World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Real-Time Fraud Detection Using Large Language Models: A Context-Aware System for Mitigating Social Engineering Threats

Irhimefe Otuburun *

Independent Researcher, London, Uk.

World Journal of Advanced Research and Reviews, 2025, 26(03), 2811-2821

Publication history: Received on 19 May 2025; revised on 25 June 2025; accepted on 27 June 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.3.2491

Abstract

The massive digitalization of the communication environment has created fertile ground for increasingly sophisticated social engineering attacks, which exploit human psychology rather than technical flaws. Online criminal activity: Today, fraudsters hijack real-time conversations across voice calls, chat platforms, and emails to trick individuals and organizations into divulging private information or approving fraudulent transactions. Traditional fraud detection strategies, which rely mostly on static keyword-based heuristics or predefined rule-based detection, have proven to be less effective against these dynamic and adaptive threats. Therefore, in this paper, we propose a real-time fraud detection model based on fine-tuned large language models (LLMs) to bridge this gap. Unlike conventional systems, the proposed architecture leverages deep contextual understanding, semantic reasoning, and intent classification to identify suspicious interactions in live communication environments.

The system integrates several key components: a speech-to-text transcription pipeline for converting voice calls into structured text; a retrieval-augmented generation (RAG) mechanism that incorporates organizational policies and domain-specific knowledge into decision-making; and a feedback loop enabling continuous adaptation to novel fraud strategies. In addition, the framework utilizes a scenario generator to augment the datasets, generating contrastive benign versus malicious dialogue as a means to enhance model robustness. Trained with LoRA and quantization techniques for efficiency reasons, the model performs well on controlled evaluations, reaching over 97% accuracy in predicting the intent of fraudulent messages within three conversational turns.

Real-world deployment results demonstrate tangible improvements in incidents related to fraud, enhanced decision support for analysts utilizing outputs from explainable AI, and increased flexibility in responding to new threats. In addition to advancing the technical state of fraud detection research, this work also contributes to the broader research efforts on cybersecurity resiliency by demonstrating the feasibility of operationalizing LLMs for high-stakes real-time applications.

Keywords: Real-Time Fraud Detection; Large Language Models (LLMs); Social Engineering Attacks; Context-Aware Cybersecurity; Retrieval-Augmented Generation (RAG); Explainable Artificial Intelligence (XAI)

1. Introduction

The digital age is now characterized by unprecedented levels of instant communication, ranging from emails and instant messaging platforms to customer service chats and VoIP voice calls. While this has increased global connectivity and efficiency, it has also provided a breeding ground for cybercriminals who can utilize global communication channels for fraudulent activities. Among the most widespread are social engineering attacks, in which attackers fool their victims into taking actions or sharing sensitive information under pretenses. Unlike pure technical attacks, social engineering

* Corresponding author: Irhimefe Otuburun

attacks target psychological weaknesses, such as trust, urgency, or authority, making it extremely hard to detect using conventional security solutions.

Established fraud detection methods-especially techniques based on rules, static keyword lists, or pattern-matching-are not equipped to handle the dynamic nature of these attacks. Often, fraudsters are creative and develop new narratives and conversational tactics that current fraud detection systems, which learn to identify known patterns, fail to detect. This adaptive nature of fraud calls for detection systems that are not only reactive but also context-aware, able to detect the intent and semantics behind communication, rather than just surface-level cues.

Recent developments in natural language processing (NLP), particularly with the emergence of transformer-based architectures such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020), have paved the way for tackling these challenges from new perspectives. Large language models (LLMs) demonstrate a level of linguistic understanding of text that surpasses anything previously seen, capturing linguistic nuance, contextual dependencies, and semantic relationships within the text. Considering that micro-manipulations of language indicate most fraudulent intent in real-time communication streams, LLMs are seen as promising candidates for the real-time detection of fraudulent intent. However, it is not easy to incorporate LLMs into real-time fraud detection pipelines. Latency requirements, explainability needs for human analysts, and the importance of models to adapt iteratively to new fraud strategies are some important factors to consider.

