In-Context Learning for Enhanced Fraud Detection in Second-Hand Marketplaces

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

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

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

Abstract

As remote second-hand trading platforms expand their user base, a small but inevitable proportion of fraudulent users emerges despite ongoing efforts by platform operators to maintain secure trading environments. While traditional machine learning approaches have been deployed to address such activities, they struggle to adapt to rapidly evolving fraud patterns. This study proposes a novel fraud detection framework leveraging Large Language Models (LLMs) and In-Context Learning (ICL) to enhance detection capabilities in these dynamic environments. We investigate two distinct inference approaches: a one-step method that directly classifies transactions using contextual information and recent fraud examples, and a two-step method that first extracts fraud patterns before making classification decisions. Our experiments utilize real-world transaction data from Danggeun Pay, comprising 1,370 transactions across seven product categories with a 1:2 fraud-to-legitimate ratio. We systematically evaluate three state-of-the-art proprietary LLMs—GPT-4.1, Gemini 2.5 Flash, and Claude Sonnet 4—while varying the number of few-shot fraud examples from 10 to 90. Results demonstrate that GPT-4.1 achieves the highest F1-score of 78.9% in the two-step inference setting with only 10 recent fraud examples, while optimal configurations vary significantly across models and performance metrics. Gemini 2.5 Flash achieves the highest precision (84.3%), and Claude Sonnet 4 demonstrates superior recall (95.5%). These findings confirm that LLM-based ICL can effectively detect fraud without requiring model fine-tuning or extensive labeled datasets, offering a scalable solution for financial technology companies. The framework's ability to adapt to emerging fraud patterns through dynamic few-shot learning makes it particularly valuable in rapidly evolving fraud landscapes.

CCS Concepts

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

*Corresponding author

ICAIF '25, Singapore

ACM ISBN 979-8-4007-0240-2/23/11
https://doi.org/10.1145/3604237.3626838

Keywords

financial fraud detection, context-based fraud detection, in-context learning, remote second-hand transactions

ACM Reference Format:
Hyunwoo Kim, Hyunmyoung Oh, and Sunghyon Kyeong. 2025. In-Context Learning for Enhanced Fraud Detection in Second-Hand Marketplaces. In *6th ACM International Conference on AI in Finance (ICAIF '25), November 15–18, 2025, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3604237.3626838

1 Introduction

The global market for second-hand products has been steadily expanding, driven in large part by the rise of online platforms that facilitate peer-to-peer transactions. Prominent marketplaces in this domain include Facebook Marketplace (worldwide), Danggeun 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 generating positive social and ecological values [9, 11].

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 second-hand marketplaces are frequently exploited by malicious actors [8, 10]. 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.

Traditional fraud detection systems have largely relied on rule-based approaches or supervised machine learning (ML) models trained on historical transaction data [15, 23]. Despite ongoing model retraining enabled by MLops platforms, these approaches exhibit clear limitations in rapidly evolving environments like second-hand 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 in the financial domain as a promising alternative to address these limitations [13, 14, 16]. In the context of second-hand 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 [1, 5, 25].

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 [21]. 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 [28]. ICL is particularly advantageous in scenarios where large-scale labeled datasets are unavailable and where fraud tactics evolve rapidly.

However, the increasing sophistication of fraudsters—who now leverage generative artificial intelligence (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 framework tailored to non-face-to-face second-hand trading environments, leveraging LLM-based In-Context Learning. We investigate two distinct inference approaches: a one-step method that directly classifies transactions using contextual information and recent fraud examples, and a two-step method that first extracts fraud patterns from historical cases before making classification decisions. To comprehensively evaluate our approach, we conduct experiments across three state-of-the-art proprietary LLMs—GPT-4.1, Gemini 2.5 Flash, and Claude Sonnet 4.0—while systematically varying the number of few-shot examples to assess their impact on detection performance. Our experimental results demonstrate the effectiveness of LLM-based In-Context Learning for fraud detection in peer-to-peer marketplace environments.

2 Related Works

The evolution of fraud detection has progressed from traditional machine learning approaches to sophisticated LLM-based methods, with recent advances in In-Context Learning opening new possibilities for adaptive fraud detection systems. This section reviews the relevant literature across three key domains: established ML-based fraud detection techniques, emerging LLM-based approaches for financial fraud detection, and the application of In-Context Learning for fraud pattern recognition. These foundational works inform our proposed framework and provide context for the advantages of LLM-based In-Context Learning in rapidly evolving fraud landscapes.

