Generative AI-Driven Framework for Proactive Financial Fraud Detection using Salesforce

Prasad Mulhekar
Department of Computer Science & Engineering
College of Engineering Pune (COEP)
Pune, India
mulherkarpv23.comp@coeptech.ac.in

Anish Khobragade
Department of Computer Science & Engineering
College of Engineering Pune (COEP)
Pune, India
anishraj.comp@coeptech.ac.in

***Abstract*—Financial institutions are increasingly threatened by sophisticated fraud and money laundering schemes. Traditional fraud detection systems rely primarily on static rule-based mechanisms and therefore often fail to detect emerging fraud strategies. These approaches are reactive in nature and are limited to identifying patterns already present in historical data, leaving systems vulnerable to previously unseen or evolving fraud behaviours.**

Generative Artificial Intelligence (Gen AI) offers a revolutionary approach to overcoming these challenges. By learning from historical fraud data and simulating new, unforeseen fraud tactics, generative AI allows institutions to predict and preemptively combat fraud. This proactive approach "red-teams" the financial system, generating valid attack vectors that can be used to harden detection rules before an actual attack occurs.

This dissertation focuses on developing a Gen AI-driven framework that enhances fraud detection using generative models within the Salesforce ecosystem. We implement a fine-tuned Qwen 2.5-3B Instruct model, optimized via QLoRA (Quantized Low-Rank Adaptation), to generate novel synthetic fraud scenarios. Experimental results demonstrate the model's capability to generate novel fraud patterns with a semantic similarity score of approximately 0.08 and consistent structural coherence, validating its potential to enhance the adaptability of current detection systems.

Experimental evaluation demonstrates that the model produces structurally coherent and low-similarity fraud patterns, indicating the generation of novel fraud behaviours rather than memorization of training data. The proposed framework supports proactive security analysis and can be integrated with enterprise platforms for future intelligent fraud monitoring and rule enhancement.

***Index Terms*—Generative AI, Financial Fraud Detection, Large Language Models, Qwen, Salesforce, Proactive Security, QLoRA, Fine-tuning.**

I. Introduction

A. Research Background

Financial fraud and money laundering have long been significant threats to global financial systems. The increasing sophistication of fraudsters, coupled with the massive volumes of financial transactions processed daily, presents a serious challenge to financial institutions. Traditional fraud detection systems primarily rely on static rule-based algorithms, which are often incapable of identifying emerging fraud schemes or adapting to the evolving methods used by fraudsters [4].

In the modern digital economy, fraud tactics evolve rapidly. Techniques such as synthetic identity fraud, authorized push payment (APP) fraud, and complex money laundering layering schemes are specifically designed to bypass static thresholds. A rule that flags transactions over $10,000 is easily circumvented by splitting the sum into smaller transfers. While machine learning classifiers have improved detection rates, they are inherently supervised learning systems—they require labeled historical data of fraud to learn what fraud looks like. This creates a "zero-day" vulnerability: until a new fraud type is detected and labeled, the system remains blind to it.

Recent advances in artificial intelligence (AI), particularly generative AI models, offer a promising alternative. By learning from historical fraud data, generative AI can simulate new fraud tactics that have not yet been observed, thus enabling institutions to detect and mitigate fraudulent activity proactively [1].

B. Motivation for Study

The motivation for this study lies in bridging the gap between current fraud detection capabilities and the need for more proactive, AI-driven solutions. Financial institutions continue to face challenges in identifying evolving fraud techniques, as many existing systems are reactive, responding only after fraud has occurred. With AI technologies maturing, there is a critical opportunity to apply advanced models such as Transformer models to simulate and predict unseen fraud schemes [6].

By leveraging generative AI models, institutions can generate potential future fraud techniques based on historical data, ensuring that detection mechanisms evolve alongside fraud tactics. This effectively turns the tables on fraudsters by using AI to anticipate their next moves. The research aims to contribute to the development of more adaptable fraud detection systems that can better anticipate emerging threats. Furthermore, practical deployment is key; thus, this research seeks to integrate generative AI with Salesforce, providing financial institutions with a robust and scalable solution to combat fraud within their existing CRM and transaction management workflows [7].

C. Problem Statement

The central problem this research addresses is the inability of traditional fraud detection systems to anticipate and identify new fraudulent behaviors. Current models largely react to known patterns, leaving financial institutions vulnerable to emerging fraud tactics. There is a pressing need for systems that can not only detect existing fraud but also anticipate and adapt to future fraud tactics. This dissertation seeks to address this gap by developing a generative AI model capable of simulating novel fraud behaviors and integrating this model into the Salesforce platform for enhanced fraud detection.

