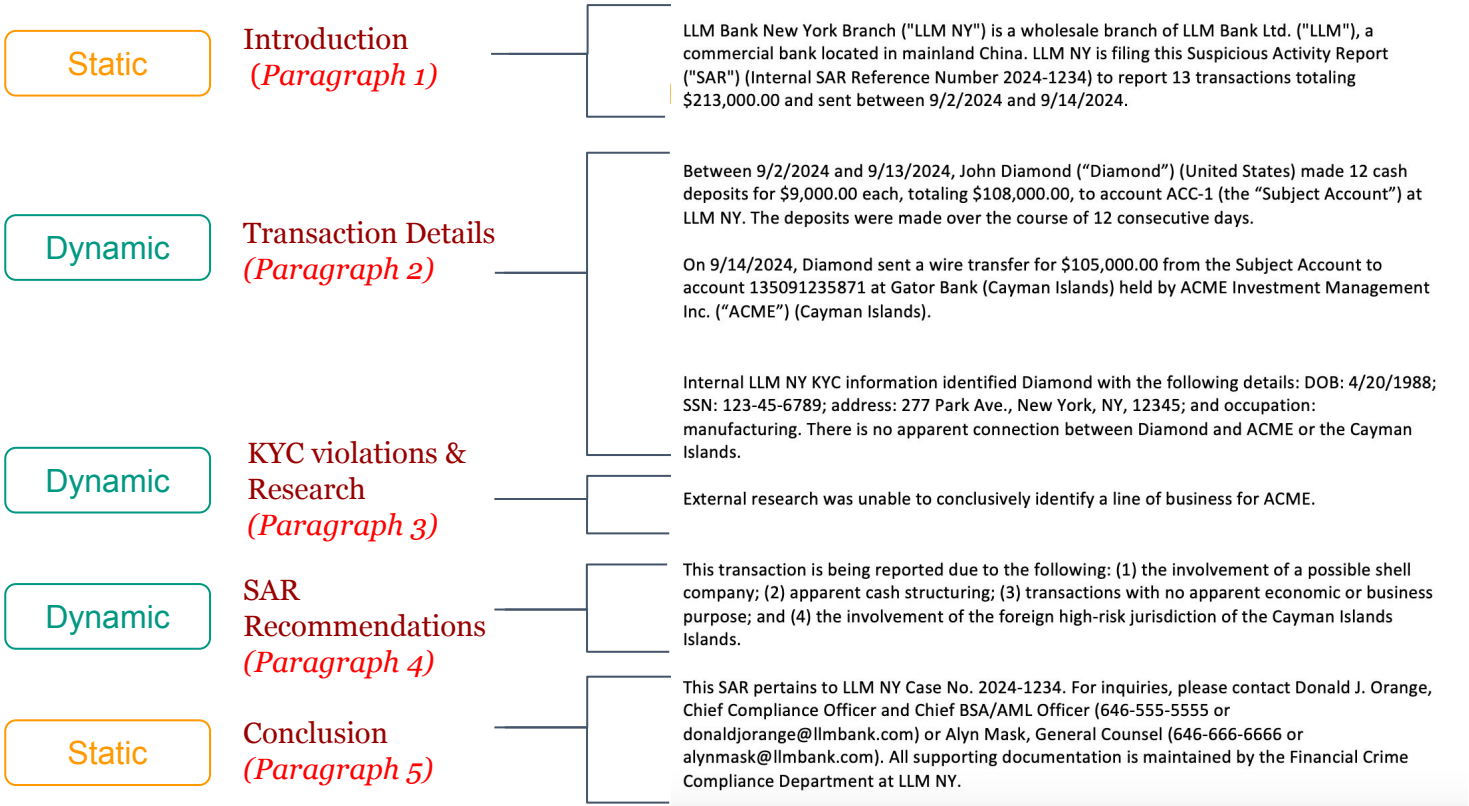
This approach integrates structured PostgreSQL transaction data and unstructured PDF guidelines to generate structured SAR narratives. It uses an embedding model for text representation and a re-rank model to prioritize relevant information. Zero-shot techniques provide exact outputs, while chain-of-thought prompting ensures logical reasoning to guide narrative construction. The output is structured into SAR sections: Introduction, Customer, Patterns, Violations, and Conclusion.



