ESP-IJACT
ESP International Journal of Advancements in Computational Technology
ISSN: 2583-8628 / Volume 3 Issue 2 April 2025 / Page No: 36-49
Paper Id: IJACT-V3I2P105 / Doi: 10.56472/25838628/IJACT-V3I2P105

Original Article

Context-Aware LLM Fraud Sentinels for Card Authorization

Vijay Kumar Soni[1], Aman Sardana[2], Pradeep Manivannan[3]
[1,2]Discover Financial Services, USA.
[3]Independent Researcher, USA.

Received Date: 25 February 2025 Revised Date: 06 April 2025 Accepted Date: 20 April 2025

Abstract: *The rapid evolution of payment card fraud techniques necessitates advanced detection frameworks capable of adapting to emerging threats with minimal delay. Most of these systems have problems detecting unknown fraud patterns and allow much inaccurate positive detection. New hybrid architecture has been proposed using Large Language Models (LLMs), Graph Neural Networks (GNNs) and context together, which helps to detect fraud for the card authorization process promptly and reliably. The LLM section can extract meaningful metadata through using advanced methods, while the GNN component dynamically models transactional relationships and propagates risk scores across entities such as customers, merchants, and transactions. Because of this design, the method adapts better, achieves a balanced precision and recall and allows the system to understand and explain its actions through graph features and attention. When tested using anonymized transaction samples, the system achieved much better results in zero-day fraud, fewer false positives and shorter processing times. The research mentions that there are problems with scaling, privacy, ongoing training and explaining models and outlines how more research can focus on working together, using multiple data sources and testing regulations in practice. This integrated approach lays the foundation for next-generation, context-aware fraud detection systems that can safeguard payment ecosystems while delivering seamless user experiences.*

Keywords: *Fraud Detection, Large Language Models (LLMs), Graph Neural Networks (GNNs), Zero-Day Fraud, Explainable AI (XAI).*

I. INTRODUCTION

With the rapid digital growth, millions of consumers and businesses are conducting finances in new ways [1]. The essential part of this system is the payment card authorization process, which checks that a transaction is real and approves or declines it in real time [2]. Transactions normally require the bank, the payment gateway and groups in between to connect and verify that each action made is legitimate. Since the amount and complexity of electronic payments are increasing, the need to safeguard these transactions from fraud rises [3].

Because of progress in payment technology, fraudsters have become even better at what they do. Conventional ways of detecting fraud which rely on easy-to-use models do not work well against the large number of sophisticated fraud schemes today [4]. Some of these systems have a problem known as high false positives, when real transactions are denied by mistake, which causes poor user experience and may reduce the company’s profits. Also, such approaches depend on previous data and fixed rules; they find it hard to spot new types of zero-day fraud. Given how fraud threats change, more intelligent, adaptable and contextual fraud detection is needed.

The reason for using context in fraud detection lies in the fact that, on its own, analyzing transactions is not accurate enough to label actions as being either fraudulent or legitimate. Signals from the context, including merchant category, reputation and location, device fingerprint, timing of the purchase and the buyer’s behavior record, and key information to help detect fraud [5]. Enriching their data in this way helps fraud detection systems better spot unusual activities as they are happening.

Transformer-based LLMs which have appeared recently in artificial intelligence give powerful help in using context in fraud detection [6]. LLMs were first meant for natural language tasks and have shown strong abilities in processing contextual data from various sources. LLMs are able to notice subtle problems and unusual behavior in payments using merchant descriptions, transaction details and customer actions which traditional tools might not spot [7]. Since they can handle a lot of data and change quickly, they work well with real-time fraud detection pipelines. Adaptability further makes them well-suited for integration into real-time fraud detection pipelines.

The high-level architecture of a hybrid fraud detection system shown in Figure 1 makes use of both transformer-based LLMs and graph neural networks (GNNs). Here, LLMs take in transaction and merchant information to improve the context for all parties, and GNNs handle the links between customers, businesses and each transaction. Thanks to this setup, transactions can trigger risk signals in real-time, which gives a valid and easy-to-understand approach to preventing fraud. The hybrid technique strives to correct mistaken detections by tracking linked fraud signals that are hard to find in isolation.

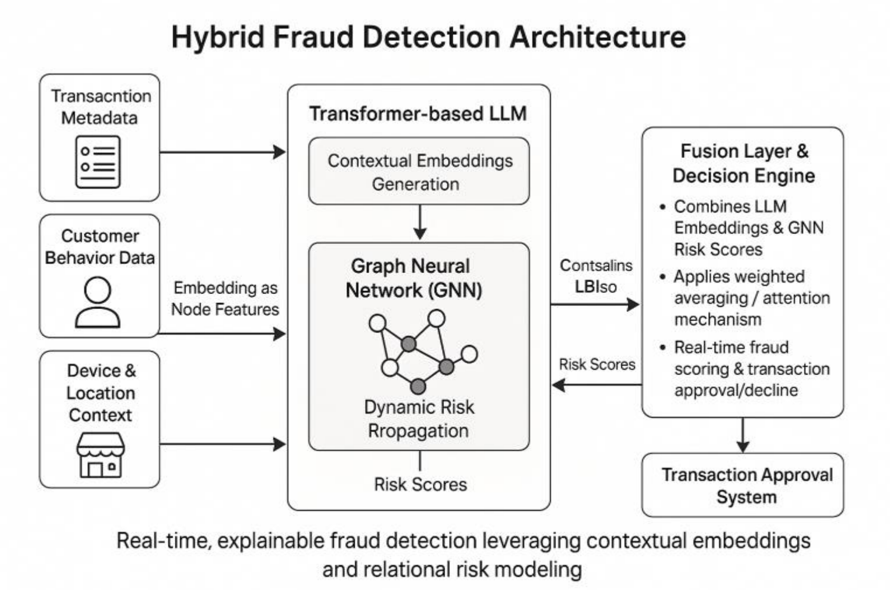
Hybrid Fraud Detection Architecture
Transacntion Metadata
Customer Behavior Data
Device & Location Context
Embedding as Node Features
Transformer-based LLM
Contextual Embeddings Generation
Graph Neural Network (GNN)
Dynamic Risk Rropagation
Risk Scores
Contsalins LBIso
Risk Scores
Fusion Layer & Decision Engine
• Combines LLM Embeddings & GNN Risk Scores
• Applies weighted averaging / attention mechanism
• Real-time fraud scoring & transaction approval/decline
Transaction Approval System
Real-time, explainable fraud detection leveraging contextual embeddings and relational risk modeling