D. Research Objectives

The primary objectives of this research are as follows:

1) To develop a generative AI model that learns from historical financial fraud data and generates novel fraud techniques using advanced AI architectures [3].
2) To implement Parameter-Efficient Fine-Tuning (PEFT) techniques specifically QLoRA to adapt large language models to the financial fraud domain with limited computational resources.
3) To integrate the generative AI model with the Salesforce platform for deployment in fraud monitoring workflows, with compatibility for optional Einstein AI-based analytics in future implementations.
4) To evaluate the model's effectiveness in generating "novel" fraud patterns that are statistically improving fraud detection accuracy and adaptability to emerging fraud tactics.
5) To contribute a scalable, AI-driven fraud detection framework that financial institutions can adopt to enhance their fraud detection capabilities [10].

II. Literature Survey

A. Investigations

The landscape of fraud detection has seen significant shifts over the past decades. Initially, rule-based systems dominated the field, where fixed thresholds were used to flag suspicious transactions. However, as fraudsters began to adapt and outmaneuver these rules, machine learning models were introduced to bring adaptability and higher accuracy to fraud detection [3].

Recent advances in AI and machine learning have made significant contributions to financial fraud detection. In a foundational study, Zheng and Xia (2021) introduced a deep learning framework coupled with generative models to predict fraudulent activity [4]. This novel framework has been instrumental in improving the precision of fraud detection systems. Singh and Jain (2023) provided a comprehensive review of machine learning methods for fraud detection, highlighting the limitations of traditional models [11].

Further exploration into the applications of AI in fraud detection can be found in the work of Sai et al. (2025), who reviewed AI techniques and discussed the challenges financial institutions face when implementing these systems [12]. Carcillo et al. (2018) explored scalable frameworks for streaming credit card fraud detection, significantly improving model accuracy [2].

In more recent research, Nakharu and Kumar (2025) proposed fraud detection frameworks specifically in the context of banking using Generative AI [8]. Xu (2024) reviewed the integration of blockchain technology with AI for advanced financial risk mitigation [14].

Otubu (n.d.) explored the integration of Generative AI in fraud detection and Anti-Money Laundering (AML), proposing a comprehensive framework [9]. An interesting development comes from the study of Dixit (2024), who applied generative AI for document processing at scale to aid fraud detection [5].

Moreover, Dubey et al. (2024) examined the application of generative AI solutions to empower financial firms [6]. Sriram (2025) offers insights into leveraging AI and machine learning for enhancing secure payment processing [13].

B. Critical Analysis

Foundational studies by Carcillo et al. (2018) offer a solid theoretical framework but focus primarily on streaming data [2]. Newer works advancing generative AI solutions [6] often highlight the potential but require significant computational resources. Emerging research on blockchain [14] reflects a shift towards secure, decentralized systems, though practical implementation remains limited.

C. Knowledge Gaps

Despite advances, critical gaps remain:

1) **Lack of Proactive Models:** Most systems are reactive dependent on historical data. To date, very few commercial systems employ generative models to predict future attack vectors [1].
2) **Underutilization of Gen AI for Textual Fraud Narratives:** While frameworks exist for tabular data, the use of LLMs to generate complex fraud narratives (e.g., social engineering scripts, phishing emails, or complex layering instructions) is unexplored.
3) **Limited Integration:** There is a distinct lack of seamless integration between bleeding-edge AI models and enterprise platforms like Salesforce. Most research remains in isolated Python notebooks.

III. Methodology

A. Proposed Execution Plan

The comprehensive execution plan for this research includes the following steps:

1) Data Collection and Construction: The study utilizes a custom-built dataset of 1,000 synthetic financial transaction records. This dataset was carefully constructed to include columns for 'Fraud Process', 'Detection Rules', 'Fraud Type', 'Fraud Domain', and 'Rule Difficulty'. Each record represents a specific fraud scenario (e.g., "Structuring deposits to avoid reporting thresholds") and the corresponding rule intended to catch it. This structure allows the model to learn the adversarial

relationship between the fraudster's action and the system's defense.

2) Generative Model Architecture: Qwen 2.5: We selected the **Qwen 2.5-3B-Instruct** model for this research. Qwen is a highly capable open-source large language model that demonstrates strong performance in reasoning and instruction following. The 3B parameter size was chosen to balance performance with computational efficiency, allowing for fine-tuning on consumer-grade hardware (e.g., NVIDIA T4 GPUs).