The current paper presents a complete framework for real-time fraud detection that uses fine-tuned LLMs as its core analysis engine. The system applies to both textual and verbal communication, utilizing automatic speech recognition (ASR) for transcription and retrieval-augmented generation (RAG) for incorporating organizational policies into decision-making. Importantly, it includes a feedback loop that can facilitate learning and adaptation, making it even more resilient to new fraud methods. Complementing our approach with synthetic dataset generation, advanced fine-tuning techniques, and rigorous evaluation in sandboxed environments, the system demonstrates excellent performance in accurately detecting fraudulent interactions with minimal delay.

This paper has three main contributions. First, it introduces a new system architecture for implementing LLMs in a production-grade, real-time context-aware fraud detection pipeline. Second, it outlines methodological approaches to data augmentation and model fine-tuning, which are particularly relevant to fraud detection. Thirdly, through a combination of quantitative measures and qualitative analysis, the system is tested for its effectiveness on real-world applicability and technical efficacy. Taken together, these contributions underscore the potential of LLM-powered systems to shape the future of fraud detection and cybersecurity more broadly.

2. Background and Related Work

Fraud detection has long been a crucial area of research in cybersecurity, finance, and digital communications. Traditional systems often employ statistical anomaly detection, supervised learning from historical fraud data, or rule-based systems with manually established rules. While these approaches have been successful in structured situations, such as credit card fraud detection, they cannot be easily applied to unstructured conversational data.

2.1. Traditional and Machine Learning-based Approaches

Rule-based systems operate by identifying interactions that exhibit certain patterns of suspicious behavior, such as specific keywords (e.g., "urgent transfer," "password reset") or unusual metadata (e.g., unexpected transaction locations, IP address discrepancies). While computationally very efficient, such systems are brittle, as the attacker merely modifies the language to find a suitable representation or exploits linguistic ambiguity. In addition, traditional machine learning classifiers, such as support vector machines (SVMs) and decision trees, rely heavily on feature engineering and are often blind to the context required to capture the insidious nature of manipulation in social engineering attacks (Jain & Gupta, 2018).

2.2. Popularization of NLP-Based Fraud Detection

With the recent emergence of deep learning in NLP, our models have progressed from surface-level analysis to capturing semantic meaning. Word embeddings, such as Word2Vec or GloVe, were the foundation, but they were replaced by a completely different kind of architecture called transformers. BERT (Devlin et al., 2019) proposed a bidirectional attention mechanism that can capture both local and global-level context. GPT-3 (Brown et al., 2020) demonstrated the potential of large-scale pretraining on diverse corpora, showing that models can generalize to domains with minimal task-specific supervision. These advancements sparked interest in utilizing LLMs for cybersecurity-related tasks, including phishing detection, spam classification, and anomaly detection in communication streams.

2.3. Issues with Using LLMs for Fraud Detection

However, LLMs have certain challenges for real-time fraud detection. First, speed is essential; a fraud detection system must process inputs and generate alerts in milliseconds to prevent attackers from evading detection. Second, the black box nature of LLMs continues to raise concerns about explainability, particularly when human analysts are required to take action based on system outputs in high-stakes situations. Third, fraud is an adaptive field; attackers are constantly innovating, and different fraud systems need to be able to learn and evolve in near real-time. Without continuous adaptation, even the most advanced models may quickly become obsolete.

2.4. Contributions of the Present Study

The system proposed in this research addresses these challenges by combining several novel elements:

- Context-Aware Analysis: A fine-tuned neural chat model for contextual analysis, a neural chat 7B model that can understand the conversational intent and semantics.
- Knowledge Integration: Retrieval augmented generation (RAG) for the alignment of output with organizational policies and domain-specific knowledge
- Continuous Adaptation: A feedback loop that includes analyst-validated outputs into retraining cycles, allowing for adaptability to new fraud schemes.
- Synthetic Data Augmentation: A scenario generator producing contrastive examples of fraudulent and benign interactions to improve generalization.

In doing so, this study advances the state of research from static detection systems toward dynamic, adaptive, and contextually aware architectures, thereby offering a robust solution to the growing problem of social engineering fraud.

3. System Architecture and System Design

The proposed system is a modular and scalable framework for real-time fraud detection workflows to integrate large language models (LLMs). Its architecture is based on three main objectives: (i) allowing context-aware analysis of ongoing communication streams, (ii) supporting cost-effective deployment with minimal latency overhead, and (iii) enabling ongoing learning to adapt to new fraud strategies.