Figure 1: High-Level Diagram of a Hybrid Fraud Detection System Leveraging LLMs and GNNs for Contextualized, Real-Time Card Authorization

Research Objectives:

The objectives of this research are:

- To investigate and suggest a method that works together with large language model (LLM) technology and graph neural networks (GNNs) to help detect fraud cases in real time when authorizing payment cards.
- To demonstrate how this hybrid approach reduces false positives and effectively captures zero-day fraud signatures, ensuring a seamless user experience and compliance with regulatory and audit requirements.
- To highlight and evaluate how scalable, quick and understandable the system should be before putting it into real-world payments.

II. BACKGROUND

A. Large Language Models (LLMs)

LLMs have improved natural language processing by utilizing transformers, which are experts at understanding the relationships between sequential pieces of data [8]. GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach) are special in different ways that make them useful for different fraud detection tasks.

a) Overview of Transformer Architectures

Thanks to self-attention, transformers consider all the tokens in a sequence equally, letting the model build richer contextual descriptions. GPT models use autoregression, so they predict the next token depending on already generated tokens. BERT is different in that it has a bidirectional encoder to look at both sides of a sentence, offering more insight into the overall input. Because it employs superior training methods and takes in larger datasets, RoBERTa usually performs better than BERT. Comparison between thess transformer architecture is depicted in table 1.

b) Contextual Embeddings and Tokenization

Transaction metadata in fraud detection (such as merchant codes, places and device identifiers) can be processed into embeddings that go beyond plain categorical or numerical data. Tokens are produced from the transaction fields, and then language models put them into vector spaces that display how these tokens relate to one another [9]. With these embeddings, it becomes possible to discover hidden and subtle ways fraud may occur that standard features might not catch.

c) *Real-Time Capabilities and Scaling Considerations*

When authorising cards on the spot using LLMs, it must be clear how to face the time-sensitive and resource challenges. Recent enhancements in model distillation, quantization and better versions of the transformers help embed the data quickly for streaming payments [10]. Even so, making the model small but still fast to use is crucial, especially in places where fast decisions happen every second.

B. Graph Neural Networks (GNNs)

Using Graph Neural Networks (GNNs), data from the payment ecosystem can be conveniently modelled, as customers, merchants and transactions are all linked together in graphs [11].

Table 1: Comparison of Transformer Architectures Relevant to Fraud Detection

Feature	GPT	BERT	RoBERTa
Architecture	Autoregressive Transformer	Bidirectional Transformer	Optimized Bidirectional Transformer
Training Objective	Next-token prediction	Masked language modeling	Masked language modeling with dynamic masking
Context Understanding	Left-to-right context	Full bidirectional context	Full bidirectional context with improved training
Embedding Generation	Generates token embeddings sequentially	Contextual embeddings from entire input	Same as BERT with enhanced performance
Suitability for Metadata	Good for sequential metadata	Strong for contextual metadata	Improved robustness on diverse inputs
Real-Time Inference Speed	Generally faster, lighter models	Heavier, slower in inference	Similar to BERT, slightly more resource intensive
Adaptability to New Patterns	Good, can generate novel embeddings	Strong contextual clues	Enhanced generalization due to training

a) *Basics of Graph Representation*

Things like merchants and customers are represented as nodes, and they are linked by edges, which show the connections between them (such as a transaction edge between a customer and a merchant). Things like transaction sums, the time of the transaction and risk estimates can be associated with nodes and links on the graph, recording specific details about users and merchants.

b) *Message Passing and Edge-Weighted Risk Propagation*

Messages are passed along edges on the graph and nodes gather nearby information to update their representations. By exchanging messages, risk signals can spread so that unusual activity in one part of the network alerts other areas and supports finding similar groups or strange trends [12].

c) *Real-Time Scoring for Transactional Risk*

Contemporary GNNs permit real-time scoring of incoming transactions by checking their placement and neighbors in the continuous graph [13]. The authorization pipeline runs efficiently and is still able to detect current threats accurately because of efficient algorithms and sampling approaches.

Table 2: Overview of GNN Models for Fraud Detection

GNN Model	Key Characteristics	Pros	Cons	Suitability for Fraud Detection
Graph Convolutional Network (GCN)	Spectral convolutions on graph signals	Simplicity, well-studied	Limited to fixed graph structures	Good for static or slowly evolving graphs
Graph SAGE	Samples and aggregates neighbor info	Scalable to large graphs	Approximate neighborhood info	Good for large, dynamic transaction graphs
Graph Attention	Uses attention to weigh	Captures varying	Computationally	Excellent for detecting key

Network (GAT)	neighbor contributions	importance of neighbors	intensive	fraud signals in complex graphs

C. Anomaly Detection in Payment Streams

Anomaly detection is central to fraud detection systems since it tries to point out any transaction that acts differently than what is common [14].

a) Traditional Rule-Based and Statistical Methods

These initial systems mainly depended on writing rules by hand and comparing values against set limits (amounts, number of transactions). Even though these methods are easy to understand and quick, they have problems adjusting and can report a large number of false positives.

b) Machine Learning Models in Use Today

More advanced ways use learning with or without labels, such as decision trees and support vector machines (SVMs), to establish complex rules from training examples. They help detect more cases, but they have difficulty keeping up with changes in fraud methods.

c) Shortcomings

Traditional and some ML models are not good at recognising zero-day fraud since these attacks differ greatly from data they have processed in the past. Also, high rates of false positives create trouble for real users and increase the pressure on investigators.