Methodologies

06

SAR structure



Approach 1

Splitting document into **6 sections** with 'Sentence Splitter'

ALERT NARRATIVE
Alert #: A-1 **Create Date:** 9/30/2024

Focal Entity: John Diamond
CIN: C-1
Review Scope: 9/2/2024 – 9/14/2024

Determination / Rationale:
Based on a review of internal and external sources, the reviewed transactions appear to potential suspicious.

Cash Structuring $10k
Rapid Movements of Funds
Large Wire to High Risk Jurisdiction

The customer made 12 cash deposits for $9,000.00 each, totaling $108,000.00 over the course of 12 consecutive days between 9/2/2024 and 9/13/2024. According to KYC information, the customer is employed in the manufacturing industry, which is not a cash-intensive business and investigation of internal and external sources did not identify a legitimate source of funds for these cash deposits. On 9/14/2024, the customer then sent a wire transfer for $105,000.000 to ACME Investment Management in the Cayman Islands. The customer's KYC information does not indicate any apparent connection between either ACME Investment Management or the Cayman Islands.

A SAR filing is recommended for the following reasons:

- The customer apparently made 12 structured cash deposits for $9,000 each over 12 consecutive days without a legitimate source of funds.
- Shortly after make the cash deposits, the customer initiated a wire transfer to an unrelated company with which the customer has no apparent connection.
- There is no apparent lawful economic purpose for the customer's activity.
- The involvement of the high risk jurisdiction of the Cayman Islands.

ALERT NARRATIVE
Alert #: A-1 **Create Date:** 9/30/2024

Focal Entity: John Diamond
CIN: C-1
Review Scope: 9/2/2024 – 9/14/2024

Determination / Rationale:
Based on a review of internal and external sources, the reviewed transactions appear to potential suspicious.

Cash Structuring $10k
Rapid Movements of Funds
Large Wire to High Risk Jurisdiction

Transaction details:
The customer made 12 cash deposits for $9,000.00 each, totaling $108,000.00 over the course of 12 consecutive days between 9/2/2024 and 9/13/2024. According to KYC information, the customer is employed in the manufacturing industry, which is not a cash-intensive business and investigation of internal and external sources did not identify a legitimate source of funds for these cash deposits. On 9/14/2024, the customer then sent a wire transfer for $105,000.000 to ACME Investment Management in the Cayman Islands. The customer's KYC information does not indicate any apparent connection between either ACME Investment Management or the Cayman Islands.

SAR Recommendation:
A SAR filing is recommended for the following reasons:

- The customer apparently made 12 structured cash deposits for $9,000 each over 12 consecutive days without a legitimate source of funds.
- Shortly after make the cash deposits, the customer initiated a wire transfer to an unrelated company with which the customer has no apparent connection.
- There is no apparent lawful economic purpose for the customer's activity.
- The involvement of the high risk jurisdiction of the Cayman Islands.

Approach 1

Static

Prompt 1

LLM Bank New York Branch ("LLM NY") is a wholesale branch of LLM Bank Ltd. ("LLM"), a commercial bank located in mainland China. LLM NY is filing this Suspicious Activity Report ("SAR") (Internal SAR Reference Number 2024-1235) to report seven (7) transactions totaling $7,227,504.80 and sent between 9/2/2024 and 9/15/2024.

Prompt 5

This SAR pertains to LLM NY Case No. 2024-1235. For inquiries, please contact Donald J. Orange, Chief Compliance Officer and Chief BSA/AML Officer (646-555-5555 or donaldjorange@llmbank.com) or Alyn Mask, General Counsel (646-666-6666 or alynmask@llmbank.com). All supporting documentation is maintained by the Financial Crime Compliance Department at LLM NY.