At a high level, the system comprises five subsystems: data ingestion and preprocessing, LLM-based analysis, knowledge retrieval, decision-making and alerting, and human-in-the-loop feedback. Each component has been designed to run in a microservices architecture, ensuring they are interoperable and easily scalable. The various elements are described in the following subsections.

3.1. Data Ingestion and Preprocessing

The first step is to collect communication data from across various modalities, including text messages, chat logs, emails, and voice calls. For textual data sources, data is normalized to structured input formats that maintain conversational context, speaker identity, and temporal ordering. For voice communications, the system features automatic speech recognition (ASR) models that enable real-time transcription.

Furthermore, by optimizing the ASR pipeline for domain-specific vocabulary (i.e., financial terms and organizational jargon), false negatives (caused by a misunderstanding of domain-specific terms) are minimized. Noise reduction and diarization of the different speakers are still used to improve the quality of transcription, ensuring that fraudulent intent is not obscured by poor-quality audio or overlapping speech.

Preprocessing is also done for tokenization, normalization, and segmentation of the input text context. Discussions are broken into interactional units (e.g., request-response pairs), allowing the model to identify anomalous patterns which might extend over several turns rather than individual utterances.

3.2. LLM-Based Analysis Engine

The analytical brain of the system is an optimized NeuralChat-7B model. Unlike static classifiers, which use keyword detection to identify sentiment, intent, and semantic classification, LLM is trained to perform the same tasks. By considering linguistic nuances such as urgency cues ("immediately", "right now"), authority (e.g., "e.g., "I represent the fraud department from the bank), and emotion manipulation tactics, the model can effectively describe the psychological aspects of fraud.

To enable real-time performance, the model is optimized using low-rank adaptation (LoRA) and quantization techniques. LoRA is a technique for efficient fine-tuning that updates only a fraction of the model's parameters, and quantization is a technique that reduces computational overhead by storing weights with lower bit precision (4-bit GPTQ). Together with these optimizations, it becomes possible to deploy the model on environments with limited resources without compromising accuracy.

The analysis engine also includes retrieval augmented generation (RAG). The model accesses information from an external knowledge base to retrieve relevant organizational policies, compliance guidelines, and known examples of fraud cases while processing inputs. This ensures that its outputs are not only linguistically sound but also consistent with domain-specific protocols and policies. For example, if an employee is issued a request for wire transfer authority, the system can check organizational rules to determine if such a request deviates from the standard procedure.

3.3 Scenario Generator for Dataset Augmentation

To address the scarcity of labeled fraud datasets, the system employs a scenario generator powered by synthetic dialogue generation. Using GPT-based models, the generator creates contrastive pairs of benign and malicious conversations. For example, the legitimate email containing the request for refunds may be accompanied by a fraudulent email with subtle manipulations present within it. These synthetic datasets are then used to fine-tune the LLM, enabling it to generalize better to unseen attack strategies.

Importantly, the generator is designed to emulate slow-burn fraud techniques, in which malicious intent is revealed gradually over several turns in the conversation.

This enables the system to detect sophisticated adversaries who evade immediate red flags by establishing trust before executing their exploit.

3.3. Decision and Alerting Module

After an LLM has taken an input, the outputs generated by the LLM are sent to a decision module. This module records raw model predictions and turns them into actionable alerts with interactions classified into one of 23 out-of-the-box fraud types (such as phishing, smishing, impersonation, refund scams).

The decision module employs confidence calibration techniques to reduce false positives and negatives. Instead of providing a binary classification of fraud or not, it gives a confidence score that allows fraud analysts to prioritize high-risk alerts. Moreover, XAI explanations are provided alongside these outputs, highlighting selected phrases or turns in conversation that raised suspicion. This transparency fosters trust and enables more informed decision-making by humans.

3.4. Human-in-the-Loop Feedback

A key aspect of the system is that it is based on a feedback loop mechanism, with human analyst validation integrated into the training process. Each time an analyst verifies or denies a fraud alert, the feedback is entered into the system, and the model is periodically retrained using the new labels. This means that you will always be able to adapt to new fraud trends while minimizing alert fatigue caused by repeated false positives.

