

FAA Framework: A Large Language Model-Based Approach for Credit Card Fraud Investigations

Shaun Shuster, Eyal Zaloof, Asaf Shabtai, Rami Puzis
Ben-Gurion University of the Negev, Israel
{shaunshu, zloofe}@post.bgu.ac.il, {shabtaia, puzis}@bgu.ac.il

Abstract—The continuous growth of the e-commerce industry attracts fraudsters who exploit stolen credit card details. Companies often investigate suspicious transactions in order to retain customer trust and address gaps in their fraud detection systems. However, analysts are overwhelmed with an enormous number of alerts from credit card transaction monitoring systems. Each alert investigation requires from the fraud analysts careful attention, specialized knowledge, and precise documentation of the outcomes, leading to alert fatigue. To address this, we propose a fraud analyst assistant (FAA) framework, which employs multi-modal large language models (LLMs) to automate credit card fraud investigations and generate explanatory reports. The FAA framework leverages the reasoning, code execution, and vision capabilities of LLMs to conduct planning, evidence collection, and analysis in each investigation step. A comprehensive empirical evaluation of 500 credit card fraud investigations demonstrates that the FAA framework produces reliable and efficient investigations comprising seven steps on average. Thus we found that the FAA framework can automate large parts of the workload and help reduce the challenges faced by fraud analysts.

Index Terms—Credit Card Fraud, Automated Fraud Investigations, Evaluation of Investigations, LLM

1. Introduction

Credit card fraud refers to the unauthorized use of funds through credit or debit cards [1]. Various techniques are used to steal the information needed to perform illegitimate online card-not-present (CNP) transactions, as well as physical transactions. These methods include skimming, phishing, malware, and man-in-the-middle attacks [2]–[6]. Given that credit card fraud losses exceeded 32 billion US dollars globally in 2023 alone [7], there is an urgent need to develop and implement effective strategies for fraud detection and investigation.

A typical fraud detection and manual investigation framework is presented in Figure 1. In most cases, banks and card networks evaluate transactions using rule-based systems and/or predictive models to identify fraud patterns and assign risk scores to the transactions [8]. If the risk score for a transaction exceeds a certain threshold, an alert is flagged

for manual investigation, indicating that the transaction requires investigation by a fraud analyst. These investigations improve analysts’ understanding regarding the nature of the transaction, while maintaining costumers’ trust and ensuring compliance with regulations. Credit card fraud investigations mainly involve gathering and examining evidence, and at the end of an investigation, a detailed report describing the investigation of the case is produced [9]–[11]. Such reports are used by a company’s senior management, law enforcement agencies, and legal teams to make informed decisions and take appropriate actions [12]. The results of an investigation are usually fed back to retrain fraud detection systems [13].

Investigation of potential fraud cases requires careful attention, specialized knowledge, and precise documentation of the outcomes. Since the number of investigated cases and their high demands can quickly exhaust workforce resources, there is a need for automation. Today, artificial intelligence and large language models (LLMs) are used in a variety of tasks that require advanced analysis capabilities, such as root cause analysis (RCA) [14], [15], cybersecurity [16], and smart policing [17].

In this article, we investigate the extent to which LLM analysis capabilities help automate diverse tasks traditionally performed by human analysts during credit card fraud alert investigations. We present a fraud analyst assistant (FAA) framework that employs LLMs to automate fraud alert investigations. The FAA framework includes tools and LLM agents used in a series of investigative steps for generating code, retrieving data from a source, analyzing and visualizing the data, and deducing evidence from the analysis and visualizations. The investigation process continues, using these tools and agents, until sufficient evidence has been collected, allowing the generation of statements that support or refute the fraudulent nature of the transaction in question. Once enough evidence is collected, the framework automatically generates a report summarizing the key findings and leads uncovered during the investigation. The generated report aims to reduce investigator workload, minimize alert fatigue, and mitigate human error by providing a concise summary of the major insights needed to resolve the case. By reducing human error, we can also increase customer trust in the accuracy and reliability of our investigations.

In our evaluation of the performance of the FAA frame-

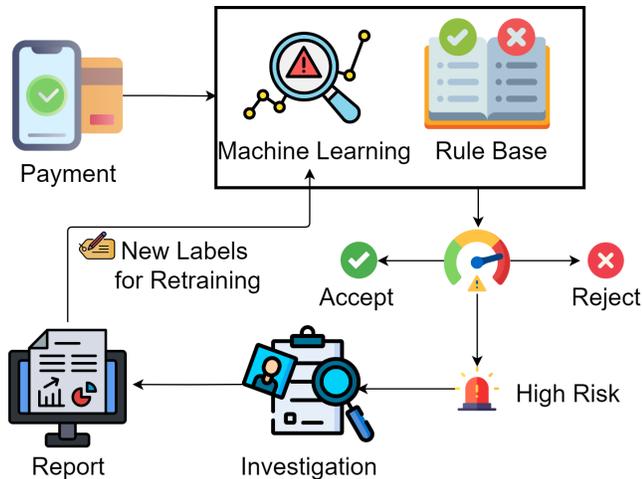


Figure 1. A typical fraud detection framework workflow. The process begins with a payment transaction that is evaluated based on predefined rules. If the transaction raises suspicion, it is further examined by a predictive model, which assesses the fraud risk level. High-risk transactions are flagged for manual investigation. A report is generated after the investigation has been performed, and the final outcome is used to label the transaction. The labeled data can be used to refine both the rule-based system and the predictive model over time [20].

work on the Sparkov and CCTD datasets [18], [19]. We evaluated the efficiency and explainability of the investigations generated using the FAA framework. Efficiency is measured as the number of investigative steps performed by the FAA framework, the number of tokens used during the investigation, and the minimal trajectory ratio. Most investigations include only 20-40% excess steps compared to a *minimal trajectory* – only the necessary steps selected retroactively. We also assessed the quality of the evidence collected by the FAA, employing a new evidence quality measure that includes four different aspects. This measure showed that 71-72% of the evidence collected during investigations performed by the FAA had a high or very high impact on fraud suspicion. Moreover, we found that all the evidence collected was relevant. In addition, there were no logical inconsistencies. Finally, 65-76% of the collected evidence provided new insights. Furthermore, we assessed the quality of the investigation by evaluating the performance of an LLM-based agent classifier tasked with predicting whether the investigation indicated a legitimate or fraudulent transaction. The classifier achieved an F1 score of 98-99%, with a precision of 97.62-98.8% and a recall of 98.4-99.2%.

The contributions of this study are as follows:

- A fully automated framework for credit card fraud investigations that performs a complete investigation and generates a detailed report.
- An approach that transforms an investigation step into well-defined tasks for an LLM.
- A new innovative and practical investigation evaluation measure.

2. Related Work

2.1. Credit Card Fraud Detection

Over the last decades, academia and industry have explored various AI-based techniques for reducing credit card fraud. These techniques encompass rule-based systems [21], [22] and a range of machine learning (ML) approaches. Recent trends have made predictive ML the de facto method for detecting fraud, transforming how businesses safeguard their assets.

The survey article [23] and the papers [24], [25] present various techniques that use ML or deep learning to address the challenge of credit card fraud detection. The academic community has proposed numerous solutions, including unique methods such as Symbolist, Bayesian, evolutionary, analogy-based, and connectionist approaches, as well as more traditional supervised and unsupervised ML models. The article also provides a comprehensive overview of deep learning techniques for detection credit card fraud.

Despite these advances, predictive ML models exhibit notable shortcomings. High performance metrics on benchmark datasets do not always translate into real-world robustness, and “black-box” models often lack the explainability required for forensic investigations. In regulated financial environments, decisions must be backed by clear evidence, which opaque models struggle to provide. Consequently, supplementary investigative frameworks are necessary to extract actionable insights and support evidence-based decision-making.