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 [17, 26].

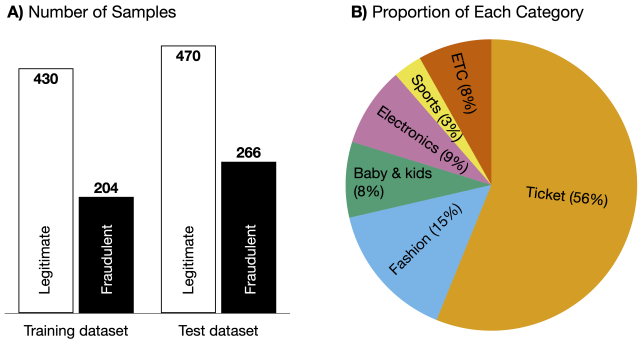

Figure 1: Overview of the sampled dataset with fraud label distribution. (A) Temporal distribution showing training dataset (April 2025: 430 legitimate, 204 fraudulent) and test dataset (May 2025: 470 legitimate, 266 fraudulent cases). (B) Category-wise distribution across seven product categories with corresponding fraud rates, highlighting the heterogeneous risk profiles across different product types.

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 second-hand 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 [12, 24].

2.2 LLM-Based Fraud Detection

With the advent of 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 [27].

Category	Fraudulent	Legitimate
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 fraudulent and legitimate transactions across seven product categories in the Danggeun Pay dataset (N=1,370 transactions, April-May 2025).

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 [7, 19]. 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 [20, 27]. The pretraining of LLMs on vast corpora enables them to develop commonsense understanding of sequences, which can be further enhanced through retrieval-augmented generation (RAG) methods to boost detection capabilities [22].

Beyond their superior performance on structured data, LLMs excel at understanding long context [2, 6] and processing unstructured contextual information alongside structured transactional features. This capability enables them to detect fraud-indicative language patterns and anomalies across extended conversation histories, transaction notes, emails, and chat logs. By analyzing contextual cues and communication patterns, LLMs can surface social engineering attempts and 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

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 [4]. This characteristic makes ICL particularly well-suited for fraud detection scenarios, where 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 [18]. 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 [3].

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 second-hand 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 legitimate. 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 ratio of approximately 1:2 (fraudulent to legitimate 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.

Table 1 shows the number of fraudulent and legitimate cases for each category, highlighting the heterogeneous fraud risk profiles. Tickets category exhibits the highest fraud rate (39.6%), while Electronics and Sports categories show relatively lower fraud rates (30.6% and 27.3%, respectively), demonstrating significant variability in fraud vulnerability across different product types.

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, listed price, and product category.
- **Transaction Details**: This feature captures transaction timestamp and transaction amount.
- **Seller Profile**: This feature includes demographic and account-level attributes such as seller age and account tenure in days.
- **Recent Seller Activity**: This feature summarizes behavioral signals over a 24-hour window preceding the listing. This includes the number of prior transactions, cumulative transaction amount, and number of unique counterparties.