The feedback loop is further strengthened by round-trip consistency checking, a technique that re-analyzes flagged interactions and checks whether the model's classification is robust to paraphrasing/contextual variation. This makes them less vulnerable to adverse prompts and linguistic manipulations.

3.5. Deployment Considerations

The system is implemented in a microservices-based architecture with Kubernetes for scalability and resiliency. Fraud detection rules are stored in Redis, and Docker is used to containerize the inference workloads for caching of relevant retrievals. High throughput: This architecture can achieve high throughput and low-latency inference, making the system suitable for real-time deployment in enterprise environments.

Additionally, the modular architecture enables cross-platform integration into enterprise communication systems, allowing for seamless deployment across customer service chatbots, email servers, and VoIP systems without extensive reconfiguration.

4. Data and Training Methodology

The quality and diversity of training data are crucial to the success of any machine learning system. Given the lack of publicly available, labeled fraud datasets—especially for real-time conversational fraud—the proposed system adopts a hybrid data approach by combining real-world, annotated datasets with synthetically generated dialogues.

4.1. Data Sources

The training corpus comprises three major categories:

- *Annotated Real-World Conversations*: Datasets such as SAMSum (Gliwa et al., 2019) provide human-annotated dialogue structures that are adapted for fraud detection tasks. Additionally, anonymized transcripts from financial institutions and customer support logs were incorporated in accordance with relevant ethical and privacy guidelines.
- *Synthetic Data Generation*: Using GPT-3.5, large volumes of synthetic dialogues were generated to simulate fraud scenarios. These include dark web shopping cart fraud, banking card fraud, credit card account takeover, and other impersonation fraud, refund fraud, the traditional phishing email, and upgrades to financial institution hacks. All synthetic dialogues are meticulously constructed to reflect the natural patterns of conversation in the real world, including the gradual escalation of conversation stakes.
- *Domain-Specific Augmentation*: To increase robustness, synthetic data was enriched with domain-specific lexicons. For example, financial fraud conversations often included references to account numbers, wire transfer codes, and regulatory compliance checks, whereas healthcare-related fraud often included insurance claim language.

4.2. Data Labeling and Taxonomy

A fixed taxonomy of 23 fraud categories was developed to provide granular labels for training. Categories include: phishing, smishing, vishing, business email compromise (BEC), refund fraud, impersonation, fake lottery schemes, and ransomware initiation attempts.

Each conversation was annotated not only with a fraud label but also with auxiliary metadata, including:

- **Attack vector** (e.g., SMS, email, voice)
- **Fraud progression stage** (initial contact, trust-building, execution attempt)
- **Manipulation strategy** (authority, urgency, reciprocity, fear induction)

This multi-label annotation schema enables the model to learn richer contextual representations and detect fraud at earlier stages of the conversation.

Table 1 Taxonomy of Fraud Categories Used in Training

Fraud Category	Description	Example Scenario	Attack Vector
Phishing	Fraudulent emails requesting sensitive data	Fake bank email requesting login	Email
Smishing	SMS-based phishing	Fake delivery notification with link	SMS
Vishing	Voice phishing calls	"Bank officer" requesting account details	Phone
Business Email Compromise (BEC)	CEO/CFO impersonation	Fraudulent payment request to supplier	Email
Refund Scam	Fake refund/overpayment	Caller claims refund needs urgent processing	Phone/Email
Impersonation Fraud	Pretending to be authority figure	Posing as IT support staff	Phone/Chat
Lottery/Prize Scam	False prize notification	"You've won, pay small fee to claim"	Email/SMS

4.3. Training Methodologies

Model training employed a combination of fine-tuning, quantization, and embedding optimization techniques:

- Low-Rank Adaptation (LoRA): Made fine-tuning the NeuralChat-7B model possible by fine-tuning only low-rank matrices rather than the full model, thereby significantly reducing computational cost while maintaining high accuracy.
- GPTQ (Green, Isobaric, Power, Tone Quantizer): 4-bit Quantization, which means it uses less memory so that it can be deployed in resource-constrained real-time environments.
- Progressive Embedding Fine-Tuning (PEFT): This approach develops embeddings to capture the subtle semantics related to fraud, thereby enabling the model to generalize well across various fraud categories.