In [26] the authors introduced the gated temporal attention network (GTAN), a semi-supervised predictive ML graph neural network model designed to tackle the challenge of limited labeled data in credit card transaction datasets by leveraging the temporal and interactional characteristics of transaction records. At the heart of this approach is the construction of a temporal transaction graph, where transactions are represented as nodes and their interrelations over time are represented as edges. This methodology addresses key shortcomings in existing fraud detection approaches, which often neglect unlabeled data, ignore categorical attributes, and require extensive feature engineering for categorical features. By utilizing both labeled and unlabeled data, the authors demonstrated how a semi-supervised graph neural network can uncover complex relationships and temporal patterns within transactions.

While the paper shows an interesting approach, it is limited to the types of connections defined by the authors which could be thought of as feature engineering, therefore might not be effective in case of domain shift. In contrast, the FAA does not require feature engineering and extracts the relevant connection to each case during the investigation.

Recent studies have further advanced the field by integrating ensemble learning and deep models, yet these too face limitations in generalization and transparency. The articles [27], [28] proposed hybrid frameworks that combine ensemble classifiers with deep neural networks to improve

detection accuracy. They emphasize that modern fraud detection methods “include classical, machine learning, deep learning, and hybrid or ensemble methods”. Although their approach outperforms simpler models, they lack explainability which is mandatory when dealing with financial fraud.

The main shortcoming in current research on fraud detection primarily focuses on predictive ML models. While predictive ML models often achieve high performance rates, they cannot be trusted blindly. Moreover, the bureaucratic nature of the financial field requires forensic proof before reaching critical decisions. Consequently, investigations are necessary to extract additional insights, gather the required evidence, and support financial decision making.

2.2. Explainability Methods in Fraud Detection

In numerous applications, explaining the derived conclusions is essential for building trust and transparency. This is particularly important in fields like healthcare, where medical professionals need to have utmost confidence in their diagnoses. As an example, doctors must understand the rationale behind an AI system’s diagnosis of a disease based on a CT scan. Given that AI systems are not infallible, gaining insights into the decision-making process can not only enhance trust but also prevent potentially fatal errors [29].

In the domain of fraud detection, explainability is important for several reasons. First and foremost, it streamlines the investigation process, enabling quicker and more accurate identification of fraudulent activity. This efficiency is important for mitigating financial losses as well as minimizing customer dissatisfaction that can arise from false positives or delayed transaction processing. Furthermore, explainability ensures that the underlying model is working correctly, providing transparency in its decision-making processes. By understanding how and why certain transactions are flagged as suspicious, stakeholders can fine-tune the model to improve its accuracy and reliability. This transparency contributes maintaining trust in the system, both from an internal perspective and in the eyes of the customers and regulatory bodies. Moreover, in an environment where fraud tactics are constantly evolving, explainability allows for the adaptation and evolution of detection strategies, ensuring that the system remains effective against zero-day fraud.

The xFraud framework [30] is designed for detecting and explaining fraudulent transactions in online retail. It features two main components: a detector and an explainer. The detector, which utilizes a graph neural network influenced by heterogeneous graph transformer (HGT) architecture [31], efficiently predicts transactions’ legitimacy by processing diverse transaction data, such as payment and shipping details. The explainer provides understandable explanations for predictions made by the detector. It blends GNNExplainer [32] with task-agnostic centrality measures, enabling decision-makers to comprehend the model’s reasoning by generating a sub-graph that allows the examiner to understand the connections.

While the paper offers interesting approach to explaining the predictions of detection models, there remains a need to

present evidence in a formal and detailed investigation report ensuring compliance with regulations in natural language. The FAA framework addresses this need by utilizing large language models to provide a step-by-step walkthrough of the investigation process, offering a level of detail that surpasses that of traditional detection models.

2.3. Investigation Challenges

As discussed in 2.1, the predictive ML models are insufficient on their own; therefore, investigations are necessary. To better understand the challenges faced by credit card fraud investigators while investigating alerts, we explored variety of sources, including articles, books, and websites.

According to a number of sources [33]–[36], the main challenges faced by credit card fraud investigators when investigating alerts are:

- 1) **Volume of Alerts:** Credit card fraud analysts must cope with a large number of alerts raised by rule-based systems and/or predictive model. Manually resolving each alert results in alert fatigue, thus the analysts may become desensitized and miss critical alerts.
- 2) **Complexity of Investigation:** When properly performed, the credit card fraud detection process is complex, requiring attention, specialized expertise, and precision. This process, which is prone to errors, involves tasks such as retrieval of relevant data, anomaly detection, pattern recognition, as well as various reasoning tasks. Thus it is crucial to perform this process correctly; otherwise, it may lead to credit loss and customer dissatisfaction.
- 3) **Documenting Evidence:** When dealing with credit card transactions and sensitive data, it is crucial to meticulously document the evidence that supports the findings of an investigation. The process must ensure that all relevant details are captured to trace back the reasoning behind declaring a transaction fraudulent or legitimate [36]. However, looking at the evidence can be deceiving, as the investigation process is long and involves numerous pieces of evidence gathered at different investigation steps. This can lead to misinterpreting evidence, particularly for inexperienced investigators, which may lead to incorrect decisions that may harm the investigation’s credibility and cause further confusion among colleagues.

In this paper, we address the challenges faced by credit card fraud detection investigators, by introducing an automated credit card fraud investigation framework.

2.4. Automated Investigation Using LLMs

In recent years, with the rise of LLMs, a few studies have proposed various techniques for automating investigation in various domains.

For example, the use of LLMs for law enforcement via classification and retrieval tasks has been explored. In [17], the authors employed LLMs with prompt engineering, fine-tuning, zero-shot, and few-shot learning to perform tasks

related to smart policing. They used historical data for human-defined tasks, such as crime classification, criminal classification, and other related tasks, combining them into a cohesive narrative about the case being investigated. However, this paper does not provide the reasoning steps to allow for varying investigation lengths. Furthermore, as it limited and cannot create a dynamic investigations making it vulnerable to concept drift as their predefined tasks may not cover all of the important aspect.

In cloud computing, root cause analysis (RCA) is part of the incident management process, focusing on identifying and addressing the underlying causes of system issues to prevent future incidents. It involves analyzing logs, data, and system behavior to trace the origin of problems. It is often challenging for on-call engineers as it demands in-depth domain knowledge and significant experience with a team's specific services. Automating RCA can lead to substantial time savings and reduce the strain on engineers during incident management. The state-of-the-art solutions in RCA use LLMs agents as their core idea [14], [15]. However, these solutions do not make use of LLMs code execution capabilities, unlike the FAA framework which dynamically analyzes each case 3.2.1.

Digital Forensic [37] introduced a general framework for integrated digital investigations that operates based on inputs of natural language. The proposed framework has the potential to be automated, although it uses human-in-the-loop for the investigation. In addition, they did not evaluate their framework in a forensic use case. Although their framework introduces code agents, in some use cases including fraud investigation, the analysis of charts and plots generated by the code agent is crucial for identifying patterns and anomalies.

In the field of cybercrime, an automated method was proposed in [38] to extract criminal information from digital evidence networks. This approach, which utilizes LLMs to learn patterns and relationships within forensic artifacts, enables the automatic construction of forensic intelligence graphs (FIGs).

In another paper [16], the authors addressed the following questions: "Can LLMs be utilized to understand cybersecurity logs, events, and threat feeds and explain them adequately to a human operator?" (Explainable AI) and "Can LLMs take security actions based on instructions from a human operator?" (Actionable AI). To answer those questions the presented a chatbot approach that can query information from a database and retrieve dated insights about the question back to the user, which significantly increases explainability. The conversational agent can also execute actions, which include executing scripts, blocking IP addresses, or other mitigation measures to address security breaches based on the threat level. Similarly, [39] used a similar approach but aimed at non-expert users, by explaining intrusion detection system (IDS) messages, for example, for network-based systems. A network-based IDS component inspects the network packets that pass a router for suspicious traffic and generates alerts. The chatbot accepts alerts from the IDS component, sends them to the

LLM component for translation into an intuitive explanation, and presents a user interface with the explanations to the user. If the user requires further support, they can use the interface to send follow-up questions to the LLM.