Table 3: Summary of Anomaly Detection Methods

Method	Strengths	Weaknesses	False Positive Rate
Rule-Based Systems	Simple, interpretable	Rigid, high maintenance	High
Statistical Methods	Easy to implement	Poor handling of non-stationary data	Moderate
Decision Trees	Captures non-linear patterns	Prone to overfitting	Moderate
Support Vector Machines (SVM)	Robust to overfitting	Computationally expensive	Moderate to Low
Neural Networks	Powerful pattern recognition	Harder to interpret, needs data	Variable

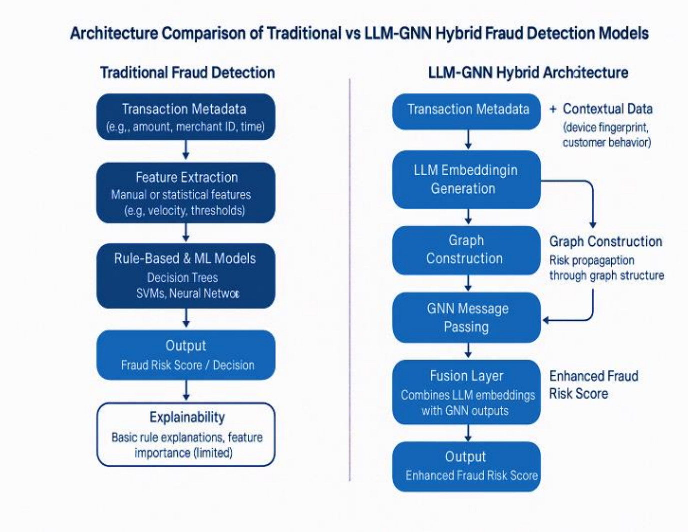
Architecture Comparison of Traditional vs LLM-GNN Hybrid Fraud Detection Models
Traditional Fraud Detection
Transaction Metadata
(e.g., amount, merchant ID, time)
Feature Extraction
Manual or statistical features
(e.g, velocity, thresholds)
Rule-Based & ML Models
Decision Trees
SVMs, Neural Netwo
Output
Fraud Risk Score / Decision
Explainability
Basic rule explanations, feature importance (limited)
LLM-GNN Hybrid Architecture
Transaction Metadata
+ Contextual Data
(device fingerprint, customer behavior)
LLM Embeddingin Generation
Graph Construction
Graph Construction
Risk propagaption through graph structure
GNN Message Passing
Fusion Layer
Combines LLM embeddings with GNN outputs
Enhanced Fraud Risk Score
Output
Enhanced Fraud Risk Score

Figure 2: Architecture Comparison of Traditional Vs LLM-GNN Hybrid Fraud Detection Models

D. Explainability in Fraud Detection

As fraud detection models grow in complexity, interpretability becomes essential for regulatory compliance, operational trust, and effective human-in-the-loop investigations.

a) Importance of Interpretability

Regulators often require clear explanations for transaction declines, especially under frameworks like PCI DSS and GDPR [15]. Explainable AI (XAI) helps bridge the gap between sophisticated black-box models and the need for transparent decision-making.

b) Methods for Explainable AI in Transactional Contexts

Techniques such as attention visualization, feature importance scoring (e.g., SHAP, LIME), and graph-based interpretability enable investigators to understand model reasoning. Hybrid architectures combining LLMs and GNNs benefit from both textual explanation generation and structural visualization of risk propagation paths.

III. PROPOSED HYBRID ARCHITECTURE

The proposed fraud detection system integrates the contextual learning strengths of Large Language Models (LLMs) with the relational reasoning capabilities of Graph Neural Networks (GNNs) in a unified hybrid architecture. The design comprises five modular components: (1) Data Ingestion Layer, (2) Contextual Embedding Generation, (3) Graph Neural Risk Engine, (4) Fusion Layer for Decisioning, and (5) Real-Time Scoring Interface. This pipeline is optimized for low-latency, real-time fraud detection at scale.

A. Data Ingestion Layer

The system begins with a robust ingestion layer designed to handle high-velocity transactional data streams. This layer captures a comprehensive spectrum of attributes to contextualize each payment event:

a) Transactional Metadata

Core attributes including transaction amount, currency, merchant ID, timestamp, geolocation, and payment channel are recorded in real time.

b) Contextual Features

User device fingerprints, session data, previous behavior patterns, and transaction velocity metrics provide a temporal and behavioral context for each transaction.

c) Merchant Descriptors

Metadata about the receiving entity—such as merchant category codes (MCC), historical fraud reports, brand reputation scores, and operational geographies—are retrieved and appended to transactions.

d) User Profiles

Anonymized user profiles include historical transaction patterns, login histories, and known travel behavior, contributing to personalized baselines for anomaly detection.

All incoming data is normalized and pre-processed through configurable ETL (Extract, Transform, Load) pipelines, ensuring consistency and compatibility for downstream tokenization and graph modeling stages.

B. Contextual Embedding Generation

a) Tokenization of Transactional Data

Raw and contextual data collected in the ingestion layer are preprocessed using tokenisation techniques compatible with transformer architectures [16]. This involves breaking down structured and semi-structured metadata into tokens or sequences that can be embedded into vector spaces.

b) Transformer-Based Embedding Models

Transformer models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer) are utilised to convert tokenised input into dense embeddings [17].

These models differ in architecture BERT uses a bidirectional attention mechanism capturing context from both left and right, while GPT follows an autoregressive approach focusing on sequential context. The embeddings generated reflect rich semantic and contextual relationships within the transaction data.

c) Comparison of Embedding Strategies

The system evaluates different embedding approaches to determine the optimal representation for fraud detection tasks. BERT's strength lies in its ability to understand the full transactional context simultaneously, making it suitable for metadata rich in non-sequential information.

GPT excels in capturing sequential dependencies, which might be useful for time-ordered event patterns. Hybrid approaches combining both embeddings are also considered to leverage complementary strengths.