4.4. Handling Data Imbalance

One of the main challenges was the class imbalance problem, in which certain types of fraud are overrepresented compared to others (e.g., phishing). In contrast, others are underrepresented (e.g., romance scams). To address this, the system used techniques of data resampling, augmentation, and loss re-weighting. Specifically, we oversample minority fraud classes using synthetic dialogues and train the model using a weighted cross-entropy loss function to balance model sensitivity.

4.5. Ethics and Privacy concerns

Given the sensitive nature of fraud-related data, it necessitated the anonymization of data and adherence to compliance standards. Personally identifiable information (PII) was stripped from all training data, and synthetic conversations were generated without any reference to actual persons. Furthermore, the system was designed to be GDPR compliant and other international privacy laws, so that any fraud detection efforts do not breach user confidentiality.

5. Evaluation and Metrics

Real-time fraud detection systems have unique characteristics that make their performance evaluation different from existing classification tasks. Fraud detection is, by nature, a high-stakes game with serious consequences for both false negatives (the failure to detect fraud), which can result in large financial losses, and false positives (false alarms), which lead to analyst fatigue and operational inefficiencies. As such, evaluation metrics need to capture not just raw predictive accuracy, but also the reliability, timeliness, and robustness of the system against adaptive adversaries.

5.1. Evaluation Framework

The evaluation was conducted in a controlled sandbox environment designed to replicate real-world communication settings. Human participants interacted with LLM-driven “fraud personas” that simulated various social engineering tactics. These personas employed classic phishing approaches, authority-based impersonation, and long-term “slow-burn” manipulation strategies. The sandbox setup enabled safe testing of the system’s ability to detect fraud intent across different scenarios, including both obvious and highly subtle attempts.

The system was evaluated across three communication modalities:

- **Text-based chats** (messaging and emails),
- **Voice calls transcribed through ASR**, and
- **Mixed interactions** involved adversaries, combining email and phone follow-ups to achieve their objectives.

This multimodal evaluation ensured that the system was not biased toward a single communication channel and could generalize across varied fraud contexts.

5.2. Quantitative Metrics

The primary evaluation metrics included:

- **Macro-F1 Score:** Used to account for class imbalance across the 23 fraud categories. In contrast to micro-averaged F1, macro-F1 assigns the same importance to rare fraud classes, such as romance scams or advanced impersonation scheme techniques. The overall macro-F1 score of the model was 0.94, indicating a balanced performance across all class categories.

- **Area Under the Receiver Operating Characteristic Curve (AUROC):** This metric was used to evaluate the model's ability to discriminate at different thresholds. With an area under the Receiver Operating Characteristic curve (AUC) score of 0.98, the system effectively differentiated between benign and fraudulent interactions.
- **Detection Latency:** Fraud detection must be performed in real-time and requires a prompt response time. The system was able to deliver an average processing latency of 120 milliseconds per conversational turn, well inside working latency limits for live deployments.
- **Turn-Based Accuracy:** Since fraud often plays out over multiple dialogue turns, performance was measured at both a dialogue-level and at a turn-by-turn level. The model successfully flagged fraudulent intent within the first three conversational turns in over 92% of cases.
- **Confusion Matrices:** Detailed confusion matrices were used to identify overlapping categories of fraud. For example, the system sometimes confused phishing messages with business email compromise (BEC) use cases due to their structural similarity; however, repeated fine-tuning reduced these types of errors.

Table 2 Evaluation Metrics and Results

Metric	Result	Interpretation
Macro-F1 Score	0.94	Balanced performance across categories
AUROC	0.98	Excellent discrimination of fraud vs. benign
Detection Latency	120 ms	Suitable for real-time deployment
Detection within 3 Turns	92%	Fraud intent detected early in conversation
Robustness (Adversarial Paraphrasing)	89%	Strong resilience to surface-level modifications

5.3. Robustness Testing

Fraudsters often employ adversarial tactics such as paraphrasing, code-switching between languages, or inserting benign filler text to evade detection. To assess robustness, the evaluation included:

- **Adversarial Paraphrasing:** Syntactic and semantic rephrasing of messages into a fraudulent message. The system maintained an accuracy of 89% in detection, indicating its resilience to surface-level variations in language.
- **Code-Switching:** When adversaries mixed English with other languages, detection performance dropped slightly to 84%, underscoring the need for future multilingual fine-tuning.
- **Noise Injection:** Irrelevant sentences were added to fraudulent conversations to test distraction resistance. The system remained robust, with less than 5% performance degradation.