While both papers demonstrate significant advancements in using chatbots and LLMs for cybersecurity, they primarily focus on initial alert handling and basic user interactions. The FAA framework extends these concepts by not only taking basic actions upon request and explaining messages from the monitoring system but also creating a full detailed investigation from the alert raised by the monitoring system.

3. Method

In this section, we outline the investigation process tailored for automated credit card fraud investigations and describe our proposed framework for automating credit card fraud investigations which employs this process.

3.1. Investigation Process

The proposed investigative process consists of a sequence of steps, each of which provides evidence that collectively forms a coherent narrative about the case. Each step is comprised of three phases: planning, information-gathering, and analysis. At the end of the investigation, the FAA framework generates a report that includes insights from each major step, particularly insights that shed light on whether the transaction is fraudulent or not.

3.1.1. Investigation Step Phases. Each investigation step is implemented using the chain-of-thought (CoT) prompt engineering technique, which instructs the LLM to perform a series of reasoning steps [40].

The FAA framework starts by planning the strategy for the investigation step with the goal of solving the fraud case, based on the available evidence. It then executes this strategy, which may involve generating charts, querying databases, or writing Python code. Then, the results are analyzed in relation to the initial objectives and prior findings.

- 1) **Planning Phase:** The purpose of this phase is to review the evidence gathered in the investigation so far and leverage LLM's domain knowledge to determine an strategy that will advance the investigation by collecting additional evidence to decide whether the transaction is fraudulent.
- 2) **Information-Gathering Phase:** During this phase, the FAA framework generates code to execute the approach chosen in the planning phase. This may include creating plots, retrieving relevant data, and performing other necessary operations required in an investigation.
- 3) **Analysis Phase:** This phase involves analyzing and interpreting the evidence gathered in the information-gathering phase. This evidence is then integrated with information from previous steps to derive meaningful insights about the case. At this point, the LLM may choose to stop performing investigation steps if it determines that enough evidence has been collected to make a decision.

3.1.2. Investigation Report. The report is an integral part of a proper investigation as there is a need to review the investigation steps taken and the evidence gathered in the investigation. The report is produced by the FAA framework after the last investigation step, it should be an informative summary of the important evidence that was collected during the investigation. The report is used by senior management, law enforcement agencies, and legal teams to make informed decisions and take appropriate action [12]. Note that the report the FAA framework provides the analyst, does not include client interviews and formal processing of each case [41] which should still be performed by the personnel in charge.

3.2. FAA Framework

The workflow of the FAA framework for conducting the investigation process, as discussed in Section 3.1, is illustrated in Figure 2. At the start of a new investigation, an initial prompt containing the transaction number (see Listing 1 in Appendix A) is sent directly to the FAA. Initially, the FAA typically verifies that the transaction is in the credit card transaction database. If found, the FAA displays the transaction details. This process corresponds to steps 0- 2 in Figure 2.

In step 3, the current investigation, along with the investigation step phases description (see Section 3.1.1) and general investigation guidelines (see Listing 3 and Listing 4 in Appendix A). This Investigation Step Prompt is responsible for guiding the FAA on how to perform an investigation step.

In the figure, starting with step 4, the arrows are in red to indicate a loop that generates investigation steps until sufficient evidence has been gathered to resolve the case.

In step 4, the FAA processes the prompt, and in step 5, the FAA displays the newly generated investigation step to the user for transparency of the process. If the FAA chose to end the investigation, in addition to the last investigation step, a report, and a decision by the FAA of whether the transaction is fraudulent is displayed. Otherwise, the framework proceeds to step 6 where the framework updates the current investigation with the newly generated investigated step. In step 7, the framework incorporates the updated investigation, along with the investigation step phases description, and the investigation guidelines.

3.2.1. Fraud Analyst Assistant. To create the FAA, we chose to use OpenAI's Assistants API [42], leveraging the GPT-4o model [43]. This assistant carries out investigation steps by utilizing the appropriate tools and agents in each part of a step, as illustrated in Figure 3. The diagram illustrates seven key steps in the communication and workflow among different components: the FAA, code execution tool, vision agent, and report-generating agent. Starting with the FAA, the flow connects to a code execution tool interacting with both a Python code interpreter and a transaction database. Further steps involve information exchange with a vision agent, which completes the cycle by communicating

back to the FAA and the report-generating agent, ultimately leading to the generation of reports.

The tasks of the FAA during an investigation step are:

- 1) Planning the investigation step (Section 1). At the start of each step, the FAA performs the planning phase by determining which investigative lead to pursue and outlining the strategy to take.
- 2) Activating the code execution tool (Section 3.2.2). This tool is used during the information-gathering phase for querying the database, computing statistics about relevant data, and generating charts and visualizations (Section 2). The output from the executed code will be returned to the FAA.
- 3) Activating the vision agent using the function calling tool (Section 3.2.3). The vision agent is activated during the information-gathering phase to extract insights from the visualizations which are provided by the FAA (Section 3.1).
- 4) Performing the Analysis Phase, that consists of analyzing the insights gathered (Section 3). Based on the output from the information-gathering phase, the FAA analyzes the insights, along with previously collected evidence, to derive new evidence.
- 5) Deciding the investigation outcome (Section 3.1). During the analysis phase, the FAA decides whether the investigation can be concluded or if an additional investigation step is required. If the investigation can be concluded, the FAA calls the report-generating agent (RGA) to generate the investigation reports (Section 3.2.3) and with the report the FAA calls the detective agent to decide whether the transaction is fraudulent. If not, the investigation process continues as shown in Figure 2.

3.2.2. Tools. Code Execution Tool. Code Interpreter allows the FAA to write and run Python code in a sandboxed execution environment. This tool can process files such as the transaction database and output the execution's result. Code Interpreter allows the FAA to run code iteratively to resolve the investigation. When the FAA writes code that fails to run, it can iterate on this code by attempting to run different code until the code execution succeeds.

Function Calling Tool. The function calling tool enables the definition of functions that the FAA can utilize. This definition comprises the function's name, a description of its purpose, and the input parameters along with their types, names, and descriptions. When the FAA invokes the tool, an event is triggered in the code that includes the function's name and its argument values.

3.2.3. Agents. Vision Agent. The vision agent is built on OpenAI's GPT-4o. GPT-4o is a multi-modal model that not only excels in text generation but also possesses image-to-text capabilities. By leveraging the descriptions provided by the FAA along with the generated image, we aim to identify trends, clusters, or anomalies within an image (see Listing 5 in Appendix A). We employed various prompt engineering

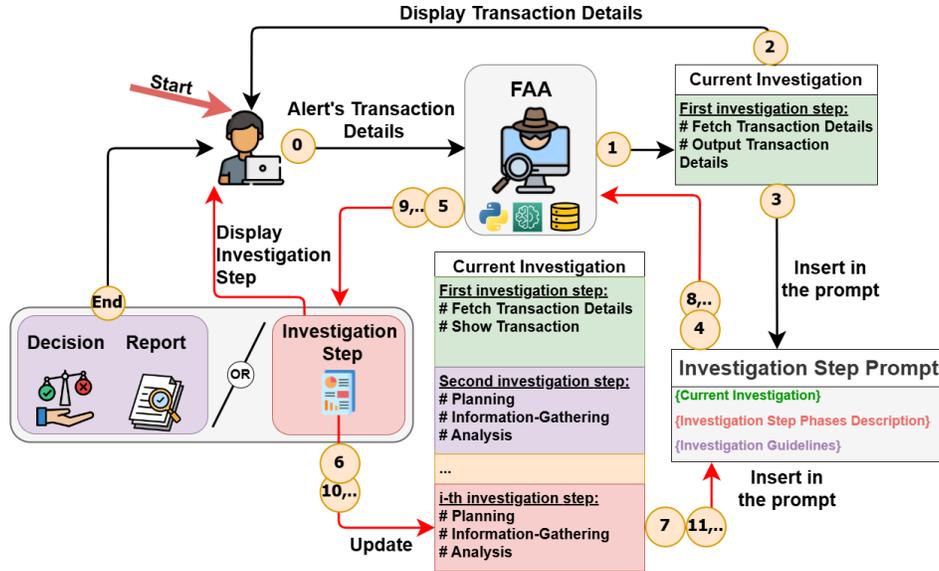


Figure 2. The FAA framework for automating fraud investigation. This includes performing a sequence of investigation steps until the FAA determines that enough evidence has been gathered, and then a report is generated. Given the report, which consists of the evidence gathered, the detective agent decides whether the investigated case is fraudulent.

