

# Large Language Models for Cryptocurrency Transaction Analysis: A Bitcoin Case Study

Yuchen Lei\*

School of Cyber Science and  
Engineering, Wuhan University  
Wuhan, China  
system@whu.edu.cn

Rafael Dowsley

Faculty of Information Technology,  
Monash University  
Melbourne, Australia  
Rafael.Dowsley@monash.edu

Yuxin Xiang\*†

Faculty of Information Technology,  
Monash University  
Melbourne, Australia  
Yuxin.Xiang@monash.edu

Qin Wang

CSIRO Data61  
Sydney, Australia  
qinwangtech@gmail.com

Jiangshan Yu

School of Computer Science,  
University of Sydney  
Sydney, Australia  
jiangshan.yu@sydney.edu.au

## Abstract

Cryptocurrencies are widely used, yet current methods for analyzing transactions heavily rely on opaque, black-box models. These lack interpretability and adaptability, failing to effectively capture behavioral patterns. Many researchers, including us, believe that Large Language Models (LLMs) could bridge this gap due to their robust reasoning abilities for complex tasks.

In this paper, we test this hypothesis by applying LLMs to real-world cryptocurrency transaction graphs, specifically within the Bitcoin network. We introduce a three-tiered framework to assess LLM capabilities: foundational metrics, characteristic overview, and contextual interpretation. This includes a new, human-readable graph representation format, LLM4TG, and a connectivity-enhanced sampling algorithm, CETaS, which simplifies larger transaction graphs.

Experimental results show that LLMs excel at foundational metrics and offer detailed characteristic overviews. Their effectiveness in contextual interpretation suggests they can provide useful explanations of transaction behaviors, even with limited labeled data.

## CCS Concepts

- Security and privacy → Intrusion/anomaly detection and malware mitigation;
- Applied computing → Network forensics;
- Computing methodologies → Machine learning.

## Keywords

LLM, Cryptocurrency, Transaction Graph, Bitcoin Network

## 1 Introduction

Large language models (LLMs) [1] have significantly boosted the productivity of daily life and have a huge impact on the research community. LLMs broadened the boundaries of numerous fields, including natural language processing (NLP) [2, 3], computer vision (CV) [4, 5], and application research [6–8]. Moreover, the applications of LLMs extend beyond traditional domains, influencing areas with social and economic implications.

One such area is the cryptocurrency ecosystem. Its growing adoption in finance, retail, and entertainment has led to a surge

in transaction volumes. However, the expansion also exposes the ecosystem to risks, such as scams and money laundering, enabled by its decentralized and pseudoanonymous nature. Current analysis methods rely on black-box models and struggle with interpretability and adaptability. In this context, applying LLMs to analyze cryptocurrency transactions offers a promising approach to bridging these gaps. By leveraging their capacity to interpret complex patterns and behaviors, LLMs can help identify illicit activities and enhance cybercrime detection efforts.

Although LLMs trained on massive datasets excel in NLP tasks, their application to graph analysis presents challenges due to structural differences between graph and text data. Recent studies investigated the possibilities of LLMs for handling graph data-related tasks, concluding affirmatively that they are capable of completing specific tasks with acceptable performance on graphs such as small graphs, citation graphs, or knowledge graphs (KGs) [9–12]. Nevertheless, measuring LLMs' capability to understand and analyze cryptocurrency transaction graphs remains impractical. They contain different information compared with the other graph types such as KGs. Taking the Bitcoin transaction network as an example, the node represents the Bitcoin address or transaction, while the edge indicates the token flows among the address nodes and the transaction nodes [13].

In addition, due to the input token limit of LLMs, how to efficiently feed larger graph data into LLMs to gain more information for potentially improving the quality of generated answers to various questions relevant to Bitcoin transaction graphs (e.g., address type prediction) continues to be an open question. To bridge the gaps in applying LLMs to transaction graph analysis, we study Bitcoin networks and address three research questions:

- **RQ1:** What graph representation formats are effective in LLMs for Bitcoin transaction graphs?
- **RQ2:** How to measure LLMs' capacity to understand or analyze Bitcoin transaction graphs?
- **RQ3:** What are the key differences between using engineered graph features and raw graph data in analyses?

We adopt quantitative methods combined with qualitative analysis to answer those research questions. For RQ1, we investigate various graph representation formats and their feasibility for LLMs.