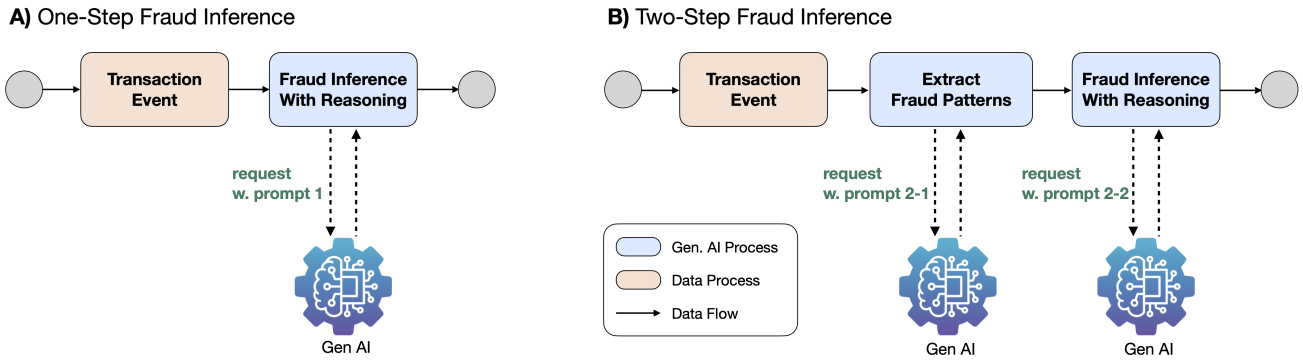

Figure 2: Overview of the proposed LLM-based fraud detection framework. (A) One-step inference approach that directly classifies transactions by incorporating target transaction details, recent fraud examples, and contextual information in a single prompt (detailed prompt structure shown in Fig. 3). (B) Two-step inference approach that first extracts fraud patterns from historical cases, then uses these patterns along with target transaction information for classification (detailed prompt structures shown in Fig. 4). Both approaches leverage dynamic few-shot learning with temporally sampled fraud examples.

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

4 Experiments

This study investigates whether an LLM can accurately classify whether the given financial transaction was fraudulent or not, given rich contextual information specific to peer-to-peer remote second-hand marketplaces. To this end, we designed a series of experiments aimed at evaluating the LLM's ability to make context-aware inferences about the legitimacy of each transaction event.

Two primary inference approaches were explored as illustrated in Fig. 2. In the one-step inference method, prompts were constructed to directly assess whether a given transaction was fraudulent. Each prompt embedded comprehensive contextual information about the target transaction along with recent fraud examples, enabling the LLM to leverage both transaction-specific features and patterns from past fraud cases.

In the two-step inference method, the fraud detection process was decomposed into two stages. In the first stage, the LLM was prompted with recent fraud examples to identify distinct fraud clusters and extract representative features for each group. In the second stage, these extracted fraud patterns were combined with the full contextual information of the target transaction in a new prompt, allowing the LLM to determine whether the transaction aligned with any known fraud patterns.

4.1 Proprietary LLM

We conducted experiments to evaluate whether state-of-the-art proprietary LLMs can classify whether the given transaction was fraudulent or not by leveraging in-context learning, using real-world examples of recent fraudulent transactions along with rich contextual information surrounding the current transaction. Specifically, we investigated the extent to which each model could comprehend the nuances of financial fraud scenarios and accurately infer whether the target transaction was fraudulent or not in a context-sensitive manner. Three leading proprietary LLMs were selected for this evaluation: OpenAI's GPT 4.1, Google's Gemini 2.5 Flash, and Anthropic's Claude Sonnet 4.0. These models were tested under identical prompt structures and inference conditions to ensure a fair comparison of their reasoning capabilities and context sensitivity in the domain of peer-to-peer financial fraud detection.

4.2 Dynamic Sampling of Fraud Examples

Fraud examples were constructed by converting the listing metadata and seller profile information of transactions labeled as fraudulent in the full dataset into a structured textual format. These examples served as representative few-shot instances for prompt-based inference. Below are two illustrative examples:

- Title of Listing: iPhone 7 Silver, Category: electronics, Listed Price: 50,000, Seller Age: 15, Seller Account Tenure: 18 days
- Title of Listing: Stokke Tripp Trapp Newborn Set (Baby Chair), Category: baby and kids, Listed Price: 130,000, Seller Age: 16, Seller Account Tenure: 1 day

These examples were dynamically inserted into the prompts used in both the one-step (see Sections 4.3 and Fig. 3) and two-step (see Section 4.4 and Fig. 4) inference experiments. For each target transaction, we chronologically sorted the full set of fraudulent transactions and selected the N most recent cases that occurred within the 24 hours preceding the transaction's timestamp. These time-filtered examples provided temporally relevant fraud signals for the LLM to reference during inference.

This dynamic sampling strategy was designed to improve the model's ability to detect emerging fraud tactics by leveraging temporally proximate examples that reflect recent behavioral patterns in peer-to-peer marketplaces.

4.3 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 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 metadata 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.

The experiments were independently conducted for each of the three proprietary LLMs—GPT 4.1, Gemini 2.5 Flash, and Claude Sonnet 4.0—by repeatedly applying the same protocol across varying values of N (10, 30, 50, 70, 90). This ensured a consistent and comparative evaluation of the one-step inference performance across all models under different prompt sizes.