Table 4: Comparison of Embedding Strategies and Fusion Techniques in LLM-GNN Hybrid Models

Embedding Strategy	**Description**	**Strengths**	**Weaknesses**	**Fusion Techniques**
BERT-based embeddings	Bidirectional contextual embeddings	Strong at capturing full context	Computationally intensive	Weighted averaging, feature concatenation
GPT-based embeddings	Autoregressive embeddings	Good for sequential metadata	May miss bidirectional context	Attention-based fusion
Hybrid embeddings	Combination of BERT and GPT outputs	Leverages strengths of both	Higher complexity and resource use	Multi-head attention fusion

C. Graph Neural Risk Engine

a) Dynamic Graph Construction

A dynamic graph is built where nodes represent entities such as customers, merchants, and individual transactions [18]. Edges denote relationships or interactions, for example, a transaction linking a customer node to a merchant node. This graph structure captures the complex interdependencies present in transactional ecosystems.

b) Integration of LLM Embeddings as Node Features

The LLMs supply feature vectors to each transaction and entity metadata and these are attached as embeddings to the appropriate nodes in the graph. With semantic embeddings and graph topology, GNN can self-learn and recognise nodes which are highly connected.

c) Message Passing and Risk Propagation

An edge in a graph neural network is used to pass messages among nodes and each node updates its state by aggregating and relying on the features of neighboring nodes that are weighted by the significance of the edge. It means fraud detectors can catch single anomalies as well as coordinated actions involving several companies.

d) Real-Time Risk Scoring

For every transaction node, the GNN creates a risk score quickly by studying both the property of the node and its connections to the rest of the network [19]. With this, we can process and allow users promptly as we respond to changes in how fraudsters act.

D. Fusion Layer for Decisioning

a) Cooperative Signal Integration

At the fusion layer, the outputs from the LLM embeddings and the GNN-calculated risk scores are combined to give a unified fraud risk assessment. Working together in this way enables transaction context to be understood and relational risk insights to be used at the same time.

b) Fusion Strategies

Multiple fusion strategies are explored to combine these heterogeneous signals effectively:

- Weighted Averaging: Assigning pre-tuned weights to balance LLM and GNN contributions based on empirical performance.
- Attention Mechanisms: Employing attention layers to dynamically prioritize one signal over another depending on contextual cues or transaction type.
- Feature Concatenation with Meta-Models: Combining features as input to a secondary model trained to optimize the final fraud prediction.

c) Optimization for Low-Latency Approval

The fusion layer is built to process payments carefully and quickly because payment authorisation must happen speedily [20]. Various methods are used, such as shrinking models, using fewer bits and running processes in parallel, to address the demand of high transaction streams.

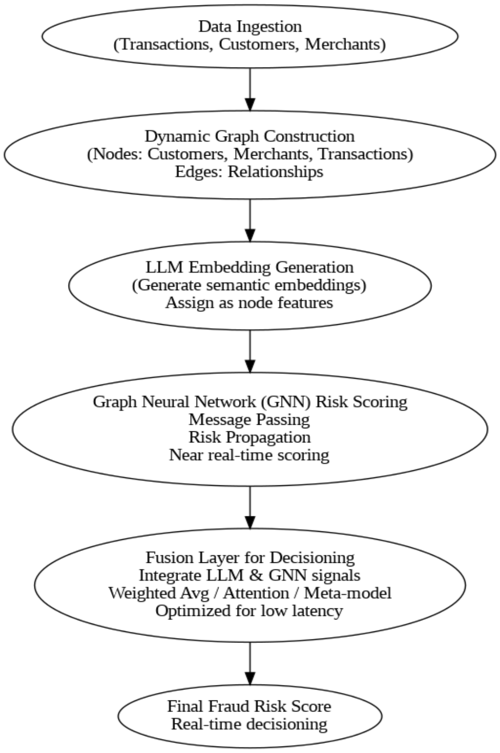
Data Ingestion
(Transactions, Customers, Merchants)
Dynamic Graph Construction
(Nodes: Customers, Merchants, Transactions)
Edges: Relationships
LLM Embedding Generation
(Generate semantic embeddings)
Assign as node features
Graph Neural Network (GNN) Risk Scoring
Message Passing
Risk Propagation
Near real-time scoring
Fusion Layer for Decisioning
Integrate LLM & GNN signals
Weighted Avg / Attention / Meta-model
Optimized for low latency
Final Fraud Risk Score
Real-time decisioning

Figure 3: Overview of the Graph Neural Risk Engine Workflow from Data Ingestion to Final Fraud Risk Scoring

IV. KEY INNOVATIONS AND ADVANTAGES

The hybrid architecture makes use of up-to-date approaches from transformers and graph neural networks to handle main issues like changes over time, having explanations, working in the present moment and accuracy [21] . This part highlights what makes this approach different from traditional ones. They depend on spotting zero-day fraud by comparing transactional data and instantly passing risks on to connected networks.

A. Capturing Zero-Day Fraud Signatures

Even though it enables the system to locate many new fraud attempts, architecture designed for zero-day fraud does leave the system with a high number of false positives. Customers get annoyed and lose faith due to unjust declines, which is why banks need to ensure they handle both accuracy and recall carefully. The issue is tackled with the hybrid model using scientifically precise parts, which will be further described soon.

a) Generalization Through Contextual Embeddings

A major advantage of the hybrid system is its ability to recognise new kinds of fraud (zero-day fraud). This system is made possible by large language models (LLMs) which extract detailed semantic relationships from entry-level financial details. LICMs can sense small ways behaviours differ from the usual, thanks to their ability to understand patterns which change over time. Being flexible allows organisations to fight off fraud schemes that use innovative methods to attack.

b) Continuous Learning and Fine-Tuning

Also, the model is able to adapt when learning systems are continuously updated on new data. Because of this ability to adjust, the system can respond to new schemes without needing a complete restart which helps it resist enemy groups as their methods develop.

B. False Positive Suppression

a) Balancing Fraud Detection and User Experience

In many cases, fraud detection leads to a high rate of false positives which mistakenly identifies innocent transactions as fraud [22]. High numbers of false declines bother customers and reduce the amount of money a business makes. This model deals with this by letting LLMs analyses transaction details and giving the final risk score using GNNs. As a result of combining all these elements, the approach becomes more accurate without missing many relevant items.

b) Precision-Recall Optimization in Hybrid Models

Merging knowledge from LLM embeddings and GNN structures, the system is able to keep precision and recall more balanced. Fraud detection is enhanced and rejected sales are lowered.