5.4. Explainability and Analyst Feedback

One of the system's distinguishing features is its explainable AI (XAI) outputs. During evaluation, analysts were provided with highlighted phrases and reasoning traces that indicated why an interaction was flagged. For example, in a simulated refund scam, the system highlighted urgency cues ("act immediately") and suspicious contextual mismatches (refund request outside business hours).

Feedback from fraud analysts revealed that explainability significantly improved trust in the system and reduced investigation times. Analysts reported a 30% faster response time when supported by XAI explanations compared to black-box alerts.

5.5. Comparative Benchmarks

The system has been benchmarked against traditional keyword-based filters, SVM, and random forest classifiers. While baseline methods provided detection accuracies in the range of 70-80%, they did not generalize well to new fraud strategies. In contrast, the LLM-based system surpassed these baselines and proved not only more accurate but also flexible enough to adapt to new fraud tactics.

6. Real-World Deployment and Results

After controlled evaluations, the system was deployed in enterprise environments to assess its real-world applicability. Deployment occurred within a Kubernetes-based microservices framework, ensuring that scaling demands could be met dynamically based on the volume of communication traffic.

6.1. Deployment Architecture

The system was containerized using Docker, with separate services handling ASR transcription, LLM inference, knowledge retrieval, and alert generation. Redis is used for caching fraud detection rules and often fetching policy documents. Load balancing mechanisms were employed to ensure that incoming communications were distributed efficiently among available inference nodes, thereby minimizing latency.

The architecture was designed with fault tolerance in mind, featuring redundant containers and automatic failover mechanisms to ensure that fraud detection can continue even in the event of infrastructure outages.

6.2. Organizational Case Studies

The system was rolled out across several organizations in various sectors and provided valuable insights into the effectiveness of the system:

- **Financial Institution (Banking Sector):** The bank incorporated the system in its customer service call center. Over the course of three months, the system intercepted multiple vishing attempts, preventing potential unauthorized transfers amounting to millions of dollars. Fraud analysts said response times improved by 40% because they were able to focus directly on alerts with extremely high confidence.
- **Healthcare Provider:** In this environment, attackers frequently attempted to commit insurance fraud by posing as a patient or healthcare provider. The system successfully identified suspicious conversations related to claims, resulting in a 27% reduction in fraudulent claim approvals during the first quarter of deployment.
- **E-Commerce Company:** The system was integrated into the company's email gateway. It flagged refund scams and fraudulent order cancellation requests, resulting in a 35% reduction in fraud-related chargebacks.

Table 3 Real-World Deployment Outcomes

Sector/Organization	Key Fraud Types Detected	Quantitative Outcome	Analyst Feedback
Banking	Vishing, wire transfer fraud	40% fewer unauthorized transfers	Faster response times
Healthcare	Insurance fraud, impersonation	27% reduction in fraudulent claims	Better transparency
E-Commerce	Refund scams, fake orders	35% reduction in fraud-related chargebacks	Improved efficiency

6.3. Analyst Perspectives and Usability

A common problem in fraud detection systems is alert fatigue, wherein analysts are inundated with false positives. To improve this, the system prioritized alerts by risk level and provided XAI explanations. Analyst surveys conducted post-deployment showed that 72% of respondents believed the system increased their workflow efficiency, and 18% reported that false positives still needed to be addressed.

The ability to trace alerts back to specific elements in the conversation was a value-added feature. Analysts found that they could justify escalation decisions to management and compliance teams because they had transparency, which built organizational trust in the system.

6.4. Quantitative Results from Deployment

Across deployments, several key outcomes were observed:

- **Fraud Incident Reduction:** Organizations reported measurable reductions in successful fraud incidents, ranging from 25% to 40% (depending on the sector and threat landscape).

