

In-Context Learning for Smarter Fraud Detection in Remote Flea Market Transactions

Hyunwoo Kim
Dangeun Pay Inc.
Seoul, Republic of Korea
peter.kim@daangnpay.com

Hyunmyoung Oh
Dangeun Pay Inc.
Seoul, Republic of Korea
hammer@daangnpay.com

Sunghyon Kyeong*
Dangeun Pay Inc.
Seoul, Republic of Korea
devyn@daangnpay.com

Abstract

this part is not ready yet.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Natural language processing**; • **Information systems** → **Enterprise information systems**.

Keywords

In-context learning, context-based fraud detection, fraud detection, remote secondhand transactions

ACM Reference Format:

Hyunwoo Kim, Hyunmyoung Oh, and Sunghyon Kyeong. 2025. In-Context Learning for Smarter Fraud Detection in Remote Flea Market Transactions. In *6th ACM International Conference on AI in Finance (ICAIF '25)*, November 15–18, 2025, Singapore. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3604237.3626838>

1 Introduction

The global market for secondhand goods has been steadily expanding, driven in large part by the rise of online platforms that facilitate non-face-to-face peer-to-peer transactions. Prominent marketplaces in this domain include Facebook Marketplace (worldwide), Dangeun Market—also known as Karrot—operating in regions such as North America, Korea, and Japan, and Mercari, which is widely used in Japan. These platforms promote the reuse of goods, contributing to environmental sustainability, and attract a growing user base who are motivated by shared social and ecological values.

To ensure secure and convenient transactions for the majority of well-intentioned users, platform providers have implemented protective measures, including escrow-based financial services. Nevertheless, the anonymity and remote nature of these platforms are frequently exploited by malicious actors. For instance, some fraudulent sellers post items at unusually low prices and fail to deliver the products, engaging in what is commonly referred to as merchant fraud. In response, platforms invest substantial effort into detecting and sanctioning such fraudulent activities.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICAIF '25, Singapore

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0240-2/23/11
<https://doi.org/10.1145/3604237.3626838>

Traditional fraud detection systems have largely relied on rule-based approaches or supervised machine learning (ML) models trained on historical transaction data. Despite ongoing model re-training enabled by MLOps platforms, these approaches exhibit clear limitations in rapidly evolving environments like secondhand marketplaces, where contextual factors heavily influence transaction dynamics. In particular, the early detection of novel fraud schemes remains structurally constrained.

Recently, Large Language Models (LLMs) have emerged as a promising alternative to address these limitations. In the context of secondhand trading, LLMs can effectively analyze unstructured textual data—such as listing titles, product descriptions, and seller profiles—to identify suspicious language patterns or detect fraud strategies that resemble previously known cases. Unlike traditional models that rely solely on structured features, LLMs excel at natural language understanding and can capture subtle linguistic cues, inconsistencies in phrasing, and tone variations that might otherwise elude human analysts.

Furthermore, LLMs possess the ability to cross-reference contextual information across multiple transactions. For example, the repeated use of similar phrases, emojis, or urgent language across listings from different user accounts may be linked to a single fraud actor. This capability is significantly enhanced through In-Context Learning (ICL), which enables LLMs to perform fraud detection tasks with minimal examples and without requiring explicit model fine-tuning. ICL is particularly advantageous in scenarios where large-scale labeled datasets are unavailable and where fraud tactics evolve rapidly.

Moreover, the increasing sophistication of fraudsters—who now leverage generative AI to craft convincing phishing messages, fabricate identities, and produce deepfake documents—underscores the urgency for platforms to deploy equally advanced AI-based defense mechanisms. This technological arms race necessitates the adoption of LLM-powered, intelligent fraud detection frameworks.

In this study, we propose a novel fraud detection approach tailored to non-face-to-face secondhand trading environments, leveraging LLM-based In-Context Learning. Specifically, we extract salient features from previously confirmed fraud cases using LLMs to analyze unstructured elements such as listing titles and seller profiles. These features are then compared against ongoing transactions to assess their likelihood of being fraudulent. Finally, we evaluate the effectiveness of our approach in comparison with traditional machine learning-based detection methods to determine its potential performance gains in real-world settings.

2 Related Works

2.1 ML-Based Fraud Detection

A wide range of studies in both industry and academia have sought to advance techniques for financial fraud detection. One line of research focuses on representing transaction histories between bank accounts as graphs, enabling the development of graph-based fraud detection models that significantly outperform traditional baselines in terms of F1 score performance [5, 11].