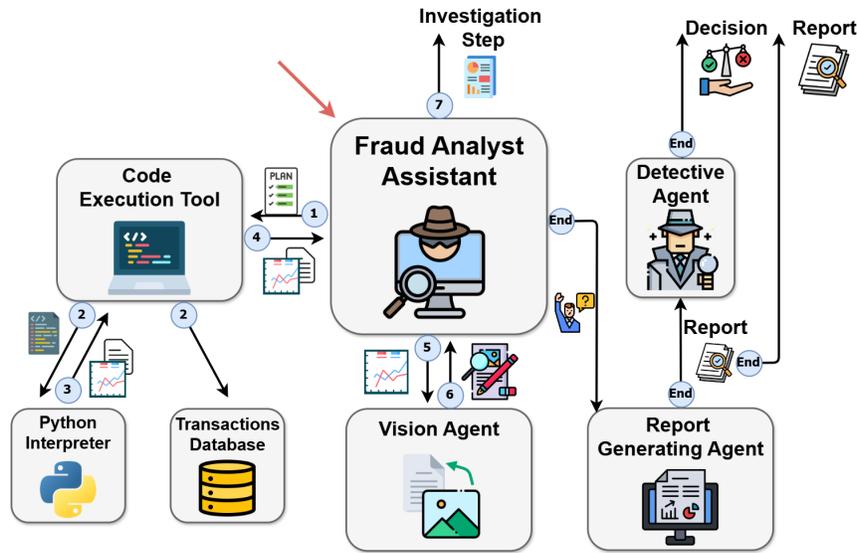


Figure 3. The communication and workflow between the FAA framework components: the FAA, tools, and agents.

techniques, including role-based prompting and zero-shot CoT prompting to the vision agent [40].

Report-Generating Agent. After the FAA performs the necessary investigation steps, the RGA generates a detailed report that summarizes the key evidence gathered during the investigation.

The generation process includes two steps:

- 1) First, the RGA generates an unfiltered report that includes a description of the steps taken during the investigation and the evidence that was collected in each step (the full prompt is available in Listing 6 in Appendix

A). In the prompt, we used the few-shot prompting technique.

- 2) Then the RGA takes the unfiltered report it generated, and extracts the evidence it found to be the most important for determining the legitimacy of the suspicious transaction. This results in the report, which is presented to the user.

Detective Agent. After the report is generated at the end of the investigation, the FAA informs the user whether, based on the main evidence gathered, the transaction is fraudulent or not. For this purpose, the FAA framework

utilizes the detective agent, which is responsible for determining whether the transaction is fraudulent based on the evidence from the report produced by the RGA (Section 3.2.3).

4. Experiments

To assess FAA’s performance in fraud investigations, we pose the following research questions:

RQ1 Is the FAA framework capable of generating a report with high-quality evidence, without compromising the efficiency of the investigation process?

The report, as detailed in Section 3.1.2, summarizes the investigation steps taken and the evidence gathered during the investigation. We define a high quality evidence as an evidence that: *i*) has high impact on fraud suspicion level, *ii*) is relevant to the investigation case, *iii*) provides new knowledge, and *iv*) logically aligns with the rest of the investigation evidence. To answer this research question, we evaluate the quality of the evidence presented in the report using the evaluation measure described in Section 4.3.3. The efficiency of the evidence extraction performed in the investigation process is measured by the number of investigation steps, the total number of tokens, and the minimal trajectory (Section 4.3.2).

RQ2 Can the report produced by the FAA framework be used for reliable fraud detection?

Our objective is to determine whether the evidence presented in the reports generated by the FAA framework can lead to accurately detecting fraud.

RQ3 What is the distribution of the investigation step categories comprising the investigations performed by the FAA?

Since each case has unique characteristics based on the behavior of the cardholder, the merchant, and a variety of other variables, our objective is to examine the categories of investigation steps that the FAA framework tends to execute.

RQ4 What is the impact of the vision agent the FAA framework?

We assess how including the vision agent in the FAA affects the results of the measures used to assess the FAA framework (RQ1) and (RQ2).

4.1. Datasets

For our experiments, we used the Sparkov and CCTD datasets [18], [19]. The sparkov dataset is a synthetic labeled dataset consisting of 1,852,394 card-not-present (CNP) transactions. The dataset labels indicate whether a transaction is fraudulent or legitimate. The dataset includes 1,842,743 legitimate transactions and 9,651 fraudulent transactions, providing a diverse sample for fraud detection.

The dataset contains various fields, including the date and time of each transaction, along with the cardholder’s credit card number. It includes important merchant details, personal and geographic information about the cardholder, and transaction-specific fields such as the amount and location of the transaction and the transaction number.

We preprocessed the dataset to address an issue where most merchant names contained the word ‘fraud,’ causing the model to inaccurately classify all cases as fraudulent, even when no fraudulent activity was present. To resolve this, we removed the word ‘fraud’ from all merchant names.

The CCTD dataset is also a synthetic labeled datasets which consists of 24,357,143 legitimate credit card transactions and 29,757 fraudulent ones. It includes details on whether the transaction was successful, detail about the credit cards used, cardholder details, merchant details, transaction location, transaction price, time, date, and whether it was fraudulent or not. We apply minimal preparation for the data, by turning the columns representing numerical or binary values into integers. We use labels to assess the reliability of decisions made based on the reports generated by the FAA framework (RQ2).

4.2. Dataset Memorization Check

Inspired by [44], which demonstrates that LLMs can memorize tabular data, we apply a similar method to demonstrate that GPT-4o did not memorize our datasets. Therefore we conducted two experiments, random feature completion and prediction for the is-fraud column. For that we randomly sampled 25 fraudulent transactions and 25 legitimate transactions.

4.2.1. Random Feature Completion. In this experiment we hid a set of three columns in each transaction, one column prediction at a time, resulting in $50 * 3 = 150$ tests. For features that could not be inferred by domain knowledge, the accuracy was 0. No feature was correctly identified. The following features were not included in this test, as they could be inferred using domain knowledge:

- city - Using the zip or the state.
- state - Using the city or the zip.
- zip - Using the city.
- gender - Using the name.
- is-fraud - To check for dataset memorization, we did not modify the merchant names. Since many merchant names contained the word ‘fraud,’ the LLM classified all transactions as fraudulent. Therefore, we removed the is-fraud column for this experiment.

4.2.2. Prediction for the is-fraud Column. In this experiment, we removed the fraud string from all merchant names, as was done in our automated investigations. We then prompted the LLM to classify the transaction. The confusion matrix (Table 1) presents the predictions of the is-fraud column. The accuracy from the confusion matrix is 0.58. Since we are working with binary labels, we can conclude that the LLM did not memorize the labels and is outputting predictions randomly.

4.3. Key Performance Indicators

To assess the performance of the FAA framework, we used standard measures when they were available and sug-

Actual	Predicted Not Fraud	Predicted Fraud	Total
Not Fraud	23	2	25
Fraud	19	6	25
Total	42	8	50

TABLE 1. CONFUSION MATRIX OF THE MEMORIZATION CHECK.

gested custom measures to assess the quality of evidence and the effectiveness of the investigation.

4.3.1. Fraud Detection Metrics. Based on the results generated by the detective agent, we would like to measure how well the FAA framework can detect whether a transaction is fraudulent (Section 3.2.3). By measuring our fraud detection performance, we can evaluate how well our detective agent performs and how clear and decisive the pieces of evidence are in the report. We use standard notions of the confusion matrix: true positive (TP), true negative (TN), false positive (FP), and false negative (FN), dissecting the actual transaction labels vs. the detective agent’s decisions. To assess fraud detection performance, we employed the metrics most commonly used in this domain: precision, recall, and F1 [23].