Dynamic

ALERT NARRATIVE
Alert #: A-1 **Create Date:** 9/30/2024

Focal Entity: John Diamond
CIN: C-1
Review Scope: 9/2/2024 – 9/14/2024

Determination / Rationale:
Based on a review of internal and external sources, the reviewed transactions appear to potential suspicious.

Cash Structuring $10k
Rapid Movements of Funds
Large Wire to High Risk Jurisdiction

Transaction details:

Prompt 2

The customer made 12 cash deposits for $9,000.00 each, totaling $108,000.00 over the course of 12 consecutive days between 9/2/2024 and 9/13/2024. According to KYC information, the customer is employed in the manufacturing industry, which is not a cash-intensive business and investigation of internal and external sources did not identify a legitimate source of funds for these cash deposits. On 9/14/2024, the customer then sent a wire transfer for $105,000.000 to

Prompt 3

ACME Investment Management in the Cayman Islands. The customer's KYC information does not indicate any apparent connection between either ACME Investment Management or the Cayman Islands.

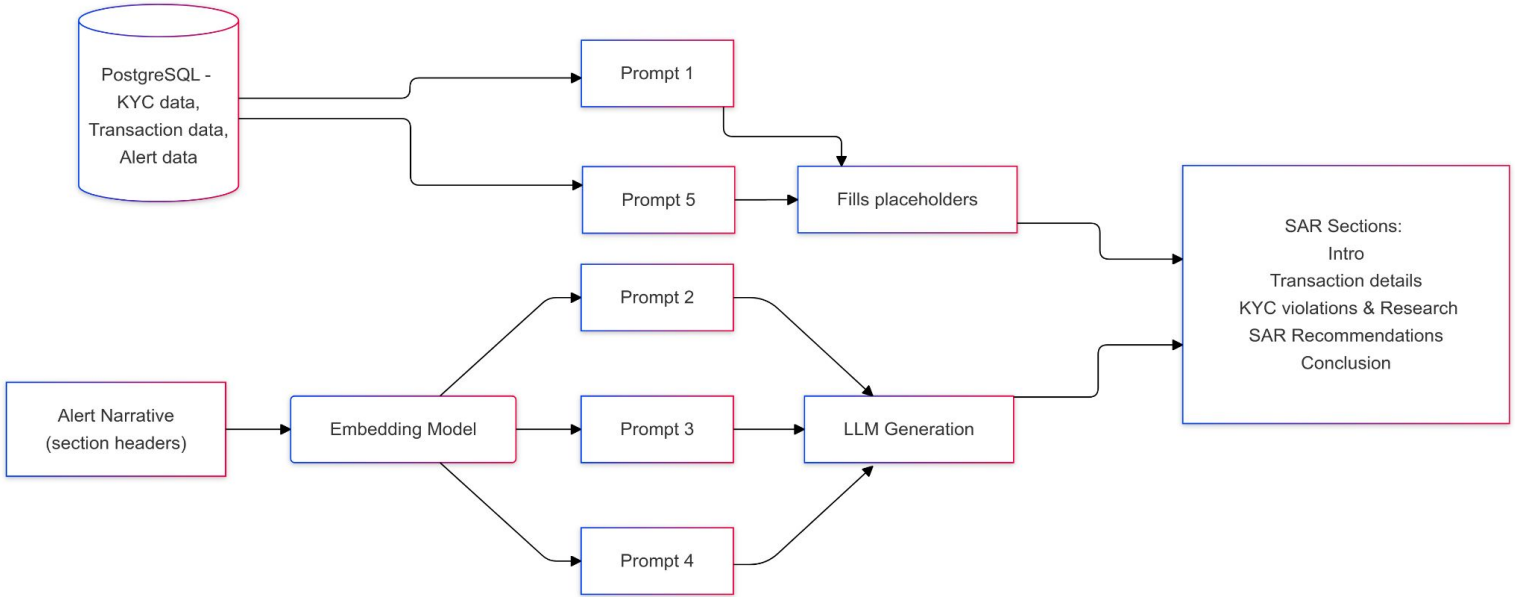
SAR Recommendation:

Prompt 4

A SAR filing is recommended for the following reasons:

- The customer apparently made 12 structured cash deposits for $9,000 each over 12 consecutive days without a legitimate source of funds.
- Shortly after make the cash deposits, the customer initiated a wire transfer to an unrelated company with which the customer has no apparent connection.
- There is no apparent lawful economic purpose for the customer's activity.
- The involvement of the high risk jurisdiction of the Cayman Islands.