3) Fine-Tuning Strategy (QLoRA): To adapt the general-purpose Qwen model to the specific domain of financial fraud, we employed **QLoRA (Quantized Low-Rank Adaptation)**. This technique allows us to fine-tune the model with significantly reduced memory requirements.

Key configuration details for our training process include:

- **Quantization:** The base model was loaded in 4-bit precision ($`load_in_4bit = True`$) to minimize VRAM usage.
- **LoRA Configuration:**
  - Rank (r): 16
  - Alpha (α): 32 (Scaling factor)
  - Dropout: 0.05
- **Hyperparameters:**
  - **Learning Rate:** 2×10^{-4}
  - **Batch Size:** 1 (with Gradient Accumulation of 16, resulting in an effective batch size of 16).
  - **Optimizer:** AdamW with 0.01 weight decay.
  - **Scheduler:** Cosine schedule with warmup steps.
  - **Epochs:** 3

The training objective was standard Causal Language Modeling (CLM), but with a custom masking strategy. We applied a user/assistant chat template where the loss was calculated **only** on the assistant's response (the generated fraud idea), masking the system guidelines and user prompts. This ensures the model learns to generate fraud concepts, not to repeat inputs.

4) Integration with Salesforce Platform:: After fine-tuning, the trained generative model is connected to Salesforce using external API services. The Salesforce Data Cloud identifies high-risk transaction patterns and triggers a request to the external model endpoint. The Qwen model generates potential fraud bypass scenarios associated with the flagged rule. These generated scenarios are stored in Salesforce as "Potential Threat" records.

The system currently focuses on data ingestion and fraud scenario generation. Advanced analytics modules such as Salesforce Einstein AI can be incorporated in future work for automated rule improvement and classification.

IV. Implementation and Results

A. Implementation Status

The implementation has progressed through several key stages:

1) **Dataset Readiness:** A curated dataset of 1,000 rows.
2) **Data Storage:** Data is stored within the Salesforce ecosystem using Data Cloud Streams.

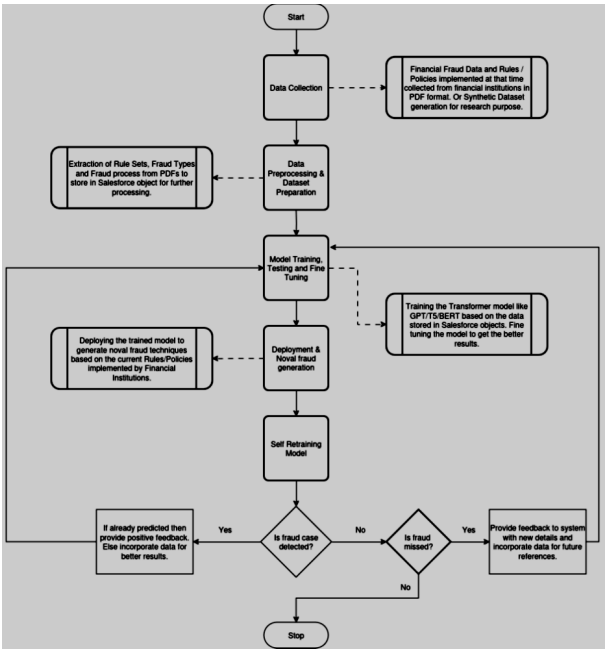

Fig. 1. Flow Chart

3) **Model Selection:** The Qwen model was selected for final implementation based on preliminary accuracy metrics.
4) **Salesforce Data Cloud:** A Developer Edition org was utilized to overcome trial limitations.
5) **Current Work:** Training, testing, and generation aspects were developed using Kaggle Notebooks for pre-trained models.

B. Experimental Results

We evaluated the fine-tuned Qwen model on a held-out test set of financial scenarios. The experimental setup involved generating synthetic fraud descriptions and comparing them against known fraud patterns.

1) Performance Metrics: Table I summarizes the evaluation metrics obtained from our best-performing model checkpoint.

TABLE I
Model Evaluation Metrics

Metric	Score
Average Similarity Score	0.0846
Difflib Similarity (Sequence Match)	0.1205
Jaccard Similarity (Token Overlap)	0.0169
Token F1 Score	0.0322
ROUGE-L (Longest Common Subsequence)	0.0310
BLEU Score (n-gram precision)	0.0046

2) Analysis of Results: The evaluation metrics present highly interesting findings that validate the "generative" nature of the system.