4.3.2. Investigation Efficiency Measures.

Number of Investigation Steps: This measure simply counts the number of investigation steps. We aim for a low number of steps.

Number of Tokens: Recent studies have tried to reduce costs by simplifying queries, which can decrease the number of input and output tokens [45], or by dynamically selecting between more affordable and expensive models [46]–[48]. Minimizing the token count is helpful for reducing costs and improving response times in both open-source and closed-source models.

To count the number of GPT-4o text tokens, we used the "o200k_base" tokenizer available in tiktoken [49].

Investigation Minimal Trajectory Ratio: In the field of reinforcement learning, the term trajectory refers to the sequence of states and actions taken by an agent during an episode. In investigations, some steps in this trajectory may lead to dead ends due to incorrect hypotheses. After completing an investigation, we can calculate the percentage of steps that provided useful insights supporting the investigation’s decision, resulting in the minimal trajectory ratio of an investigation. We denote the set of supporting decision steps in an investigation I as $SDS(I)$ and the set of all investigation steps in investigation I as $STEPS(I)$. Note that "minimal" in this context refers to the subset of the trajectory that is useful. In contrast, "minimum" would refer to the optimum trajectory, which represents the most efficient path the investigation could have taken, rather than analyzing the trajectory that was followed.

$$\text{Min-trajectory-ratio}(I) = \frac{|SDS(I)|}{|STEPS(I)|}$$

where I is an investigation.

4.3.3. Evidence Quality Score. In the FAA framework, the report plays a major role, as it summarizes the investigation steps taken and the evidence gathered during the investigation (see Section 3.1.2). Therefore, we would like to use the generated report for the evaluation of the framework. However, since we want to evaluate the entire investigation process, there is a need to review all of the evidence gathered – not just the evidence selected in the evidence filtering phase as described in Section 3.2.3. Therefore, in our evaluation, we use the unfiltered report. In addition, we implemented this measure using a few-shot prompting technique to prompt an LLM to provide a rating on a five-point Likert scale for each piece of evidence on 4 different aspects as follows:

- 1) **Impact on fraud suspicion level:** This evidence affects the level of fraud suspicion, either by increasing or decreasing it.
- 2) **Relevant to investigation case:** This evidence was relevant to the investigation case.
- 3) **Providing new knowledge:** This evidence either provided me with new insights or confirmed previously unverified information about the case.
- 4) **Logical alignment:** The evidence logically aligns with the rest of the report.

4.4. Quality of Generated Evidence and Efficiency Assessment (RQ1)

To address the first research question, we assessed the FAA framework’s quality and efficiency in producing automated investigations, by employing the evidence quality score 4.3.3, as well as investigation efficiency measures 4.3.2.

Investigation Efficiency Assessment. For the efficiency assessment, we analyzed the minimal trajectory ratio, input tokens, output tokens, and number of investigation steps. Presented in table 2 we found that for the minimal trajectory measure, most investigations were in the range of 72-83% with the majority requiring 7 investigation steps on average. Although the investigations contained mostly necessary steps, unnecessary steps were revealed that did not contribute significant evidence toward deciding the case. Furthermore, the total number of tokens used was substantial, averaging above 205k tokens in a single investigation when utilizing the vision agent.

Quality of Generated Evidence. Figure 4 presents the results of our evaluation of the quality of the evidence generated by the FFA framework on a 5-point Likert scale across four aspects: impact on fraud suspicion level, relevant to investigation case, providing new knowledge, and logical alignment. The results indicate varying score across these aspects. In the investigations we conducted, all of the pieces of evidence were found relevant to the investigation case and logically aligned with the rest of the investigation, as the evaluation rated them as "agree" or "strongly agree" on those aspects. More than three-quarters of the evidence had a high or very high impact on the aspect of the level

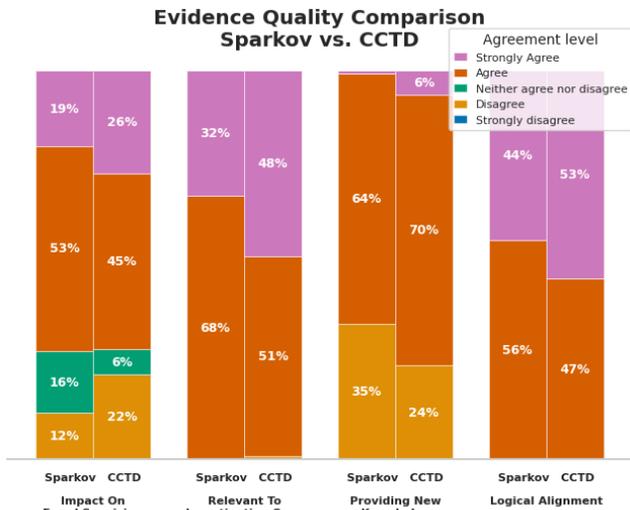


Figure 4. The quality of the evidence found by the FAA framework.

of fraud suspicion. All of the evidence pieces were rated as relevant, with no logical inconsistencies, and majority of the evidence provided new insights. Note that the distribution in the categories; except for impact on suspicion level, does not include neither agree nor disagree. This can be explained by the fact that those categories are binary. For example, in logical alignment, something can either align or not, and there is not an option for "maybe." This indicates that the evidence provided in these areas was consistently perceived as relevant and logically consistent. However, while the 'impact on fraud suspicious level' and the 'providing new knowledge' aspects were 71-72.14% and 65-76%, respectively, these results are raising some concern as they show that some part of the evidence evidence had questionable contribution to the investigations.

These findings highlight the strengths of our FAA framework in generating structured and logical investigative steps while maintaining a high efficiency, which can be seen by, the high minimal trajectory, and a reasonable number of steps. Moreover, those results also point out things we can improve, particularly in ensuring that the investigations performed consistently provide new knowledge and impact the fraud suspicion level significantly.

4.5. Reliability of Detection Based on FAA Report (RQ2)

To address the second research question, we utilized the report containing all the key evidence generated by the RGA 3.2.3. An LLM was tasked with classifying transactions solely based on the evidence represented in the report 3.2.3. Table 3 shows the results across standard metrics: F1, precision, and recall scores. Due to budget constraints that limited our usage of the OpenAI's Assistant API, our FAA framework was evaluated on a total of 500 transactions, consisting of 250 fraudulent and 250 legitimate transactions.

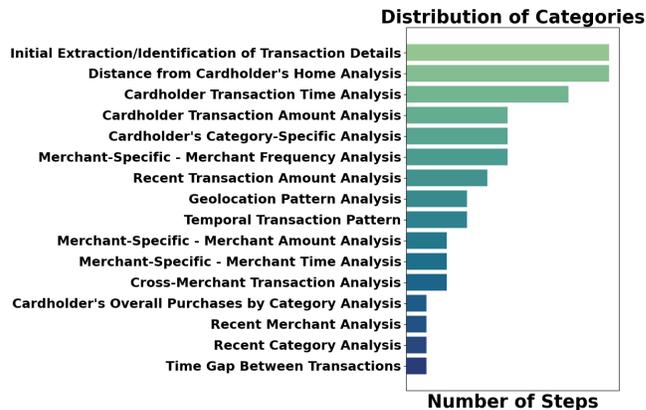


Figure 5. Categories of the investigation steps performed by the FAA framework.

For the Sparkov dataset, the FAA framework (with and without the vision agent) outperforms all baseline methods in terms of F1 score, precision, and recall, establishing a new state-of-the-art for this dataset. Notably, the FAA framework achieves an F1 score of 0.9801, significantly higher than previous ensemble and graph-based approaches.

In contrast, on the CCTD dataset, our framework still performs competitively—with an F1 score of 0.99—but does not surpass the best-performing method. This slight difference can be attributed to a fundamental design distinction: unlike prior models, our FAA framework does not rely on training over the same data distribution, making it more generalizable but less optimized for static datasets. Nonetheless, the FAA model still achieves exceptionally high recall (0.992) and precision (0.988), underscoring its robustness even in unfamiliar data settings.