Approach 1


PostgreSQL - KYC data, Transaction data, Alert data
Prompt 1
Prompt 5
Fills placeholders
SAR Sections:
Intro
Transaction details
KYC violations & Research
SAR Recommendations
Conclusion
Prompt 2
Alert Narrative (section headers)
Embedding Model
Prompt 3
LLM Generation
Prompt 4

Approach 1 : Taking into consideration 5 rules

Data Preparation

PDF:
Prepare Alert Narrative by splitting doc into **6 sections** it using 'Sentence Splitter' tool in Llama index.

PostgreSQL Database:

Extracting & storing Placeholders for:

Prompt 1

Prompt 5

Embedding and Prompting

Embedding Model:
Text in are converted into embeddings using 'sentence-transformers/all-mpnet-base-v2'

Prompting:
Zero-Shot Learning

Prompt 2

Prompt 3

Prompt 4

Combining Prompts

Prompt 1

Prompt 2

Prompt 3

Prompt 4

Prompt 5

SAR Generation

The *Mistral Nemo 12B* model processes the combined prompts to generate the SAR report.

The output is structured into five key sections:

1. *Introduction*
2. *Transaction Details*
3. *KYC Violation & Research*
4. *SAR Recommendations*
5. *Conclusion*

Rules :

1. *Cash Structuring*
2. *Rapid Movement of Funds*
3. *Large Wire to High Risk Jurisdiction*
4. *Concentration Account*
5. *New Account*

Approach 1 :Results

Original SAR

LLM Bank New York Branch ("LLM NY") is a wholesale branch of LLM Bank Ltd. ("LLM"), a commercial bank located in mainland China. LLM NY is filing this Suspicious Activity Report ("SAR") (Internal SAR Reference Number 2024-1234) to report 13 transactions totaling $213,000.00 and sent between 9/2/2024 and 9/14/2024.

Between 9/2/2024 and 9/13/2024, John Diamond (“Diamond”) (United States) made 12 cash deposits for $9,000.00 each, totaling $108,000.00, to account ACC-1 (the “Subject Account”) at LLM NY. The deposits were made over the course of 12 consecutive days.

On 9/14/2024, Diamond sent a wire transfer for $105,000.00 from the Subject Account to account 135091235871 at Gator Bank (Cayman Islands) held by ACME Investment Management Inc. (“ACME”) (Cayman Islands).

Internal LLM NY KYC information identified Diamond with the following details: DOB: 4/20/1988; SSN: 123-45-6789; address: 277 Park Ave., New York, NY, 12345; and occupation: manufacturing. There is no apparent connection between Diamond and ACME or the Cayman Islands.

External research was unable to conclusively identify a line of business for ACME.

This transaction is being reported due to the following: (1) the involvement of a possible shell company; (2) apparent cash structuring; (3) transactions with no apparent economic or business purpose; and (4) the involvement of the foreign high-risk jurisdiction of the Cayman Islands Islands.

This SAR pertains to LLM NY Case No. 2024-1234. For inquiries, please contact Donald J. Orange, Chief Compliance Officer and Chief BSA/AML Officer (646-555-5555 or donaldjorange@llmbank.com) or Alyn Mask, General Counsel (646-666-6666 or alynmask@llmbank.com). All supporting documentation is maintained by the Financial Crime Compliance Department at LLM NY.