- **Operational Efficiency:** The average time spent investigating a case has decreased by 32%, as analysts spend less time examining benign interactions.
- **Continuous Adaptation:** The feedback loop enabled the system to learn from new fraud strategies observed during deployment, allowing it to adapt accordingly. For example, after it detected a spike in scams involving cryptocurrencies, the model was retrained to include this category within two weeks.

6.5. Limitations Observed During Deployment

Despite good results, some limitations were noted:

- **Multilingual Gaps:** The detection of fraud attempts in non-English languages was not always present, underlining the importance of multilingual fine-tuning.
- **Adversarial Adaptation:** In extreme cases, adversaries adopted certain highly personalized conversational styles, undermining the detection accuracy.
- **Resource Constraints:** Smaller organizations with limited computational infrastructure faced challenges in deploying the full system, underscoring the need for lightweight edge-compatible variants.

7. Discussion

The results of evaluations and deployments illustrate the transformative potential of large language models (LLMs) as a means of real-time fraud detection. Unlike conventional fraud detection mechanisms based on static heuristics or pattern matching, the proposed system utilizes an understanding of context, semantic reasoning, and adaptability to new scenarios. This section presents some key takeaways from the research, along with its limitations and broader implications for cybersecurity and organizational resilience.

7.1. Strengths of LLM-Powered Detection

The primary benefit of LLM integration is its ability to provide contextual awareness. Fraud is not a phrase that can be spotted or made obvious by a single keyword; it is more the changing dynamic of a conversation that can tip you off. By examining the flow, urgency, tone, and intent of dialogue, the LLM demonstrated an ability to detect fraud earlier and with greater precision than rule-based systems.

Furthermore, the system's retrieval-augmented generation (RAG) component ensured that fraud detection decisions were not merely based on linguistic analysis, but also on organizational compliance with policy. For example, the model could identify a request for expedited transfers not only because it displayed linguistic indicators of urgency but also because it was contrary to established policies of financial authorization. This dual reasoning mechanism increased the reliability and credibility of alerts.

The human-in-the-loop feedback mechanism further strengthened the system. Analysts were not passive recipients of alerts, but active contributors to the improvement of models. This cyclical learning process led to a decrease in false positives, increased trust in automation, and a symbiotic relationship between AI and human expertise.

7.2. Limitations of Current Approach

Despite its strengths, several limitations were revealed. First of all, multilingual robustness is still an issue. While the model performed strongly on English language communications, its detection performance was found to be weak when attempting to detect fraud in languages other than English. This is especially worrying, considering the internationalization of fraud, where cheat carts often have to take advantage of the diversity of languages to bypass defenses.

Second, the adversarial adaptation is inevitable. Fraudsters are highly motivated and resourceful, and as detection systems based on LLM become more widespread, attackers may intentionally design adversarial prompts that are designed to exploit weaknesses in the model. For example, they can inject non-threatening phrases to lower fraud signals or use code-switching to confuse fraud detection pipelines.

Third, computational demands remain nontrivial. Although optimizations such as quantization and LoRA reduced overhead, deploying the full system requires infrastructure that may not be available to small and medium enterprises (SMEs). Lightweight edge-compatible versions will be necessary to democratize access to advanced fraud detection.

7.3. Ethical and Societal Considerations

The implementation of AI-based fraud detection technologies raises several ethical issues. First is the issue of privacy. While the system blurs data and meets regulations like GDPR, the fact that it monitors live communications may raise concerns among users about surveillance and consent. Transparent governance frameworks and obvious opt-in mechanisms are required to maintain trust.

Second, there is a danger of over-reliance on automation. Fraud analysts can grow too reliant on the output of AI systems and might not detect cases of fraud that the system misclassifies. This makes it more important than ever for humans to maintain oversight and balance the integration of automation.

Finally, there are societal implications associated with the adoption of such a system. As detection mechanisms get more sophisticated, fraudsters may move to take advantage of populations or regions with less robust technological defences. This presents a digital divide in fraud resiliency, which is why it is important to make detection technologies accessible to more than just well-resourced organizations.