In terms of the quality of the generated evidence, overall, we observed that the evidence provided in the reports generated by the FAA framework are reliable for determining whether a transaction is fraudulent. However, it is important to note that our framework is designed for investigation purposes and might be found to be more expensive when compared to other detection models.

4.6. Distribution of Investigation Steps (RQ3)

To address this research question we analyzed the investigations we performed and categorized the investigation steps performed in the investigations?

From Figure 5 we can learn that our FAA framework acknowledges the importance of investigating cardholder behavior and transactions. The core steps focus on transaction detail extraction and geolocation analysis, reflecting their foundational role in early fraud detection. High-frequency use of cardholder and merchant behavior analysis highlights the system's emphasis on behavioral profiling. Less frequent categories, such as recent activity and timing patterns, provide complementary signals for edge-case detection. Moreover, it is capable of performing relatively unique steps that were performed throughout the investigations.

TABLE 2. COMPARISON OF THE FAA FRAMEWORK UNDER VISION-ENABLED AND NON-VISION SETTINGS ACROSS THE SPARKOV AND CCTD DATASETS.

Dataset	Setting	Minimal Trajectory Ratio (%)	Input Tokens (K)	Output Tokens (K)	Number of Investigation Steps
Sparkov	w/o Vision	0.76	110	4.1	5.5
	w/ Vision	0.80	205	5.05	6.0
CCTD	w/o Vision	0.8	130	4.3	7.0
	w/ Vision	0.83	250	5.1	7.3

TABLE 3. PERFORMANCE COMPARISON OF VARIOUS CREDIT CARD FRAUD DETECTION MODELS ACROSS SPARKOV AND CCTD DATASETS.

Dataset	Model	F1 Score	Precision	Recall
Sparkov	Enhancing Credit Card Fraud Detection through Advanced Ensemble Learning Techniques [50]	0.85	0.87	0.84
	Identifying Fraudulent Credit Card Transactions Using Ensemble Learning [51]	0.88	–	–
	Leveraging Graph-Based Learning for Credit Card Fraud Detection [52]	0.61	0.92	0.42
	Proposed framework without vision agent (FFA)	0.972	0.968	0.976
	proposed framework (FFA)	0.9801	0.9762	0.984
CCTD	A machine learning based credit card fraud detection using the GA feature selection [25]	0.998	0.999	0.998
	Credit Card Fraud Detection Using SMOTE and AdaBoost [24]	0.99	0.9828	0.9810
	Proposed framework Without vision agent (FFA)	0.978	0.972	0.984
	Proposed Framework (FFA)	0.99	0.988	0.992

4.7. Impact of Vision Agent on the Investigations (RQ4)

To examine the contribution of the Vision Agent within our framework, we re-evaluated the same set of transactions used in the previous research questions. This time, we excluded the Vision Agent and relied solely on the built-in vision functionality of the Assistant API, allowing us to compare its impact on performance and investigation quality.

Figure 6 compares the quality of the between investigations conducted without the vision agent and those with it, each stacked bar represents the agreement level (strongly agree, agree, neither agree nor disagree, disagree) for different aspects of evidence quality: impact on fraud suspicion level, relevance to investigation case, providing new Knowledge, and logical alignment.

The results indicate that including the vision agent leads to higher levels of agreement in terms of evidence quality, particularly in the aspects of providing new knowledge and logical alignment, demonstrating the Vision agent’s positive impact on our FAA framework, for both Sparkov and CCTD datasets.

Table 2 explores the trade-offs between token usage and trajectory efficiency under the conditions of “w/o vision” and “w/ vision”. The addition of vision capabilities enhances the minimal trajectory ratio, suggesting more efficient paths, but it also significantly increases both the input and output tokens. This trade-off highlights that while the vision agent improves trajectory optimization, as seen in the higher minimal trajectory ratio, it requires more computational

resources in terms of token usage. The impact on the number of steps remains moderate, with minimal variations observed between the two conditions. These findings underscore the benefits and costs associated with integrating the vision agent into the investigative process, emphasizing the need to balance efficiency gains with resource expenditures.

Table 3 shows an improvement in the detection rate performance when incorporating the vision agent. On the Sparkov dataset, the proposed framework with the vision agent achieved an F1 score of 0.9801, compared to 0.972 without vision capabilities. The CCTD dataset showed an even greater improvement, with the F1 score increasing from 0.978 to 0.99. These results highlight the effectiveness of incorporating the vision agent into the FAA framework.

5. Conclusions

In this paper, we introduce an innovative approach to automate credit card fraud investigations using LLMs. Our FAA framework leverages a combination of planning, information-gathering, and analysis phases to dynamically investigate suspicious transactions. By integrating the FAA, Vision, and the RGA, the FAA framework automates critical steps in the investigation, and provides detailed and structured reports that enhance analysts’ decision-making capabilities.

Addressing the Challenges. Our FAA framework tackles several core challenges inherent in credit card fraud investigations:

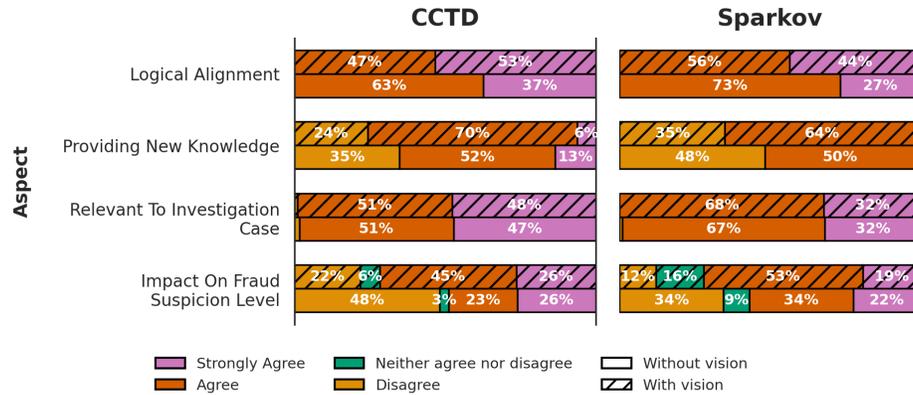


Figure 6. Comparison of the quality of evidence in investigations performed without vision agent and with vision agent across four aspects.

- 1) **Volume of Alerts:** By automating investigation steps, our model reduces the manual workload of fraud analysts, allowing them to focus on more complex cases and reducing alert fatigue associated with handling a large volume of alerts.
- 2) **Complexity of Investigation:** The FAA framework's systematic approach addresses the complexities involved in fraud investigations. Each step is planned, executed, and analyzed, ensuring a thorough examination of multiple aspects of each transaction it investigates.
- 3) **Documenting Evidence:** The FAA framework automates the creation of detailed and well-structured reports. These reports offer a summary of the key evidence gathered during the investigation, such summary can help the analyst avoid incorrect decisions. Making the report especially valuable for senior management, law enforcement agencies, and legal teams which relies on the significant evidence and the decision. We evaluate the report based on impact on fraud suspicious, relevance, logical alignment, and new information of its evidences.

By addressing these challenges, the FAA framework enhances the efficiency, accuracy, and transparency of credit card fraud investigations, Leveraging the power of LLMs, the FAA framework not only reduces the burden on fraud analysts but also improves the overall efficiency of fraud investigation efforts.

For future work, we aim to study how incorporating continuous learning could impact the performance of the FAA framework framework. Currently, the framework does not account for past investigations to enhance its performance over time.

This is intriguing because, although the framework generates reliable and efficient automatic investigations with high-quality evidence, it differs from a human fraud analyst, as human analyst learns about current trends in credit card fraud through experience, but the FAA framework is not aware of emerging fraud patterns since it does not learn from new investigations.