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: Prompt structure for the one-step fraud inference approach. The prompt incorporates comprehensive contextual information about the target transaction, recent fraud examples, and high-risk fraud indicators. The red text enclosed in double curly brackets (e.g., `{{current_transaction}}`, `{{recent_fraud_examples}}`) denotes dynamic input parameters that were updated at each LLM invocation based on the specific transaction being evaluated.

4.4 Two-Step Fraud Inference

In the two-step fraud inference approach, the evaluation of whether a target transaction was fraudulent or not 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.

The same two-step inference procedure was independently applied to all three proprietary LLMs—GPT 4.1, Gemini 2.5 Flash, and Claude Sonnet 4.0—repeatedly across different values of N (10, 30, 50, 70, and 90), enabling a model-by-model comparison under identical experimental conditions.

4.5 Evaluation Metrics

To comprehensively assess the performance of the proposed fraud inference approaches, we report three widely adopted evaluation metrics: precision, recall, and F1-score. These metrics are particularly important in real-world fraud detection scenarios, where both false positives (misclassifying a legitimate transaction as fraudulent) and false negatives (failing to detect an actual fraud) entail significant practical consequences. Precision quantifies the proportion of transactions predicted as fraudulent that are indeed fraudulent. It reflects the model's effectiveness in minimizing false alarms and is particularly valuable when the cost of incorrectly flagging legitimate users is high.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

Recall measures the proportion of actual fraudulent transactions that are correctly identified by the model. It captures the model's sensitivity to fraud cases and is especially critical when the cost of missing fraudulent behavior is high.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

F1-score is the harmonic mean of precision and recall, providing a balanced metric that accounts for both types of classification errors. It is particularly suitable when neither precision nor recall can be compromised, as is often the case in fraud detection systems.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