Simultaneously, increasing attention has been paid to fraud in peer-to-peer transactions within online marketplaces, where financial transactions often accompany interpersonal exchanges. A prominent example is merchant fraud, in which a scammer lists trending products at unusually low prices, receives payment, and fails to deliver the goods. This type of fraud is especially prevalent in remote secondhand platforms. Some studies have addressed this issue by analyzing fraudulent seller accounts and building machine learning-based detection models using features derived from transaction histories and product listings [4, 10].

2.2 LLM-Based Fraud Detection

With the advent of large language models (LLMs), researchers and practitioners have actively explored their potential for financial fraud detection. Traditional methods—such as logistic regression, random forests, and neural networks—have long been applied to detect fraud (e.g., in credit card transactions), but these models face limitations when dealing with highly imbalanced datasets and evolving fraud patterns [12].

Recent studies suggest that Transformer-based LLMs are better suited for capturing long-range dependencies and subtle correlations in transaction data, leading to improved detection performance [3, 7]. For example, Yu et al. (2024) demonstrate that Transformer-based models outperform conventional machine learning approaches in terms of accuracy and are particularly effective at identifying rare fraudulent cases [8, 12]. The pretraining of LLMs on vast corpora enables them to form a form of commonsense understanding of sequences, which can be further enhanced through retrieval-augmented generation (RAG) methods to boost detection capabilities [9].

Moreover, LLMs have proven useful in processing unstructured data alongside structured transactional features. Butler (2025) highlights that LLMs can detect fraud-indicative language and anomalies in textual sources such as transaction notes, emails, and chat logs. This capacity allows them to surface social engineering attempts or abnormal phrasing in online interactions—types of fraud that often evade detection by traditional rule-based or statistical systems.

2.3 In-Context Learning for Fraud Pattern Recognition

In-context learning (ICL) has emerged as a powerful paradigm that enables LLMs to perform tasks without explicit fine-tuning. Introduced by Brown et al. (2020) with the release of GPT-3, ICL allows a model to generalize to new tasks using only a prompt containing a few labeled examples [2]. This characteristic makes ICL particularly well-suited for fraud detection scenarios, where

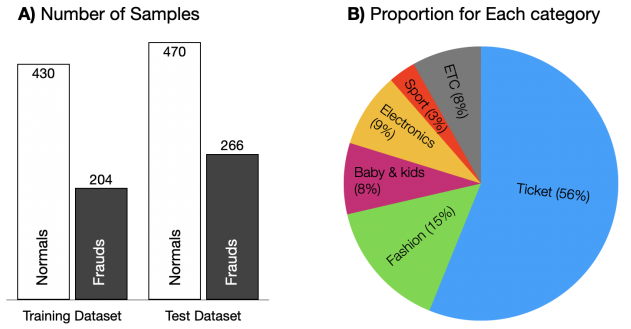


Figure 1: Overview of the sampled dataset with fraud label distribution

only a small number of examples of emerging fraud types may be available.

Through ICL, LLMs can implicitly learn patterns from a few in-context examples and adapt to new fraud types in real time. Liu et al. (2024) apply this concept to graph-based anomaly detection, using a handful of normal nodes as context to identify outliers in unseen graphs without additional training [6]. Similarly, Bhattacharya et al. (2025) propose a system that converts structured transaction features (e.g., amount, location, device information) into natural language descriptions and feeds them into an LLM along with a few labeled examples, enabling accurate classification of novel transactions as fraudulent or legitimate [1].

3 Datasets

This study leverages proprietary real-world transaction data provided by Danggeun Pay Inc., a financial technology company that operates the official payment infrastructure for Danggeun Market Inc.—a widely used local community platform in South Korea. The platform supports a variety of services, including secondhand goods trading, real estate listings, part-time job postings, and more. Within this ecosystem, Danggeun Pay facilitates peer-to-peer (P2P) payments, enabling the collection of fine-grained transactional records that are particularly rich in behavioral signals relevant to fraud detection.

The dataset comprises transaction-level records labeled as either fraudulent or normal. Each transaction is augmented with accompanying listing metadata as well as detailed behavioral features extracted from the seller’s historical activity. The dataset was curated for the express purpose of facilitating machine learning research on fraud detection in P2P commerce and offers a comprehensive foundation for studying behavioral patterns in online trust-mediated environments.