References

- [1] T. P. Bhatla, V. Prabhu, and A. Dua, "Understanding credit card frauds," *Cards business review*, vol. 1, no. 6, pp. 1–15, 2003.
- [2] Y. Irvin-Erickson, "Identity fraud victimization: a critical review of the literature of the past two decades," *Crime Science*, vol. 13, no. 1, p. 3, 2024.
- [3] K. J. Barker, J. D'amato, and P. Sheridan, "Credit card fraud: awareness and prevention," *Journal of financial crime*, vol. 15, no. 4, pp. 398–410, 2008.
- [4] S. Akter, S. Chellappan, T. Chakraborty, T. A. Khan, A. Rahman, and A. A. Al Islam, "Man-in-the-middle attack on contactless payment over nfc communications: design, implementation, experiments and detection," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 6, pp. 3012–3023, 2020.
- [5] K. K. Peretti, "Data breaches: What the underground world of "carding" reveals," *Santa Clara Computer & High Tech. LJ*, vol. 25, p. 375, 2008.
- [6] A. Mishra, B. B. Gupta, and D. Gupta, "Identity theft, malware, and social engineering in dealing with cybercrime," in *Computer and Cyber Security*. Auerbach Publications, 2018, pp. 627–648.
- [7] S. BARBARA, "Payment card fraud losses reach 32.34 billion," 2022, accessed: 2024-07-25. [Online]. Available: <https://www.globenewswire.com/news-release/2022/12/22/2578877/0/en/Payment-Card-Fraud-Losses-Reach-32-34-Billion.html>
- [8] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol. 16, pp. 449–475, 2013.
- [9] B. Nikkel, "Fintech forensics: Criminal investigation and digital evidence in financial technologies," *Forensic Science International: Digital Investigation*, vol. 33, p. 200908, 2020.
- [10] S. Raghavan, "Digital forensic research: current state of the art," *Csi Transactions on ICT*, vol. 1, pp. 91–114, 2013.
- [11] M. Reith, C. Carr, and G. Gunsch, "An examination of digital forensic models," *International Journal of digital evidence*, vol. 1, no. 3, pp. 1–12, 2002.
- [12] F. C. Academy, "The investigation team and their roles: The important role of investigation team," 2024, accessed: 2024-07-27. [Online]. Available: <https://financialcrimeacademy.org/the-investigation-team-and-their-roles/>
- [13] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learning strategy," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3784–3797, 2018.

- [14] D. Roy, X. Zhang, R. Bhavé, C. Bansal, P. Las-Casas, R. Fonseca, and S. Rajmohan, "Exploring llm-based agents for root cause analysis," in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 2024, pp. 208–219.
- [15] Y. Chen, H. Xie, M. Ma, Y. Kang, X. Gao, L. Shi, Y. Cao, X. Gao, H. Fan, M. Wen *et al.*, "Automatic root cause analysis via large language models for cloud incidents," in *Proceedings of the Nineteenth European Conference on Computer Systems*, 2024, pp. 674–688.
- [16] M. Kaheh, D. K. Kholgh, and P. Kostakos, "Cyber sentinel: Exploring conversational agents in streamlining security tasks with gpt-4," *arXiv preprint arXiv:2309.16422*, 2023.
- [17] P. Sarzaeim, Q. H. Mahmoud, and A. Azim, "A framework for llm-assisted smart policing system," *IEEE Access*, 2024.
- [18] N. Brandon, "Sparkov data generation," GitHub repository, 2016. [Online]. Available: https://github.com/namebrandon/Sparkov_Data_Generation
- [19] E. Altman, "Credit card transactions," 2019, apache License 2.0. [Online]. Available: <https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions>
- [20] D. Cheng, X. Wang, Y. Zhang, and L. Zhang, "Graph neural network for fraud detection via spatial-temporal attention," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3800–3813, 2020.
- [21] J. Kumar and V. Saxena, "Rule-based credit card fraud detection using user's keystroke behavior," in *Soft Computing: Theories and Applications: Proceedings of SoCTA 2021*. Springer, 2022, pp. 469–480.
- [22] K. RamaKalyani and D. UmaDevi, "Fraud detection of credit card payment system by genetic algorithm," *International Journal of Scientific & Engineering Research*, vol. 3, no. 7, pp. 1–6, 2012.
- [23] A. Cherif, A. Badhib, H. Ammar, S. Alshehri, M. Kalkatawi, and A. Imine, "Credit card fraud detection in the era of disruptive technologies: A systematic review," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 1, pp. 145–174, 2023.
- [24] E. Ileberi, Y. Sun, and Z. Wang, "Performance evaluation of machine learning methods for credit card fraud detection using smote and adaboost," *IEEE Access*, vol. 9, pp. 165 286–165 294, 2021.
- [25] —, "A machine learning based credit card fraud detection using the ga algorithm for feature selection," *Journal of Big Data*, vol. 9, no. 1, p. 24, 2022. [Online]. Available: <https://doi.org/10.1186/s40537-022-00573-8>
- [26] S. Xiang, M. Zhu, D. Cheng, E. Li, R. Zhao, Y. Ouyang, L. Chen, and Y. Zheng, "Semi-supervised credit card fraud detection via attribute-driven graph representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 14 557–14 565.
- [27] M. Dhandore, C. Agrawal, M. Meena, and I. Journal, "Enhancing credit card fraud detection through advanced ensemble learning techniques and deep learning integration," vol. 5, pp. 2582–6948, 09 2024.
- [28] J. Jemai, A. Zarrad, and A. Daud, "Identifying fraudulent credit card transactions using ensemble learning," *IEEE Access*, vol. 12, pp. 54 893–54 900, 2024.
- [29] P. Gohel, P. Singh, and M. Mohanty, "Explainable ai: current status and future directions," 2021.
- [30] S. X. Rao, S. Zhang, Z. Han, Z. Zhang, W. Min, Z. Chen, Y. Shan, Y. Zhao, and C. Zhang, "xfraud: explainable fraud transaction detection," *Proceedings of the VLDB Endowment*, vol. 15, no. 3, p. 427–436, Nov. 2021. [Online]. Available: <http://dx.doi.org/10.14778/3494124.3494128>
- [31] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," 2020.
- [32] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," 2019.
- [33] A. of Certified Fraud Examiners, *Fraud Examiners Manual*. ACFE, 2024, accessed: 2024-07-17. [Online]. Available: <https://www.acfe.com/training-events-and-products/cfe-exam-preparation/cfe-exam-prep---study-on-your-own-options/fraud-examiners-manual>
- [34] Chargebacks911, "Credit card fraud investigation: How it works & why it's important," 2023, accessed: 12 September 2024. [Online]. Available: <https://chargebacks911.com/credit-card-fraud-investigation/>
- [35] W. V. P. Bart Baesens, Véronique Van Vlasselaer, *FRAUD ANALYTICS*. Wiley, 2015, accessed: 2024-09-12. [Online]. Available: <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119146841?msocid=34b4c9be4ab86f9a36b2dbec4bc56ef9>
- [36] C. IQ, "Investigation report example: How to write an investigative report," <https://www.caseiq.com/resources/ultimate-guide-to-writing-investigation-reports/>, 2024, accessed: 2024-07-28.
- [37] A. Wickramasekara and M. Scanlon, "A framework for integrated digital forensic investigation employing autogen ai agents," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 2024, pp. 01–06.
- [38] H. Zhou, W. Xu, J. Dehlinger, S. Chakraborty, and L. Deng, "An llm-driven approach to gain cybercrime insights with evidence networks," *usnix*, 2024.
- [39] V. Jüttner, M. Grimmer, and E. Buchmann, "Chatids: Explainable cybersecurity using generative ai," 2023.
- [40] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," *arXiv preprint arXiv:2310.14735*, 2023.
- [41] C. IQ, "How to conduct an effective fraud investigation: Your complete guide," accessed: 2024-08-07. [Online]. Available: <https://www.caseiq.com/resources/how-to-conduct-a-fraud-investigation-the-complete-guide/#analyze>
- [42] OpenAI, "Openai assistants overview," 2024, accessed: 2024-08-14. [Online]. Available: <https://platform.openai.com/docs/assistants/overview>
- [43] —, "Hello gpt-4o," accessed: 14-Aug-2024. [Online]. Available: <https://openai.com/index/hello-gpt-4o/>
- [44] S. Bordt, H. Nori, and R. Caruana, "Elephants never forget: Testing language models for memorization of tabular data," *arXiv preprint arXiv:2403.06644*, 2024.
- [45] S. Shekhar, T. Dubey, K. Mukherjee, A. Saxena, A. Tyagi, and N. Kotla, "Towards optimizing the costs of llm usage," *arXiv preprint arXiv:2402.01742*, 2024.
- [46] F. Bang, "Gptcache: An open-source semantic cache for llm applications enabling faster answers and cost savings," in *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, 2023, pp. 212–218.
- [47] D. Ding, A. Mallick, C. Wang, R. Sim, S. Mukherjee, V. Ruhle, L. V. Lakshmanan, and A. H. Awadallah, "Hybrid llm: Cost-efficient and quality-aware query routing," *arXiv preprint arXiv:2404.14618*, 2024.
- [48] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, and I. Stoica, "Routellm: Learning to route llms with preference data," *arXiv preprint arXiv:2406.18665*, 2024.
- [49] OpenAI, "tiktoken," 2024, accessed: 2024-08-14. [Online]. Available: <https://github.com/openai/tiktoken>
- [50] M. D. Dhandore, M. C. Agrawal, and M. P. Meena, "Enhancing credit card fraud detection through advanced ensemble learning techniques and deep learning integration."
- [51] J. Jemai, A. Zarrad, and A. Daud, "Identifying fraudulent credit card transactions using ensemble learning," *IEEE Access*, 2024.
- [52] S. Harish, C. Lakhanpal, and A. H. Jafari, "Leveraging graph-based learning for credit card fraud detection: a comparative study of classical, deep learning and graph-based approaches," *Neural Computing and Applications*, vol. 36, no. 34, pp. 21 873–21 883, 2024.