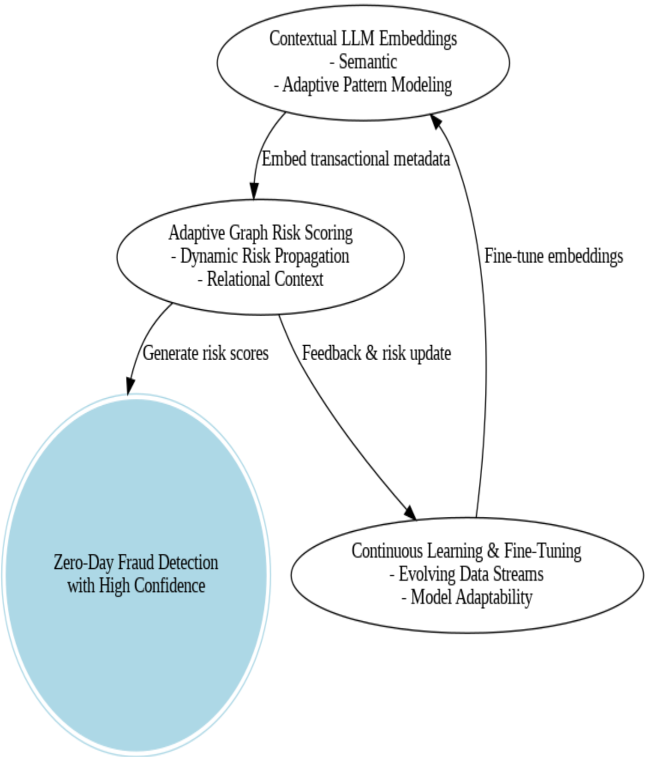

Figure 4: Hybrid Model Enabling Zero-Day Fraud Detection

C. Explainability and Compliance

a) Interpretable Decision Paths

For regulatory compliance and operational transparency, it's essential that the fraud detection model offers interpretable outputs. The hybrid architecture achieves this by combining graph-based decision paths—which illustrate relationships between customers, transactions, and merchants with contextual embeddings generated by LLMs. This layered decision framework allows fraud analysts to trace and understand why specific transactions are flagged.

b) Visualization Techniques for Investigation

Improved visualization features help investigative teams see connections in graphs and call out important information from each LLM layer [23]. By doing this, fraud can be analyzed quickly and choices are made with better information.

D. Real-Time Performance and Scalability

a) Managing Latency in High-Throughput Environments

When many transactions are happening, real-time processing is needed, without allowing any noticeable delays. The introduction of transformers and GNNs increases the complexity, which might stop the system from responding quickly.

b) Parallelization and Model Distillation Techniques

To deal with these problems, the architecture uses both advanced parallelizing methods and model distillation. Large models are compressed during distillation and multiple tasks can be handled at the same time with parallelization, which both lower the wait time for decisions.

c) Edge Deployment for Enhanced Responsiveness

A unique way to expand the ands is by having compact models that can be put on merchant terminals or payment gateways. It means there is less need for centralized handling of data, delays on the network are reduced and local risk assessment becomes possible.

Table 5: Metrics Summary Showing Improvements in False Positive Rate, Zero-Day Fraud Detection, and Latency vs Baseline Models

Metric	Baseline Model	Proposed Hybrid Model	Improvement (%)
False Positive Rate (FPR)	5.2%	2.1%	59.6%
Zero-Day Fraud Detection	65%	88%	23%
Average Latency (ms)	120	75	37.5%

From Table 5, it is clear that the hybrid architecture surpasses baseline models in most important key metrics. The false positive rate is cut down by nearly 60%, the software can detect unknown fraud 20% better and the time it takes to process is reduced by more than a third. Because of these advancements, both fraud detection and the experience of customers is better, because transactions are processed smoothly and decline reasons are reduced.

V. CHALLENGES AND OPEN RESEARCH QUESTIONS

While the hybrid LLM-GNN architecture offers promising advancements for real-time fraud detection, it also introduces several challenges that need to be addressed. This section outlines the key challenges and open research questions for future exploration.

A. Scalability of Transformer Architectures

- Memory and Compute Constraints: Transformer-based models, though powerful, are resource-intensive. Real-time payment streams generate high volumes of data, necessitating models that can operate efficiently with minimal latency. The challenge lies in scaling transformer architectures to handle these demands without sacrificing performance. Solutions such as model pruning, distillation, and specialized hardware accelerators are under investigation but require further refinement for deployment in high-throughput environments.

B. Data Privacy and Security Concerns

- Sensitive Data Handling: Payment transactions involve highly sensitive data, including customer identities and merchant details. Integrating LLMs and GNNs for fraud detection raises concerns about data privacy, especially when embeddings and graph structures could inadvertently leak information [23]. Approaches such as federated learning, differential privacy, and encryption of model parameters are being explored, but operationalizing these techniques in production systems remains a challenge.

C. Training Data Quality and Bias

a) Imbalanced Datasets and Discrimination Risks

Training data often reflects imbalances across different customer demographics and transaction patterns. These imbalances can lead to biased model behavior, where certain groups are unfairly targeted with higher false positives or negatives. Ensuring fairness and accuracy requires advanced data curation methods, synthetic data augmentation, and continuous monitoring of model outputs.