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_2_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": "Reason 1 (in Korean) || Reason 2 (in Korean) || Reason 3 (in Korean)"
}
```

Figure 4: Prompt structures for the two-step fraud inference approach. (A) The first-step prompt is designed to cluster recent fraud examples and extract representative fraud patterns from historical cases. (B) The second-step prompt incorporates both the extracted fraud patterns from step A and full contextual information of the target transaction to make final classification decisions. The red text enclosed in double curly brackets represents dynamic input parameters that are updated for each LLM invocation. The {{recent_fraud_examples}} parameter contains listing metadata and seller profile information from previously labeled fraudulent transactions, {{current_transaction}} includes comprehensive contextual information about the target transaction under evaluation, and {{response_from_prompt_2_1}} encompasses the fraud pattern information generated from the first step (i.e., output of Prompt 2-1).

5 Experimental Results

We conducted comprehensive experiments to evaluate the effectiveness of our proposed LLM-based fraud detection framework across both one-step and two-step inference approaches. Our evaluation utilized the Danggeun Pay dataset described in Section 3, systematically varying the number of few-shot fraud examples from 10 to 90 across three state-of-the-art proprietary LLMs. The results, summarized in Fig. 5, demonstrate significant performance variations across models, inference approaches, and the number of contextual examples provided. Notably, GPT-4.1 consistently outperformed other models in both inference settings, while the optimal number of few-shot examples varied considerably across different LLMs and performance metrics.

5.1 One-Step Fraud Inference

In the one-step fraud inference experiments, a single prompt was constructed by embedding comprehensive contextual information about the target transaction, recent fraud examples, and a description of high-risk fraud indicators. The LLM was then asked to classify whether the given transaction was fraudulent or not. To evaluate the impact of the number of fraudulent examples provided as contextual input, we varied this number and averaged the resulting F1-scores across trials. The average F1-scores across all values of N were as follows: GPT-4.1 achieved 73.9%, Gemini 2.5 Flash reached 70.3%, and Claude Sonnet 4 recorded 63.0%. For GPT-4.1, the highest F1-score (77.5%) was achieved with 10 recent fraud examples. Gemini 2.5 Flash performed best with 70 recent fraud examples, yielding an F1-score of 71.4%, while Claude Sonnet 4 achieved its peak performance of 64.0% at 90 recent fraud examples.

5.2 Two-Step Fraud Inference

In the two-step fraud inference experiments, classification of a transaction as fraudulent or not was performed using a two-stage LLM prompting strategy. In the first stage, the model was prompted with recent fraud examples to identify fraud clusters and extract representative fraud features. In the second stage, the model was presented with both these extracted patterns and the full contextual information of the target transaction to make a binary classification decision. As in the one-step setting, we varied the number of recent fraud examples and measured F1-scores across repeated trials. The average F1-scores across all values of N were as follows: GPT-4.1 at 76.6%, Gemini 2.5 Flash at 60.0%, and Claude Sonnet 4 at 49.2%. GPT-4.1 achieved its highest F1-score (78.9%) at $N = 10$. Gemini 2.5 Flash performed best at $N = 70$, with 61.6%, while Claude Sonnet 4.0 peaked at $N = 10$ with 54.1%.

5.3 Metric-Wise Best Performing LLM

We identified the best-performing configuration for each proprietary LLM with respect to the individual evaluation metrics of precision, recall, and F1-score. The highest precision was achieved by Gemini 2.5 Flash in the two-step fraud inference setting, where 70 recent fraud examples were provided as contextual input. Under

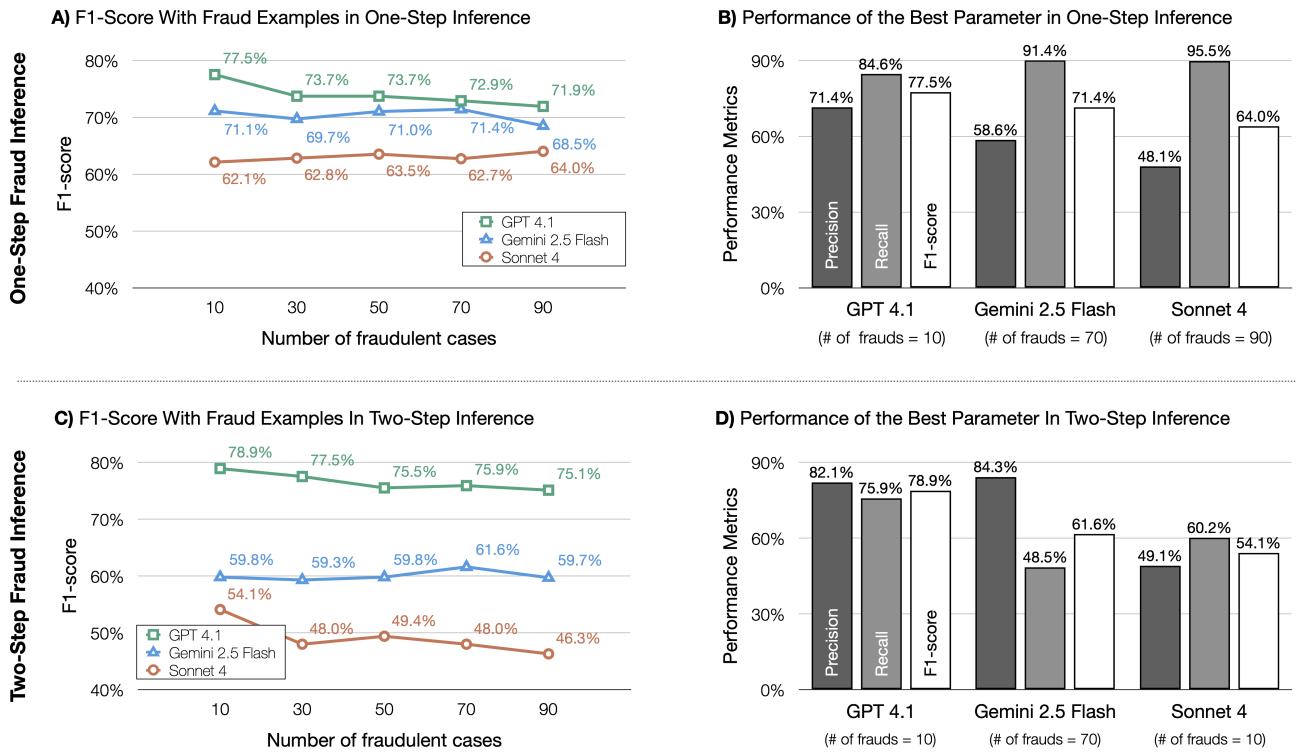

Figure 5: Experimental results comparing one-step and two-step fraud inference approaches across three proprietary LLMs. (A) One-step inference results showing F1-score performance for GPT-4.1, Gemini 2.5 Flash, and Claude Sonnet 4 with varying numbers of few-shot fraud examples (N=10, 30, 50, 70, 90). (B) Best performing LLM for each metric in one-step inference. (C) Two-step inference results demonstrating the impact of fraud pattern extraction on classification performance across the same experimental conditions with (A). (D) Best performing LLM for each metric in two-step inference.