Appendix

1. Prompts

```
A new alert for a transaction we suspect might be fraudulent has come to our system. This is
the transaction number you need to investigate: trans_num (str): '{trans_num}'
```

Important:

You must stop after each step of investigation!

Listing 1. The message is given to the FAA at the beginning of an investigation. In this message, we inform it of a new alert and provide it with the suspicious transaction number. Additionally, we instruct the FAA to stop after completing the first step and await further instructions.

```
Please continue investigating the transaction we are examining!
```

```
Remember to perform just one investigation step at a time.
```

Listing 2. The first part of the prompt is used for continuing the investigation. Here, we instruct the FAA to proceed and inform it to perform one step at a time.

```
Please remember that the format of a step is as follows:
```

1. Planning Phase:

```
Use your current evidence and domain knowledge to generate new ideas to determine the
next steps in assessing whether the transaction is fraudulent or not.
```

2. Information-Gathering Phase:

```
Implement and execute the code necessary to carry out the steps outlined in the planning
phase. This may include creating a plot and retrieving relevant data.
```

3. Analysis Phase:

```
This phase involves analyzing and interpreting the newly gathered evidence. This evidence
is then integrated with information from previous steps to derive meaningful
insights about the case. At this point, the LLM may choose to stop if it determines
that enough evidence has been collected to make a decision.
```

Listing 3. This is part of the new investigation step prompt, and it is placed between the 'Next Investigation Step' message and the 'Investigation Guidelines.' It will provide the FAA with the specific investigation step process it should perform.

NOTICE:

- Remember, a step should be relevant to the transaction we are investigating!
 - If, and only if, you believe there are no new directions to examine that might change your mind about whether the transaction is fraudulent or not, you should end the investigation.
 - Do NOT use more than 1 visualization in a step.
 - You MUST USE your additional tool each time you finish plotting a chart or charts in the information-gathering phase and use this information in the analysis phase.
- ```
image_to_text function:: This function receives a description (you must include
information about the case of investigation).
```

Listing 4. The last part of the prompt is used to perform an investigation step. This part guides the assistant on when to end the investigation, ensures that only one chart is used per step, and explains when and how to use function calling for vision analysis.

### 1.5. Vision Agent Prompt.

```
You are a fraud analyst with over 15 years of experience in fraud investigation, specializing
in financial fraud, including credit card fraud. Currently, you are collaborating with
an experienced credit card fraud investigator on an ongoing case. Your task is to provide
a clear, step-by-step explanation of a complex diagram derived from code-based analysis,
focusing on data points, trends, patterns, anomalies, and outliers that may indicate
fraudulent activities.
```

```
If the chart highlights aspects of the investigation:
```

```
Compare the insights from the chart with the case at hand, identifying significant
differences and determining if the case belongs to a specific cluster.
```

Listing 5. This is the prompt we provide to the vision agent; it includes instructions on how to analyze the images we provide. It is further discussed further in the vision agent section.

Generate a report for my investigation, summarizing the steps taken and the newly gathered concrete evidence in each step.

The report should follow this structure:

- \*\*Step [Step Number]\*\* [Brief description of the step]
  - \*\*Evidence\*\*
    - [Description of evidence 1]
    - [Description of evidence 2]
    - [Additional evidence as needed]

Example:

Report:

- \*\*Step 1:\*\* Initial analysis of transaction history.
  - \*\*Evidence\*\*
    - The transaction amount is within the upper typical range for similar merchants.
- \*\*Step 2:\*\* Geographical analysis.
  - \*\*Evidence\*\*
    - The distance between the cardholder's house location and the transaction's location is ~65 km.
- \*\*step 3:\*\*...

Listing 6. This prompt, which is used by the RGA, it is used for generating the unfiltered report.

You are asked to provide a ranking for each category on an Likert scale (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree) regarding the points in the report.

ASPECTS:

- Does this point affect the level of fraud suspicion, either by increasing or decreasing it? Both outcomes should be scored positively, as long as the point alters the suspicion level.
- This point was relevant to the investigation case.
- This point provided me with new knowledge or confirmed unbased knowledge about the case.
- This point logically aligns with the rest of the report.

NOTICE:

- You must give a rating to ALL of the evidence in ALL of the steps in the report.
- Please remember that inaccurately scoring a bad report as a good report may affect the future of your employment, so be critical about your scoring method.
- The report consists of both supporting fraud evidence and supporting legitimate evidence; both types of points should be rated by the same standards.
- The order of the evidence in the report shouldn't matter for your scoring.
- When ranking each aspect, try to isolate it and not take any other aspects into consideration.

Example JSON Rankings

The unfiltered report:

- Step 1:
  - Evidence:
    - The transaction amount is within the upper typical range for similar merchants.
- Step 2:
  - Evidence:
    - The distance between the cardholder's house and the transaction location is under 100 km.
- Step 3:
  - Evidence:
    - It falls within usual transaction hours, not an outlier.

```json

```

{"report_evaluation" : [
  {
    "evidence": "The transaction amount is within the upper typical range for similar
    merchants."
    "impact_on_fraud_suspicion_level": "agree,"
    "relevant_to_investigation_case": "strongly agree,"
    "providing_new_knowledge": "disagree,"
    "logical_alignment": "agree"
  },
  {
    "evidence": "The distance between the cardholder's house and the transaction location
    is under 100 km."
    "impact_on_fraud_suspicion_level": "neither agree nor disagree,"
    "relevant_to_investigation_case": "agree,"
    "providing_new_knowledge": "agree,"
    "logical_alignment": "Agree"
  },
  {
    "evidence": "It falls within usual transaction hours, not an outlier."
    "impact_on_fraud_suspicion_level": "agree,"
    "relevant_to_investigation_case": "strongly agree,"
    "providing_new_knowledge": "Disagree,"
    "logical_alignment": "Strongly Disagree"
  },
}]
'''
Now, here is the report that I want you to evaluate and provide a rating for:
{report_to_evaluate}

```

Listing 7. This prompt is used to produce the evidence quality score