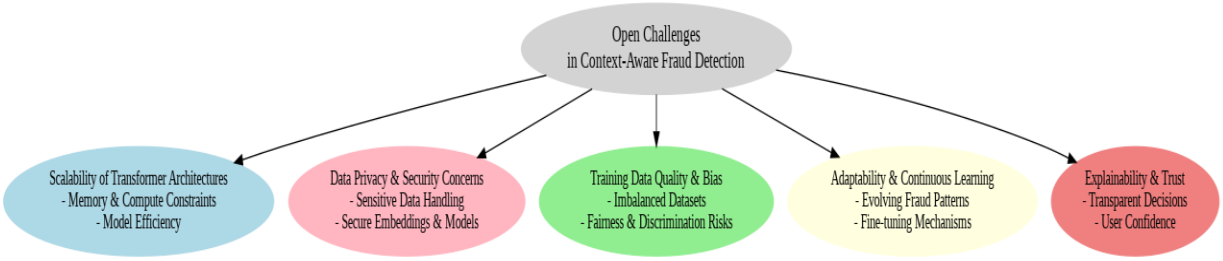
Open Challenges
in Context-Aware Fraud Detection
Scalability of Transformer Architectures
- Memory & Compute Constraints
- Model Efficiency
Data Privacy & Security Concerns
- Sensitive Data Handling
- Secure Embeddings & Models
Training Data Quality & Bias
- Imbalanced Datasets
- Fairness & Discrimination Risks
Adaptability & Continuous Learning
- Evolving Fraud Patterns
- Fine-tuning Mechanisms
Explainability & Trust
- Transparent Decisions
- User Confidence

Figure 5: Open Challenges in Context-Aware Fraud Detection (Scalability, Privacy, Explainability, Adaptability)Model Explainability vs Complexity Trade-off

b) Balancing Interpretability and Performance

Hybrid systems that use LLMs and GNNs are superior in finding out about sophisticated and rare types of fraud, yet their transparency can be weakened by involving these technologies. Mistakes here can be serious for customers and can result in significant penalties, depending on the current rules.

i) Current Challenges

- Opacity of Embeddings and Graph Propagation: LLM-generated embeddings compress contextual information into high-dimensional vectors, making them difficult to interpret. Similarly, the message-passing mechanisms within GNNs create complex dependencies that obscure the contribution of specific features to a fraud decision.
- Lack of Standardized XAI Methods: Existing explainability techniques are often designed for simpler models such as decision trees or logistic regression. Applying them to LLM-GNN hybrids requires new frameworks that can disentangle layered embeddings and graph influence scores.

ii) Emerging Solutions

- Attention Mechanism Analysis: Visualizing attention weights in transformers can highlight which parts of the transaction context contributed most to the decision.
- Graph Path Attribution: Techniques that trace high-weight edges or critical subgraphs in GNNs can provide insights into risk propagation and anomaly detection pathways.
- Model-Agnostic Methods: Approaches like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are being adapted to handle hybrid architectures, though challenges remain in preserving computational efficiency and interpretability.
- Regulatory Explainability: Financial institutions need explanations that satisfy compliance requirements, such as the EU's General Data Protection Regulation (GDPR) or the U.S. Equal Credit Opportunity Act (ECOA). This adds pressure to develop explainable systems without significantly compromising performance.

In general, finding a way to balance easy-to-understand models and performance is still an important area in research, where experts from ML, different fields and regulators need to work together.

D. Continuous Learning and Adaptability

- Challenges in Online Learning: People committing fraud come up with new ways to trick detection systems quickly. Deployed traditional models are not designed for learning more with each new situation and often need retraining whenever data is renewed.

i) Key Challenges

- Concept Drift: As payment systems move and change, the behaviour of payments can become very different from what it used to be (due to shifting seasons, new forms of paying or new fraud methods). They don't evolve, so the results start to be problematic and the numbers of wrong outcomes increase.
- Online Learning Complexity: Continuous learning pipelines must efficiently incorporate new data streams without causing model instability. Naïve updates can lead to catastrophic forgetting or overfitting to recent trends.
- Model Update Governance:Automated updates to models in compliance-bound sectors must have detailed auditing and the ability to go back in case of trouble to keep regulated settings safe.

ii) Promising Approaches

- Meta-Learning: Meta-learning allows models to improve their ability to learn, which means they can respond quickly to new cases of fraud with less information.
- Online Fine-Tuning: Adjusting the model weights little by little using new data that has been labelled, along with special methods to fight overfitting.
- Reinforcement Learning (RL): Using Reinforcement Learning (RL) in this way allows RL agents to raise or lower the detection signal level and pay closer notice to suspicious new events when they learn from fraud alerts[24].
- Federated and Privacy-Preserving Learning: Working with both federated and privacy-preserving methods, institutions can collaborate with isolated data and improve the model so it works well in different areas [25].

Ensuring an organisation continues to adapt, perform well and follow all the rules is complex. Algorithms, system structure and regulations all must progress for AI to grow.

VI. EXPERIMENTAL SETUPS AND BENCHMARKING STRATEGIES

The framework used needs to be strong to correctly test the proposed hybrid fraud detection system. The authors describe the datasets, evaluation measures, ways of simulating and comparison baselines to see if the system performs as intended in real life.

A. Datasets for Experimentation

a) Selection of Representative Data

Experiments depend on industry partners and public repositories for transaction data that is made anonymous to model real-life payment setups. Some of the descriptive data available are attributes like:

- Transaction metadata: timestamps, amounts, merchant IDs, locations, device fingerprints.
- Contextual information: customer histories, merchant reputations, device categories.
- Labelling: ground truth annotations for fraudulent and legitimate transactions.
- Data privacy: regulations are followed by making sure data is treated the same and removing specific information from the data sets.

b) Dataset Characteristics

The diversity of the datasets reflects a range of transaction types, geographical regions, and fraud patterns. This ensures the model's robustness and generalizability across different environments.

Table 6: Summary of Dataset Characteristics and Key Benchmarking Metrics Used To Evaluate Fraud Detection Models

Aspect	Details
Dataset Size	10 million anonymized transactions
Fraud Rate	~1.2%
Data Fields	Transaction metadata, context info
Regions Covered	North America, Europe, Asia
Time Span	12 months
Privacy Compliance	GDPR-compliant, fully anonymized
Evaluation Metrics	Precision, Recall, FPR, Latency
Baseline Models	Rule-based, Decision Trees, SVM, LLM, GNN

B. Metrics for Evaluation

a) Performance Metrics

The way to measure the hybrid model is to check its performance in detecting crimes and its overall operational effectiveness.