Generated SAR

Introduction	LLM Bank New York("LLM NY") is a wholesale branch of LLM Bank Ltd, a commercial bank located in mainland China. LLM NY is filing this Suspicious Activity Report ("SAR") (Internal SAR Reference Number) A-1, to report 13 transactions totaling $213000.00 between 09/02/2024 and 09/14/2024.
Transaction Details	Between 9/2/2024 and 9/13/2024, John Diamond made 12 cash deposits, each for $9,000.00, totaling $108,000.00. Subsequently, on 9/14/2024, John Diamond sent a wire transfer of $105,000.00 to ACME Investment Management in the Cayman Islands.
KYC violations & Research	LLM NY KYC information identifies John Diamond as employed in the manufacturing industry. External research was unable to conclusively identify a line of business for ACME Investment Management.
SAR Recommendations	These transactions are being reported due to the following: (1) Apparent cash structuring, (2) Apparent rapid movements of funds, (3) Apparent large wire to a high risk jurisdiction, (4) Transactions with no apparent economic or business purpose.
Conclusion	This SAR pertains to LLM NY Case No A-1. For inquiries, please contact Donald J. Orange Chief Compliance Officer and Chief BSA/AML Officer (646-555-5555) or donaldjorange@llmbank.com or Alyn Mask, General Counsel (646-555-5555) or alynmask@llmbank.com. All supporting documentation is maintained by the Financial Crime Compliance Department at LLM NY.

Approach 1

Observations

- To minimizing hallucinations - specific instructions in the prompts. This also helped achieve the **formatting** we wanted from the LLM. For example:
  - "…Follow the instructions in Requirements"
  - "Do not add your own knowledge"
  - "…write logical sentences with a natural flow"
  - "Use exact formatting:
    - Number each reason with parentheses: "(1)", "(2)", etc.
    - Start each rule violation with "Apparent" "
  - Using "Summarize" gave the output closest to what the client requested, but using "Write a comprehensive summary" and "Generate a comprehensive summary" performed poorly.

Recommendation For Approach 1

1. Further Fine-tuning LLM parameter
   - Examine how to leverage additional parameters of the RAG query engine
     - Using *context_template*
     - Using *filters*

2. Evaluate and experiment with different combination of rules in Alert Narratives, along with Customer and Alert numbers, using the same criteria prompts.

3. Examine ways to reduce the context length for RAG prompt without altering the output

Approach 2

Data Preparation

PDF Guidelines: Pre-defined documents segmented into sections like Introduction, Customer Information, Patterns, and Violations provide the structural template for SARs.

PostgreSQL Database: Stores structured data, such as transaction records, for efficient querying to support SAR generation.

Embedding and re-rank

Embedding Model: Text and query data are converted into embeddings using the all-MiniLM-L6-v2 model for efficient representation.

Cross-Encoder Re-Ranking: Embeddings are refined using the ms-marco-MiniLM-L6-v2 cross-encoder to ensure relevance and precision.

Prompting Strategies

Zero-Shot (Exact Format + RAG): Uses a single prompt to generate each SAR section.

Chain of Thought (Exact Format + RAG): Breaks the generation process into multiple steps, ensuring detailed and structured responses.

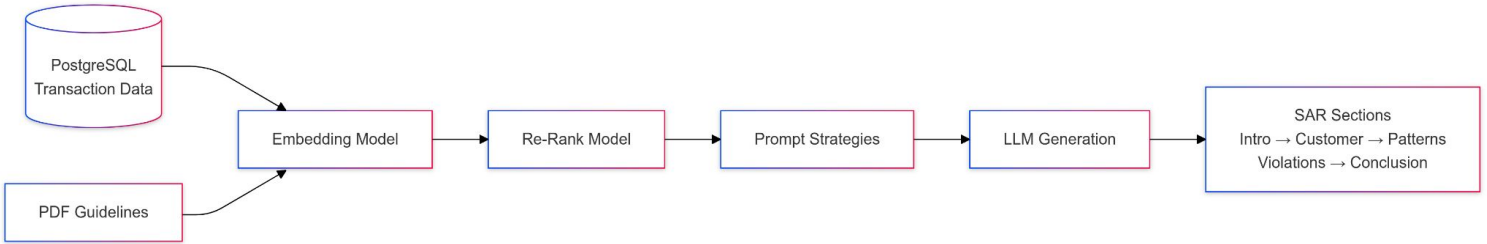
Chain of Thought (Guided Format + RAG): Sequential prompts guide the model for comprehensive and formatted responses.