this configuration, the model attained a precision of 84.3%, indicating strong capability in minimizing false positives and correctly identifying legitimate transactions.

The best recall was observed for Claude Sonnet 4.0 in the one-step fraud inference setting, with 90 recent fraud examples included in the prompt. This setup resulted in a recall of 95.5%, reflecting the model's high sensitivity to fraudulent transactions and effectiveness in minimizing false negatives.

The highest overall F1-score, the harmonic mean of precision and recall, was recorded by GPT-4.1 in the two-step fraud inference setting using only 10 recent fraud examples. This configuration yielded an F1-score of 78.9%, demonstrating a strong balance between detecting fraud and avoiding false positives.

6 Conclusions

This study demonstrates the effectiveness of LLM-based In-Context Learning for fraud detection in remote second-hand marketplace environments. Our comprehensive evaluation across three state-of-the-art proprietary LLMs—GPT-4.1, Gemini 2.5 Flash, and Claude Sonnet 4—reveals that both one-step and two-step inference approaches can achieve substantial fraud detection performance without requiring model fine-tuning or extensive labeled datasets.

The experimental results show that GPT-4.1 consistently outperforms other models, achieving the highest F1-score of 78.9% in the two-step inference setting with only 10 recent fraud examples. Notably, the optimal configuration varies significantly across models and performance metrics: while GPT-4.1 excels in F1-score performance, Gemini 2.5 Flash achieves the highest precision (84.3%), and Claude Sonnet 4 demonstrates superior recall (95.5%). These performance variations are illustrated in Fig. 5, which demonstrates the model-specific optimization requirements. These findings confirm that the number of contextual fraud examples and the inference strategy must be carefully calibrated based on the specific LLM architecture and desired performance characteristics.

The practical implications of this research extend beyond academic interest, offering a scalable solution for financial technology companies operating peer-to-peer marketplaces. Unlike traditional machine learning approaches that require extensive feature engineering and continuous retraining, our LLM-based framework can adapt to emerging fraud patterns through dynamic few-shot learning, making it particularly valuable in rapidly evolving fraud landscapes. The temporal sampling strategy employed in our experiments—extracting recent fraud cases within 24-hour windows—provides a realistic approach to maintaining contextual relevance while managing computational costs.

Additionally, our experiments requested not only binary fraud classification (Y/N) but also reasoning explanations for each decision. While the accuracy of these reasoning explanations has not yet been systematically evaluated, they present significant potential for practical application in fraud detection system (FDS) monitoring tasks within financial companies, enabling analysts to better understand and validate automated decisions. A comprehensive qualitative evaluation of these reasoning capabilities represents an important avenue for future research.

However, several limitations require consideration for future research. The reliance on proprietary LLMs may present scalability and cost challenges for widespread deployment, suggesting the need for investigation into open-source alternatives. Additionally, the evaluation was conducted on a single platform's data, and cross-platform generalizability requires validation. Future work should investigate the robustness of these approaches against adversarial attacks, as sophisticated fraudsters may attempt to exploit LLM-based detection systems. Furthermore, model-specific prompt engineering strategies merit further investigation, as our results suggest that different LLMs may benefit from tailored prompt structures, instruction formats, and contextual information presentation methods to maximize their respective strengths in fraud detection tasks.

References