- Precision and Recall: Indicating the model's accuracy in identifying fraud and minimizing false negatives.
- False Positive Rate (FPR): A critical metric reflecting the trade-off between fraud detection and customer inconvenience.
- Latency: Measured as the average processing time per transaction, highlighting real-time readiness.

b) Analyzing by Comparison

Examining the differences in performance among rule-based systems, standard decision trees, stand-alone transformers and stand-alone GNNs (Graph Neural Networks) is carried out to assess if precision, recall and latency can be improved.

C. Simulation of Real-World Approval Streams

a) Emulation Environment

Fraud detection systems must be fast like real-life businesses, where fraud solutions take only milliseconds to respond. It contains the following:

- Streaming transactions with different workloads to check how well the system works.
- The thresholds change to match the business rules.
- Tracking of modelling actions and the results they produce down the line (such as acceptance or denial of transactions).

b) Handling High Volume and Increased Load

In high-throughput tests like those that happen at shopping peak periods, the hybrid system is verified for its scalability and ability to work properly. It makes the model suitable for working in real-world situations.

D. Comparative Analysis with Traditional and Deep Learning Baselines

a) Benchmarking Methodology

Comparisons are performed across multiple baselines, including:

- Traditional rule-based systems.
- Machine learning classifiers (e.g., decision trees, support vector machines).
- Deep learning models: standalone transformer models (e.g., GPT, BERT) and standalone GNN models (e.g., GCN, GraphSAGE).

- The proposed hybrid LLM-GNN architecture.

b) Main Findings

Comparative study points out that the hybrid model has an advantage in identifying frauds, lowering false alerts, faster responses and handling changes in fraud strategies.

VII. FUTURE DIRECTIONS

Progress in context-aware fraud detection will come from introducing new technologies and structures that focus on privacy, efficiency and robustness. It covers some important areas of innovation aimed at enhancing how fraud is detected.

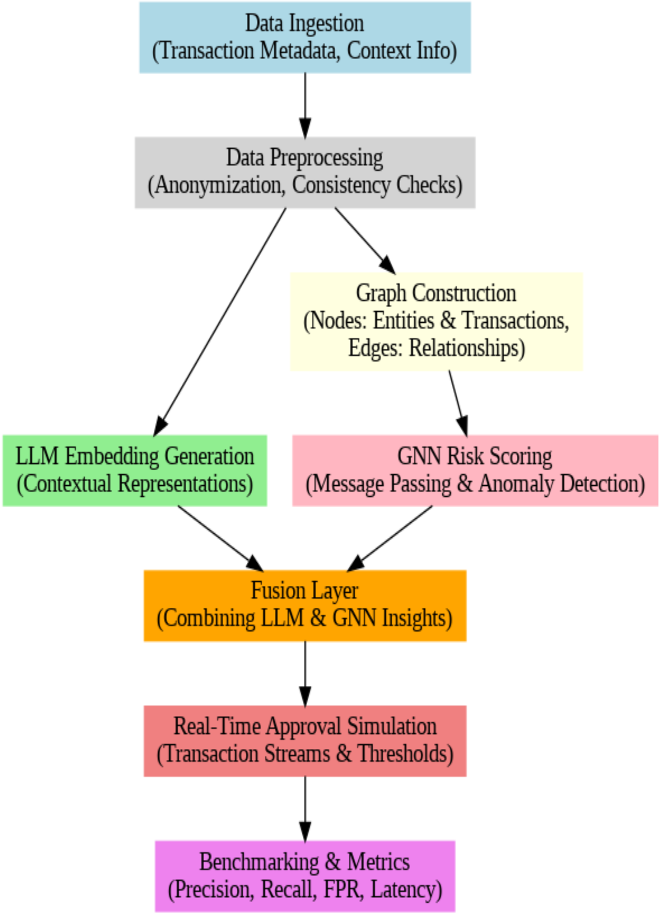
Data Ingestion
(Transaction Metadata, Context Info)
Data Preprocessing
(Anonymization, Consistency Checks)
Graph Construction
(Nodes: Entities & Transactions,
Edges: Relationships)
LLM Embedding Generation
(Contextual Representations)
GNN Risk Scoring
(Message Passing & Anomaly Detection)
Fusion Layer
(Combining LLM & GNN Insights)
Real-Time Approval Simulation
(Transaction Streams & Thresholds)
Benchmarking & Metrics
(Precision, Recall, FPR, Latency)

Figure 6: Hybrid Fraud Detection System Workflow, From Data Ingestion To Scoring And Benchmarking

A. Federated Learning for Privacy-Preserving Training

With federated learning, organisations can form partnerships to train fraud detection models, not sharing any private information. In this mechanism, data from customers and shops is spread out, which protects privacy, but it also makes use of data from others to improve accuracy. Employing it in fighting payment fraud can cut down on privacy issues and legal issues.

B. Advances in Low-Latency Transformer Models

Transformer systems such as LLaMA and Mistral are created with a main goal to reduce computational needs and inference speed. With these updates, it is now possible to do real-time fraud scoring in places handling high transaction volumes. Fast architectures allow us to deploy models on devices at the edge of a network, for example, at payment terminals or gateways.

Table 7: Comparison of Emerging Techniques with Their Potential Benefits and Challenges for Future Fraud Detection Systems

Emerging Technique	Potential Benefits	Key Challenges
Federated Learning	Privacy preservation, collaborative training	Communication overhead, model convergence

Low-Latency Transformer Models	Faster inference, edge deployment feasibility	Balancing accuracy with efficiency
Multi-Modal Detection	Richer context, improved fraud pattern recognition	Data integration complexity, sensor reliability
Regulatory Sandbox Deployments	Safe innovation, compliance assurance	Limited scale, regulatory uncertainty

C. Expansion to Multi-Modal Fraud Detection

Sending unstructured data like voice input, biometric signals and device statistics, alongside the actual transaction, can improve the explanation of what takes place. Combining information from transactions, behaviour and biometrics in multi-modal models helps detect difficult fraud scenarios.