Text Generation

The *LLaMA 3.2-3B* model processes the refined context and prompts to generate the SAR report.

The output is structured into five key sections:

1. *Introduction*
2. *Customer Information*
3. *Patterns*
4. *Violations*
5. *Conclusion*



Approach 2

I recommend adopting the Chain-of-Thought-guided LLM with a score of 7 out of 10 because, even though the zero-shot model scored 8 out of 10, the CoT-guided approach better aligns with human-like analytical reasoning in SAR generation.

Output Example:

LLm Generated · Query and Rag · LLm Static Text

Introduction

LLM Bank New York Branch (LLM NY), a wholesale branch of LLM Bank Ltd. (China-based), is filing this SAR (Internal SAR Reference Number A-6) to report a series of structured cash transactions (SAR type: Cash Structured Transactions) totaling $80,000 conducted by Mickel Angelo between 2024-11-01 and 2024-11-10.

Customer Information

During this period, Mickel Angelo, a resident of US, made 10 deposits of $8,000 each into account ACC-6 at LLM NY. These deposits were made consecutively over 10 days, suggesting an effort to structure cash deposits below reporting thresholds.

Patterns

According to LLM NY's KYC information, Mickel Angelo's occupation is in Car Sales Commission, a sector not typically associated with cash-intensive transactions. No legitimate source of funds was identified for these deposits, raising concerns regarding the origin of the deposited funds.

Violations

This SAR is being filed solely on the grounds of apparent cash structuring, with indications that these deposits may be intended to evade regulatory reporting requirements.

Conclusion

This SAR pertains to LLM NY Case No. 2024-1234. For inquiries, please contact Donald J. Orange, Chief Compliance Officer and Chief BSA/AML Officer at 646-555-5555 or donaldjorange@llmbank.com, or Alyn Mask, General Counsel, at 646-666-6666 or alynmask@llmbank.com. All supporting documentation is maintained by the Financial Crime Compliance Department at LLM NY.

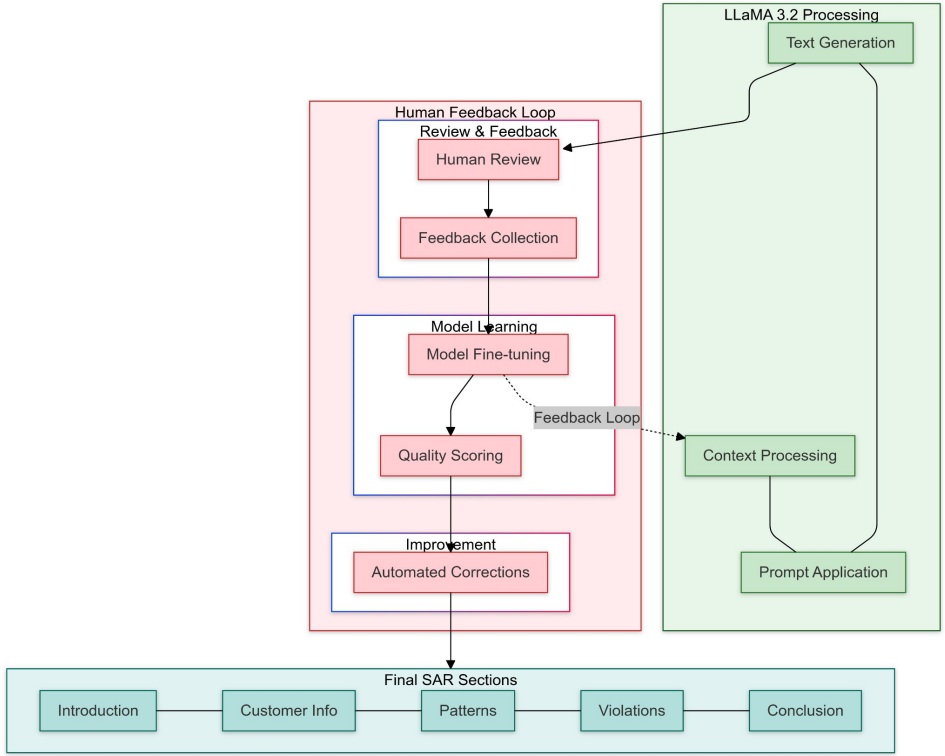
Approach 2

Recommendation For Approach 2

Human Feedback with Reinforcement Learning

This process combines AI and human oversight to generate high-quality Suspicious Activity Reports (SARs) for the bank. AI (LLaMA 3.2) drafts SAR sections like "Customer Info" and "Violations," while banking experts review and provide feedback. The AI learns from this feedback, improving its accuracy and compliance over time.