3.1 Training and Test Split

The dataset includes transactions conducted over a two-month period, from April to May 2025. To ensure balanced model training and fair evaluation across fraud classes, stratified sampling based on ground-truth fraud labels was applied. As shown in Fig. 1A, a total of 1,370 transactions were selected, with a class distribution

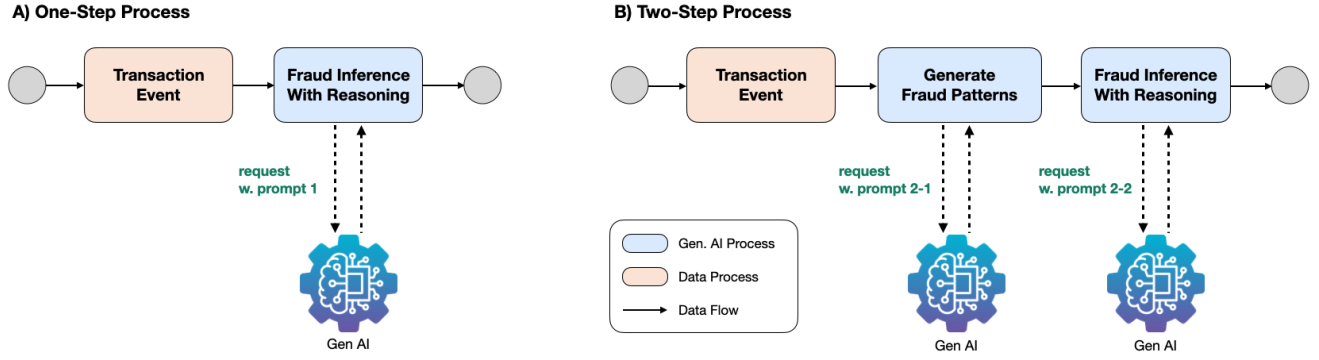


Figure 2: Overview of the experimental processes. A) One-step inference approach. B) Two-step inference approach.

ratio of approximately 1:2 (fraudulent to normal transactions), providing a sufficiently diverse dataset for evaluating fraud detection performance.

The dataset was temporally partitioned based on the transaction date. Transactions that occurred in April 2025 were designated as the training dataset (430 legitimate cases; 204 fraudulent cases), while those from May 2025 were allocated to the test dataset (470 legitimate cases; 266 fraudulent cases). The training dataset was not primarily used for direct model training but rather for extracting recent fraud cases during the experimental process. The test dataset was utilized to evaluate the performance of the proposed fraud detection methods.

3.2 Category-Wise Distribution

The sampled transactions span seven major product categories, each exhibiting distinct fraud risk profiles as shown in Fig. 1B. Table 1 summarizes the distribution of fraudulent and legitimate transactions across these categories.

Notably, categories with high liquidity and resale value—such as tickets—exhibit a disproportionately high rate of fraudulent activity. This heterogeneity underscores the importance of incorporating category-specific behavioral patterns into fraud detection models.

3.3 Feature Overview

Each transaction instance in the dataset is represented by a set of features grouped into four key dimensions:

- **Listing Metadata:** This feature includes the listing title (title), listed price (price), and product category (category).

| Category | Fraud | Normal |
|-------------------------|-------|--------|
| Tickets | 305 | 464 |
| Fashion & Miscellaneous | 25 | 184 |
| Baby & Kids | 78 | 37 |
| Electronics | 37 | 84 |
| Sports | 12 | 32 |
| Others | 13 | 99 |

Table 1: Category-wise distribution of the dataset by fraud label

- **Transaction Details:** This feature captures transaction times-tamp (tx_dttm) and transaction amount (tx_amt).
- **Seller Profile:** This feature includes demographic and account-level attributes such as seller age (seller_age) and account tenure in days (seller_account_tenure).
- **Recent Seller Activity:** This feature summarizes behavioral signals over a 24-hour window preceding the listing. This includes the number of prior transactions (recent_tx_cnt), cumulative transaction amount (recent_tx_amt_sum), and number of unique counterparties (recent_unique_buyers).

These feature groups collectively capture both static attributes and dynamic behavioral cues, facilitating a comprehensive analysis of user behavior for fraud detection.

1. Role

You are a fraud detection expert working at an remote flea market.

2. Evaluation Flow

Step 1.
Step 2.
...

3. Current Transaction

{{current_transaction}}

4. High-risk Indicators

- Mismatch between seller's age and item category
- ...