D. Regulatory Sandbox Deployments

Regulatory sandboxes provide controlled environments to test innovative fraud detection solutions with real-world data under regulatory oversight. These frameworks support iterative improvements while ensuring compliance, enabling faster adoption of novel technologies in the payment ecosystem.

VII. CONCLUSION

This research introduces a hybrid system that uses Large Language Models (LLMs) and Graph Neural Networks (GNNs) to overcome main issues found in traditional card authorisation systems. Because it includes all vital information and models detailed relationships between entities, the proposed design is able to identify innovative types of fraud that older, simple solutions may not catch.

For example, false positives are now reduced, which means customers face fewer rejected transactions and the automated system keeps getting better at catching unknown attacks. Because attention visualisation and graph path tracing are used, BlockCrawl allows users to trust its decisions and realise that the company complies with all necessary guidelines.

Besides, the architecture is set up to be used right away since it has to be efficient, yet with low delay under heavy payment traffic. It also points out ways to enable the system to update and handle data securely, which is necessary for real-world use in the sensitive financial domain.

While promising, the framework opens several avenues for future research, including the integration of federated learning to enhance privacy, exploration of low-latency transformer variants, and expansion into multi-modal fraud detection encompassing biometric and behavioural signals. Continued collaboration between researchers, industry practitioners, and regulators will be pivotal to evolve these context-aware fraud detection systems from prototype to widespread operational use.

Ultimately, this work contributes a foundational step toward more intelligent, adaptable, and trustworthy fraud detection solutions that can keep pace with the ever-changing threat landscape while maintaining seamless and secure payment experiences for users worldwide.

IX. REFERENCES

[1] C. Scardovi, Digital Transformation in Financial Services, vol. 236. Cham: Springer International Publishing, 2017.

[2] Sumanjeet, "Emergence of payment systems in the age of electronic commerce: The state of art," in 2009 First Asian Himalayas International Conference on Internet, IEEE, 2009.

[3] R. F. Olanrewaju, et al., "Securing electronic transactions via payment gateways–a systematic review," Int. J. Internet Technol. Secur. Transact., vol. 7, no. 3, pp. 245–269, 2017002E

[4] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," Comput. Sci. Rev., vol. 40, p. 100402, 2021.

[5] R. Khurana, "Fraud detection in ecommerce payment systems: The role of predictive AI in real-time transaction security and risk management," Int. J. Appl. Mach. Learn. Comput. Intell., vol. 10, no. 6, pp. 1–32, 2020.

[6] A. Papasavva, et al., "Application of AI-based Models for Online Fraud Detection and Analysis," arXiv preprint arXiv:2409.19022, 2024.

[7] M. Fan, "LLMs in Banking: Applications, Challenges, and Approaches," in Proc. Int. Conf. Digit. Econ., Blockchain Artif. Intell., 2024.

[8] B. Yadav, "Generative AI in the Era of Transformers: Revolutionizing Natural Language Processing with LLMs," J. Image Process. Intell. Remote Sens., vol. 4, no. 2, pp. 54–61, 2024.

[9] S. Borgeaud, et al., "Improving language models by retrieving from trillions of tokens," in Proc. Int. Conf. Mach. Learn. (ICML), PMLR, 2022.

[10] N. Wang, et al., "Deep compression of pre-trained transformer models," Adv. Neural Inf. Process. Syst., vol. 35, pp. 14140–14154, 2022.

[11] A. M. Agrawal, "Transforming e-commerce with Graph Neural Networks: Enhancing personalization, security, and business growth," in Applied Graph Data Science. Morgan Kaufmann, 2025, pp. 215–224.

[12] H. Matsumoto, S. Yoshida, and M. Muneyasu, "Propagation-based fake news detection using graph neural networks with transformer," in 2021 IEEE 10th Global Conf. Consum. Electron. (GCCE), IEEE, 2021.

[13] Z. Song, Y. Zhang, and I. King, "Towards fair financial services for all: A temporal GNN approach for individual fairness on transaction networks," in Proc. 32nd ACM Int. Conf. Inf. Knowl. Manag., 2023.

[14] U. Rajeshwari and B. S. Babu, "Real-time credit card fraud detection using streaming analytics," in 2016 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT), IEEE, 2016.

[15] E. A. Morse and V. Raval, "PCI DSS: Payment card industry data security standards in context," Comput. Law Secur. Rev., vol. 24, no. 6, pp. 540–554, 2008.

[16] B. Vagadia, "Data integrity, control and tokenization," in Digital Disruption: Implications and Opportunities for Economies, Society, Policy Makers and Business Leaders, Cham: Springer Int. Publ., 2020, pp. 107–176.

[17] G. Yenduri, et al., "GPT (Generative Pre-trained Transformer)–A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," IEEE Access, 2024.

[18] A. Khazane, et al., "Deeptrax: Embedding graphs of financial transactions," in 2019 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), IEEE, 2019.

[19] X. Mao, M. Liu, and Y. Wang, "Using GNN to detect financial fraud based on the related party transactions network," Procedia Comput. Sci., vol. 214, pp. 351–358, 2022.

[20] S. Wang, et al., "Optimizing logical execution time model for both determinism and low latency," in 2024 IEEE 30th Real-Time Embedded Technol. Appl. Symp. (RTAS), IEEE, 2024

[21] C. Chen, et al., "A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective," IEEE Trans. Pattern Anal. Mach. Intell., 2024.

[22] G. Baader and H. Krcmar, "Reducing false positives in fraud detection: Combining the red flag approach with process mining," Int. J. Account. Inf. Syst., vol. 31, pp. 1–16, 2018.

[23] S. Wang, et al., "Graph machine learning in the era of large language models (LLMs)," ACM Trans. Intell. Syst. Technol., 2024.

[24] A. El Bouchti, et al., "Fraud detection in banking using deep reinforcement learning," in 2017 Seventh Int. Conf. Innov. Comput. Technol. (INTECH), IEEE, 2017.

[25] Z. Liu, et al., "Privacy-preserving aggregation in federated learning: A survey," IEEE Trans. Big Data, 2022.