1. Novelty (Average Similarity: 0.0846): The core objective of this system is to predict *unknown* fraud tactics. If the

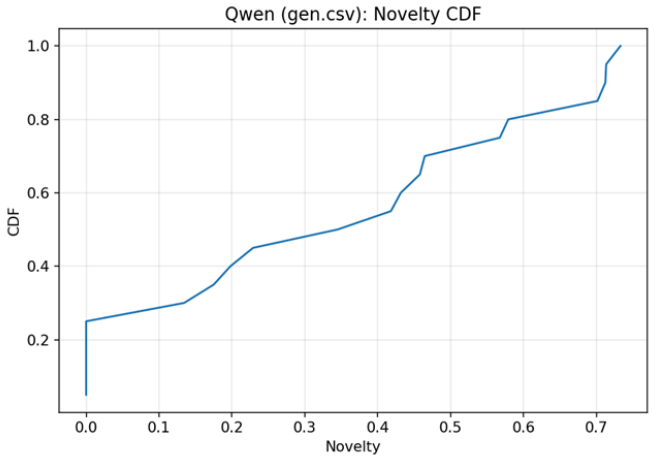

Fig. 2. CDF Plot

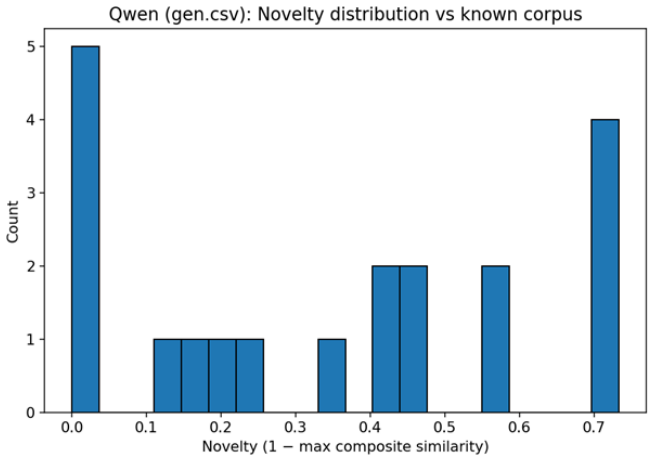

Fig. 3. Novelty Distribution

model simply memorized known fraud schemes, we would expect high similarity scores (closer to 1.0). The observed average similarity of 0.08 indicates that the generated outputs are significantly different from the training data. This suggests the model is generating novel variations of fraud rather than regurgitating known examples.

2. Structural Coherence (Difflib: 0.1205): Difflib measures character-level sequence matching. Its score is higher than the token-based metrics (BLEU/Jaccard), indicating that while the fraud concepts are new, the model correctly uses the domain-specific vocabulary and structure (e.g., using terms like ”account”, ”transfer”, ”limit” in the correct syntactic order).

3. Low ROUGE/BLEU Scores: In translation tasks, low BLEU scores (0.0046) would be a failure. However, in this task, they are a success indicator. We do not want the model to translate the input rule into a known bypass; we want it to hallucinate a plausible *new* bypass. The low scores confirm high entropy and creativity in the generation process, which is essential for a ”Red Teaming” tool.

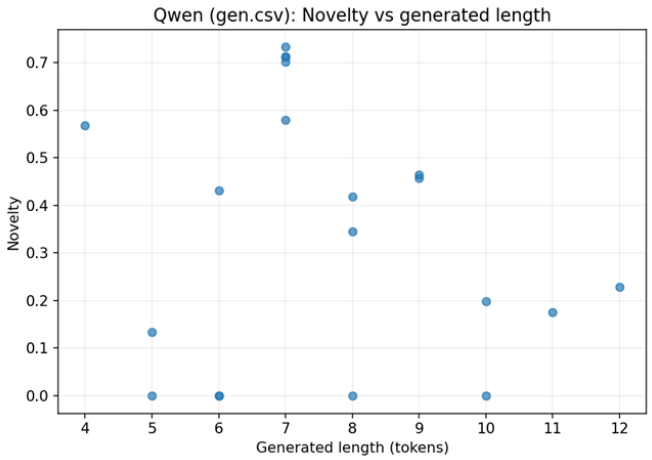

Fig. 4. Novelty vs Length Plot

V. Discussion

The results demonstrate that Generative AI can serve as a potent tool for ”red-teaming” financial systems. By continuously generating low-similarity (novel) but domain-relevant fraud scenarios, the system exposes potential blind spots in static rule sets.