4. Recent Fraud Examples

{{recent_fraud_examples}}

5. Output Format

Your response must match the following format exactly:

```
```json
{
 "fraud": "Y" or "N",
 "reasoning": "Reason 1 (in Korean) || Reason 2 (in Korean) || Reason 3 (in Korean)"
}
```

Figure 3: Partial description for one-step approach. The red text enclosed in double curly brackets denotes dynamic input parameters that were updated at each LLM invocation.

## 4 Experiments

This study conducted experiments to determine whether each transaction event in the training dataset was fraudulent. Two primary inference approaches were illustrated in Fig. 2.

First, a one-step inference method was applied (Fig. 2A). Prompts were constructed to enable the LLM to infer the likelihood of fraud in the current transaction by comparing it against recent fraud examples and high-risk fraud factors.

Second, a two-step inference method was implemented (Fig. 2B). In the first step, recent fraud examples which inserted into LLM prompts were clustered to identify distinct fraud types and extract representative features for each cluster. In the second step, the model evaluated whether the current transaction exhibited behavioral similarities to the identified fraud patterns, thereby inferring its likelihood of being fraudulent.

### 4.1 One-Step Fraud Inference

In the one-step fraud inference approach, the prompt was constructed by inserting comprehensive contextual information about the transaction under evaluation, recent fraud examples, and a description of high-risk fraud indicators. The large language model (LLM) was then prompted to determine whether or not the transaction under evaluation is fraud. The prompt format used in the one-step fraud inference experiment is illustrated in Fig. 3. In this figure, the red text enclosed in double curly brackets denotes dynamic input parameters that were updated at each LLM invocation.

The `{{current_transaction}}` slot was filled with full contextual information about the transaction, incorporating all features described in Section 3.3. The `{{recent_fraud_examples}}` slot included listing attributes and seller profile information from past transactions that had been labeled as fraudulent.

To investigate the impact of the number of few-shot examples on inference performance, we varied the number of recent fraud examples ( $N$ ) inserted into the prompt. Specifically, for each target transaction, we extracted the most recent  $N$  fraud cases that occurred within the 24-hour period preceding it. Experiments were conducted with  $N = 10, 30, 50, 70, 90$  to evaluate the performance of the one-step inference method under varying prompt sizes.

### 4.2 Two-Step Fraud Inference

In the two-step fraud inference approach, the evaluation of a target transaction's likelihood of being fraudulent was conducted through a two-stage invocation of the LLM (Fig. 2B). In the first step, the LLM was prompted with recent fraud examples to perform fraud clustering and extract representative features for each identified cluster (see the descriptive LLM prompt in Fig. 4A). The fraud features generated in this step—derived in a data-driven manner—were subsequently used as input for the second LLM invocation.

In the second step, the LLM was presented with a new prompt (Fig. 4B) that incorporated both the fraud patterns extracted in the first step and the full contextual information of the target transaction as described in Section 3.3. Based on this prompt, the LLM was tasked with assessing whether the given transaction exhibited fraudulent characteristics or not.

As in the one-step inference approach, we investigated how the number of few-shot examples influenced both the quality of

feature extraction and the performance of fraud inference. For each transaction, we selected the most recent  $N$  fraud cases that occurred within a 24-hour window prior to the transaction. Experiments were conducted with  $N = 10, 30, 50, 70$ , and  $90$  to evaluate the impact of the number of examples on inference performance.

## 5 Experimental Results

this part is not ready yet.

### 5.1 Performance of fine-tuning models

this part is not ready yet.

## 6 Conclusions

this part is not ready yet.

## References

- [1] Indranil Bhattacharya and Ana Mickovic. 2024. Accounting fraud detection using contextual language learning. *International Journal of Accounting Information Systems* 53 (2024), 100682. doi:10.1016/j.jaccinf.2024.100682