[1] Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-Shot In-Context Learning. In *Advances in Neural Information Processing Systems*, Vol. 37. Curran Associates, Inc., 76930–76966. doi:10.48550/arXiv.2404.11018

[2] Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2025. In-Context Learning with Long-Context Models: An In-Depth Exploration. arXiv:2405.00200

[3] 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.accinf.2024.100682

[4] 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

[5] Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs Are Few-Shot In-Context Low-Resource Language Learners. arXiv:2403.16512

[6] Longze Chen, Ziqiang Liu, Wanwei He, Yunshui Li, Run Luo, and Min Yang. 2024. Long Context is Not Long at All: A Prospector of Long-Dependency Data for Large Language Models. arXiv:2405.17915

[7] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. 2021. Pareto Self-Supervised Training for Few-Shot Learning. arXiv:2104.07841

[8] Zhuoli Chen, Shunan Guo, and Zonglin Mo. 2024. *How Second-Hand Trading Platforms Facilitate Online Fraud: A Case Study of China's Largest Second-Hand Marketplace.* Ph. D. Dissertation. The University of Hong Kong. http://hdl.handle.net/10722/352837

[9] Suvrat Dhanorkar. 2019. Environmental Benefits of Internet-Enabled C2C Closed-Loop Supply Chains: A Quasi-Experimental Study of Craigslist. *Management Science* 65, 2 (2019), 660–680. doi:10.1287/mnsc.2017.2963

[10] Patrik Gerdelius and Hugo Sjönneby. 2024. *Detecting Fraudulent User Behaviour: A Study of User Behaviour and Machine Learning in Fraud Detection.* Technical Report UPTEC STS 24003. Uppsala University, Analysis and Partial Differential Equations. 41 pages.

[11] Moacir Godinho Filho, Gilberto Miller Devós Ganga, Fabiana Leticia Lizarelli, Claudia Lorena Cárdenas Blaz, and Thais Moreira Tavares. 2024. Circular Economy via Chat: Evaluation of Adoption and Use of WhatsApp Instant Messaging Platform for Trading Second-Hand Products. *Journal of Cleaner Production* 460 (2024), 142510. doi:10.1016/j.jclepro.2024.142510

[12] 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

[13] Allen H. Huang, Hui Wang, and Yi Yang. 2023. FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research* 40, 2 (2023), 806–841. doi:10.1111/1911-3846.12832

[14] Wenxi Huang, Zhangyi Zhao, Xiaojun Chen, Qin Zhang, Mark Junjie Li, Hanjing Su, and Qingyao Wu. 2024. A Payment Transaction Pre-training Model for Fraud Transaction Detection. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24).* Association for Computing Machinery, New York, NY, USA, 932–941. doi:10.1145/3627673.3679670

[15] Arshiya Khanum, K S Chaitra, Brijesh Singh, and C Gomathi. 2024. Fraud Detection in Financial Transactions: A Machine Learning Approach vs. Rule-Based Systems. In *2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE).* 1–5. doi:10.1109/IITCEE59897.2024.10467759

[16] Jean Lee, Nicholas Stevens, and Soyeon Caren Han. 2025. Large Language Models in Finance (FinLLMs). *Neural Computing and Applications* (2025). doi:10.1007/s00521-024-10495-6

[17] 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

[18] 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

[19] 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

[20] Weimin Lyu, Songzhu Zheng, Lu Pang, Haibin Ling, and Chao Chen. 2023. Attention-Enhancing Backdoor Attacks Against BERT-based Models. arXiv:2310.14480

[21] Yushan Pan. 2020. Cyber Trust in the Norwegian Online Flea Market: An Ethnographic Study on Fraud. In *HCI International 2020 - Posters.* Springer International Publishing, 589–596. doi:10.1007/978-3-030-50732-9_76

[22] 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

[23] Pradheepan Raghavan and Neamat El Gayar. 2019. Fraud Detection using Machine Learning and Deep Learning. In *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE).* 334–339. doi:10.1109/ICCIKE47802.2019.9004231

[24] 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

[25] Shiguang Wu, Yaqing Wang, and Quanming Yao. 2025. Why In-Context Learning Models are Good Few-Shot Learners?. In *The Thirteenth International Conference on Learning Representations.* https://openreview.net/forum?id=iLUcsecZJp

[26] 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

[27] 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

[28] Peng Zhao and Shuyuan Jin. 2024. Fewshing: A Few-Shot Learning Approach to Phishing Email Detection. In *2024 IEEE 4th International Conference on Software Engineering and Artificial Intelligence (SEAI).* 371–375. doi:10.1109/SEAI62072.2024.10674290