The integration with Salesforce proved feasible. The Developer Edition organization allowed for the ingestion of model outputs,where the platform can store and review generated fraud scenarios for security analysis. The main limitation remains the computational resource required for real-time fine-tuning, suggesting a periodic retraining pipeline is more practical than a real-time one.

Future enhancements will look towards **Salesforce AgentForce**. This ”agentic” capability could allow the system to not only identify potential threats but to autonomously create sandboxed environments to test if these theoretical fraud vectors would actually bypass the current live rules, creating a fully autonomous defense loop.

VI. Conclusion

This research successfully demonstrated the application of the Qwen Large Language Model for proactive financial fraud detection. By generating novel fraud scenarios (validated by low similarity metrics), the framework provides a method to anticipate future threats. Integration with Salesforce bridges the gap between theoretical AI capability and enterprise deployment. We demonstrated that using QLoRA fine-tuning allowed for effective model adaptation without prohibitive computational costs.

References

[1] Algoanalytics.com. (2025). ”Generative AI for Financial Fraud Detection: Strengthening Security Measures.”.

[2] F. Carcillo, A. Dal Pozzolo, Y.-A. Le Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, "SCARFF: A scalable framework for streaming credit card fraud detection with spark," *Information Fusion*, vol. 41, pp. 182–194, 2018. doi: 10.1016/j.inffus.2017.09.005.

[3] Y. Chen, C. Zhao, Y. Xu, C. Nie, and Y. Zhang, "Deep Learning in Financial Fraud Detection: Innovations, Challenges, and Applications," *Data Science and Management*, 2025. doi: 10.1016/j.dsm.2025.08.002.

[4] S. A. Chowdhury and M. Delowar, "Next Generation Financial Security: Leveraging AI for Fraud Detection, Compliance and Adaptive Risk Management," *Well Testing*, vol. 34, no. S3, pp. 61–79, 2025.

[5] S. Dixit, "Generative AI-Powered Document Processing at Scale with Fraud Detection for Large Financial Organizations," 2024. doi: 10.36227/techrxiv.172651542.23422356/v1.

[6] S. S. Dubey, V. Astvansh, and P. K. Kopalle, "EXPRESS: Generative AI Solutions to Empower Financial Firms," *Journal of Public Policy & Marketing*, 2024. doi: 10.1177/07439156241311300.

[7] FTI Consulting. (2025). "Working Smarter, Not Harder: Generative AI's Edge in Financial Crime Detection." [online]: https://www.fticonsulting.com/insights/articles/working-smarter-not-harder-generative-ais-edge-financial-crime-detection.

[8] S. Nakharu and P. Kumar, "Fraud Detection in Banking Using Generative AI," 2025. doi: 10.5281/zenodo.17634095.

[9] R. Otubu, "Integrating Generative AI in Fraud Detection and Anti-Money Laundering: A Comprehensive Review and Framework Proposal," *Journal of Emerging Trends in Engineering and Applied Sciences*, vol. 15, no. 5, pp. 2141–7016, n.d.

[10] I. Ridwan, "Formulating Advanced Data-Driven Architectures Leveraging Machine Learning, Systemic Analytics, and Predictive Insights for Proactive Financial Threat Detection and Mitigation," *International Journal of Science and Engineering Applications*, vol. 10, 2021.

[11] B. Saha, N. Rani, and S. K. Shukla, "Generative AI in Financial Institution: A Global Survey of Opportunities, Threats, and Regulation," *arXiv.org*, 2025.

[12] S. Sai, K. Arunakar, V. Chamola, A. Hussain, P. Bisht, and S. Kumar, "Generative AI for Finance: Applications, Case Studies and Challenges," *Expert Systems*, vol. 42, no. 3, 2025. doi: 10.1111/exsy.70018.

[13] H. K. Sriram, "Leveraging AI and Machine Learning for Enhancing Secure Payment Processing: A Study on Generative AI Applications in Real-Time Fraud Detection and Prevention," *SSRN Electronic Journal*, 2025. doi: 10.2139/ssrn.5203586.

[14] T. Xu, "Leveraging Blockchain Empowered Machine Learning Architectures for Advanced Financial Risk Mitigation and Anomaly Detection," *World Journal of Innovation and Modern Technology*, vol. 7, no. 4, pp. 90–100, 2024. doi: 10.53469/wjimt.2024.07(04).11.