- [2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]. <https://arxiv.org/abs/2005.14165>
- [3] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. 2021. Pareto Self-Supervised Training for Few-Shot Learning. arXiv:2104.07841 [cs.CV]. <https://arxiv.org/abs/2104.07841>
- [4] Fahim Hasan, Sourov Kumar Mondal, Md. Rayhan Kabir, Md Abdullah Al Mamun, Nur Salman Rahman, and Md. Sagar Hossen. 2022. E-commerce Merchant Fraud Detection using Machine Learning Approach. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*. 1123–1127. doi:10.1109/ICCES54183.2022.9835868
- [5] Junhong Lin, Xiaojie Guo, Yada Zhu, Samuel Mitchell, Erik Altman, and Julian Shun. 2024. FraudGT: A Simple, Effective, and Efficient Graph Transformer for Financial Fraud Detection. In *Proceedings of the 5th ACM International Conference on AI in Finance (Brooklyn, NY, USA) (ICAIF '24)*. Association for Computing Machinery, New York, NY, USA, 292–300. doi:10.1145/3677052.3698648
- [6] Shuo Liu, Di Yao, Lanting Fang, Zhetao Li, Wenbin Li, Kaiyu Feng, XiaoWen Ji, and Jingping Bi. 2024. AnomalyLLM: Few-shot Anomaly Edge Detection for Dynamic Graphs using Large Language Models. arXiv:2405.07626 [cs.LG]. <https://arxiv.org/abs/2405.07626>
- [7] Xiaopeng Liu, Yan Liu, Meng Zhang, Xianzhong Chen, and Jiangyun Li. 2019. Improving Stockline Detection of Radar Sensor Array Systems in Blast Furnaces Using a Novel Encoder–Decoder Architecture. *Sensors* 19, 16 (2019), 3470. doi:10.3390/s19163470
- [8] Weimin Lyu, Songzhu Zheng, Lu Pang, Haibin Ling, and Chao Chen. 2023. Attention-Enhancing Backdoor Attacks Against BERT-based Models. arXiv:2310.14480 [cs.LG]. <https://arxiv.org/abs/2310.14480>
- [9] Anubha Pandey. 2024. Retrieval Augmented Fraud Detection. In *Proceedings of the 5th ACM International Conference on AI in Finance (Brooklyn, NY, USA) (ICAIF '24)*. Association for Computing Machinery, New York, NY, USA, 328–335. doi:10.1145/3677052.3698692
- [10] Shini Renjith. 2018. Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach. *International Journal of Engineering Trends and Technology (IJETT)* 57, 1 (March 2018), 48–53. doi:10.14445/22315381/IJETT-V57P210
- [11] Yeeun Yoo, Jinho Shin, and Sunghyon Kyeong. 2023. Medicare Fraud Detection Using Graph Analysis: A Comparative Study of Machine Learning and Graph Neural Networks. *IEEE Access* 11 (2023), 88278–88294. doi:10.1109/ACCESS.2023.3305962
- [12] Chang Yu, Yongshun Xu, Jin Cao, Ye Zhang, Yixin Jin, and Mengran Zhu. 2024. Credit Card Fraud Detection Using Advanced Transformer Model. In *2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom)*. 343–350. doi:10.1109/MetaCom62920.2024.00064

**A) Extract Fraud Patterns (Prompt 2-1)**

**1. Role**  
You are a fraud detection expert working at an online second-hand marketplace.

**2. Instructions**  
Analyze the recent fraud cases and identify clusters of fraud patterns based on at least two features per cluster.  
...

**3. Recent Fraud Examples**  
`{{recent_fraud_examples}}`

**4. Output Format**  
Your response must match the following format exactly:  
`` `json  
{  
 "clusters": [  
 {  
 "fraud\_type": "Name of Fraud Pattern Cluster 1 (in English)",  
 "core\_features": "Feature 1 (in English) | Feature 2 (in English) | ..."  
 },  
 {  
 "fraud\_type": "Name of Fraud Pattern Cluster 2 (in English)",  
 "core\_features": "Feature 1 (in English) | Feature 2 (in English) | ..."  
 },  
 ...  
 ]  
},  
`` `

**B) Fraud Inference (Prompt 2-2)**

**1. Role**  
You are a financial fraud detection expert working for an online secondhand marketplace.

**2. Evaluation Flow**  
Step 1:  
Step 2:  
...

**3. Current Transaction**  
`{{current_transaction}}`

**4. High Risk Indicator**  
- Similarities between the current transaction and known fraud patterns  
- ...

**5. Fraud Patterns**  
`{{response_from_prompt_1}}`

**7. Output Format**  
You are a JSON-only responder. Your task is to evaluate whether the current transaction is fraud or not based on the provided input and reasoning criteria.  
`` `json  
{  
 "fraud": "Y" or "N",  
 "reasoning": "이유1 (in Korean) || 이유2 (in Korean) || 이유3 (in Korean)"  
},  
`` `

**Figure 4: Illustrative description for Two-step prompts. A) First step prompt to cluster frauds and extract features. B) Second step prompt to determine whether the current transaction is fraudulent. The red text enclosed in double curly brackets denotes dynamic input parameters that were updated at each LLM invocation.**