The result is faster, more accurate, and compliant SAR generation, reducing manual effort and enhancing risk management.



Results and Performance Analysis

07

Results and Performance Analysis

Based on the client's insights, the key recommendation is as follows:

The preferred model depends on the feasibility of implementing distinct Models/RAGs for each individual rule.

Approach 1 (Preferred Model for the Large banks)

The **Mistral model** is more effective as a versatile model, capable of generating SAR narratives across various rules. The choice of which model to use largely hinges on the practicality of deploying separate models or RAGs for each rule. If that isn't practical, the Mistral model appears to be the better option. For a bank with numerous rules, the Mistral model would likely be the preferred choice.

Approach 2 (Preferred Model for the Branch)

The **LLaMA model** is better suited for this specific branch, as it performs exceptionally well on the **cash structuring rule** and can be independently configured with a Retrieval-Augmented Generation (RAG) system for each rule. Given the branch has a manageable number of rules, this approach is feasible and ensures optimal performance.

In conclusion, The Agricultural Bank of China see both models as viable proofs of concept for use in the future.

Risk Assessment

08

Risk Assessment

Data Privacy

- Training Phase
- Interaction Phase
- Running-time Phase

Omitted information

- Limitation of training data set
- Example: External information

Hallucination

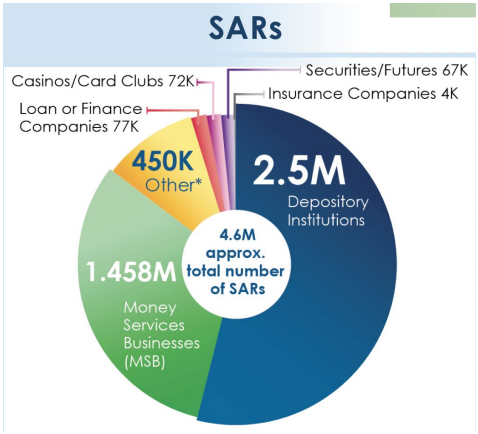
- Inaccurate and irrelevant
- Three ways: RAG, Querying with LLM, and Directly querying

Business Value Proposition

09

Business Value Proposition

[FinCEN Bank Secrecy Act Data - FY 2023]



- Reduction in SAR expenses
- Automating a time-consuming task
- Leveraging LLMs for their narrative generation capabilities

FinCEN Total number of SARs in FY 2023 - by Banks	Cost – per hour for SAR expenses($)			
2,500,000	37			
ABC NY- No. of SARs filed per month	Min Time Taken to write SAR - 1 hour	Max Time taken to write SAR - 2 days (16 working hours)	Min Cost ($)	Max Cost ($)
10,000	1	16	370,000	5,920,000
20,000	1	16	740,000	11,840,000
Larger Banks - JP Morgan				
40,000	1	16	1,480,000	23,680,000

Conclusion & Recommendations

10

Conclusion & Recommendations

- Running larger model on HPC at GWU
- Optimizing prompts
- Further fine-tuning the parameters of the query engine used for RAG
- Fine tuning the embedding in a model to cater to the specific use case
- Running other evaluations and benchmarks besides human evaluation
- Gen-AI Risk Management:
    - Model Risk :Information integrity, confabulations, Data privacy, Information security are risks by LLMs in this use case. How would we evaluate and limits these in production setting?

- Understanding Currency Transaction Reports (CTRs), filing process and report structure

Appendix

11

Appendix

Solutions: Initial Methods tried and tested