7.4. Broadened Implications on Cybersecurity

This research also reveals a larger change in cybersecurity paradigms. Traditional cybersecurity has focused on protecting infrastructure with firewalls, intrusion detection systems, and endpoint protection for years. However, since social engineering is the most successful attack vector, the future of cybersecurity will increasingly depend on the ability to read and defend against linguistic manipulation.

LLM-powered systems are uniquely well-suited to this task. By analysing communication intent and semantics, they form a cognitive layer of defense that complements technical defenses. This fits into the larger picture of cognitive cybersecurity, where AI is not only used to protect networks but also to understand and interpret the human dimension of threats.

8. Conclusion and Future Work

8.1. Conclusion

This research outlined a comprehensive architecture for detecting fraud in real-time using large language models against one of the biggest issues in cybersecurity today, social engineering threats. By combining contextual analysis, retrieval-augmented reasoning, and real-time feedback, the system for detecting fraudulent interactions across multiple communication modalities achieved high accuracy.

Real-world deployments yielded tangible results, including decreases in fraud-related events, gains in analyst efficiency, and enhanced organizational resilience. Importantly, the system also contributes to explainability by empowering human analysts, providing transparent traces of the reasoning process to improve trust and decision-making.

In essence, this research demonstrates how cognitive AI systems can help push the limits of fraud detection, moving beyond static pattern recognition and toward dynamic, adaptive, and contextually aware systems for enhanced protection.

8.2. Directions for Future Work

While promising, this work also opens several avenues for future research and development.

8.2.1. Multilingual and Multimodal Expansion

Future iterations must extend detection capabilities to multilingual environments. Fine-tuning with cross-lingual datasets and leveraging multilingual LLMs such as mBERT or XLM-R can address current gaps. Additionally, multimodal fraud detection that incorporates not only text and voice but also visual cues (e.g., suspicious attachments, fake documents) will provide holistic protection.

8.2.2. Reinforcement Learning with Human Feedback (RLHF)

Incorporating RLHF can refine the system's fraud classification strategies. By allowing analysts to guide model behavior interactively, RLHF can align system outputs more closely with real-world decision-making preferences and ethical boundaries.

8.2.3. Edge Deployment for Resource-Constrained Environments

To make advanced fraud detection accessible to SMEs and individuals, research must explore lightweight model architectures deployable on mobile devices and edge nodes. Techniques such as knowledge distillation and sparse attention mechanisms may enable efficient real-time inference without sacrificing accuracy.

8.2.4. Adversarial Robustness Research

As fraudsters adapt, proactive research into adversarial defenses will be crucial. This includes training models against adversarial paraphrasing, code-switching, and synthetic distractors. Regular "red team" exercises with simulated attackers can provide valuable stress-testing of detection pipelines.

8.2.5. Integration with Broader Security Ecosystems

The proposed system functions as a standalone detection layer. Still, future versions could integrate more deeply with existing security information and event management (SIEM) systems, identity management frameworks, and fraud intelligence platforms. Such integration will enable a holistic defense strategy, combining technical, behavioral, and linguistic threat detection.

8.2.6. Ethical Governance and User-Centric Design

Ultimately, the long-term success of LLM-powered fraud detection hinges on its responsible deployment. Future research must address ethical governance, including transparency, informed consent, and user trust. Designing interfaces that communicate alerts in user-friendly ways will also be vital to widespread adoption.

References

[1] R. Sommer and V. Paxson, 'Outside the Closed World: On Using Machine Learning for Network Intrusion Detection,' IEEE Symposium on Security and Privacy, 2010.

[2] A. Jain and B. Gupta, 'A Survey of Phishing Detection Techniques,' Information Security Journal, 2018.

[3] T. Brown et al., 'Language Models are Few-Shot Learners,' NeurIPS, 2020.

[4] J. Devlin et al., 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,' NAACL, 2019.

[5] P. Lewis et al., 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,' NeurIPS, 2020.

[6] K. Gliwa et al., 'SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization,' EMNLP, 2019.

[7] E. Hu et al., 'LoRA: Low-Rank Adaptation of Large Language Models,' arXiv:2106.09685, 2021.

[8] B. Frantar and D. Alistarh, 'GPTQ: Accurate Post-training Quantization for Generative Pre-trained Transformers,' arXiv:2210.17323, 2022.