\*Yuchen Lei and Yuxin Xiang contributed equally to this work.

†Yuxin Xiang is the corresponding author.

To reduce the token consumption of raw graphs, we propose a novel representation format called LLM4TG based on the characteristics of LLMs. As for RQ2 and RQ3, we propose three levels for measuring the understanding of transaction graph:

- **Level 1 - foundational metrics:** LLMs can determine the basic information of the graph such as the in-degree and output token amount of a node.
- **Level 2 - characteristic overview:** LLMs can figure out the highlighted characteristics of the graph, e.g., a node has a significantly large out-degree.
- **Level 3 - contextual interpretation:** LLMs can classify cryptocurrency address types for addresses without labels based on labeled address samples.

**Contribution.** Our work, for the first time, evaluates LLMs' capabilities in analyzing transaction graphs for real-world cryptocurrencies, especially Bitcoin. We make the following contributions:

- We present a layered framework (§3) with three levels of understanding for measuring LLMs' ability to analyze transaction graphs in cryptocurrency networks.
- We propose a text-based graph representation format, denoted LLM4TG (§3.2). It reduces redundant data and provides a human-readable syntax that naturally supports processing by LLMs.
- We design a Connectivity-Enhanced Transaction Graph Sampling algorithm (CETrAS) (§3.3) for graph summarization, targeting cutting off the less important nodes in middle-size transaction graphs while enhancing the critical connections.
- We applied our framework to evaluate LLMs' capabilities in real-world cryptocurrency transaction analysis using two datasets (BASD, BABD) extracted from the Bitcoin ledger and three mainstream models (GPT-3.5, GPT-4, GPT-4o), providing both quantitative assessment and qualitative analysis of their performance.

## 2 Technical Warmups

**Transaction graphs for cryptocurrencies.** Cryptocurrency transaction graphs [14–17] represent the flow of digital currency between entities on blockchain networks such as Bitcoin. Each node in the graph typically represents a transaction or wallet address, and edges represent the transfer of cryptocurrency between these nodes. These graphs are crucial for analyzing the behavior of users, identifying patterns such as fraud or money laundering, and understanding the overall dynamics of the cryptocurrency market.

**Large language models.** LLMs [1] like OpenAI's GPT (Generative Pre-trained Transformer) [18] series are advanced AI models trained on vast amounts of text data. They excel in generating coherent and contextually relevant text based on the input they receive and can perform a variety of tasks without specific task-oriented training. While primarily designed for natural language processing, LLMs can be adapted to non-textual tasks such as analyzing structured data, including graphs [19, 20]. By converting data into a format that mimics natural language or structured prompts, researchers can leverage LLMs' powerful generative and interpretive capabilities to perform complex analyses like those required for understanding cryptocurrency transaction graphs.

Within this paper, we focus on three mainstream LLM models: GPT-3.5 handles general tasks with a limited token capacity of 16K and training data up to mid-2021. GPT-4 increases the token limit

to 128K and improves its ability to tackle complex analyses and research. GPT-4o maintains the same token capacity as GPT-4 while significantly enhancing efficiency.

**Tokens in LLMs.** The term *token* indicates the smallest unit of text processed by LLMs. It can vary from individual characters, such as letters and punctuation, to more complicated units, such as words or subwords. LLMs analyze and generate text by processing token sequences. Each LLM has a maximum token limit per input, which poses a significant challenge when working with extensive data sets, such as large transaction graphs.

The token limit in LLMs like GPT-3 or GPT-4 [18] constrains data analysis in a single query, posing challenges especially in analyzing large and complex cryptocurrency transaction graphs. This limitation may result in a loss of context or missing critical data. To overcome this, researchers employ strategies such as data compression to simplify input data, selective sampling to focus on the most relevant sections of graphs, custom tokenization to optimize data representation within token limits, and iterative processing to break the analysis into manageable parts.

## 3 Methodology

### 3.1 Quick Overview

We present our framework in Fig. 1. The first step is to construct the Bitcoin transaction graph using the historical on-chain data. The second step is to extract corresponding subgraphs (or the relevant subgraphs sampled by CETrAS, §3.3) and graph features to Bitcoin addresses with labels. Finally, raw graphs (formatted as LLM4TG, §3.2) and graph features are input to LLMs with proper prompts according to the tasks on different understanding levels.

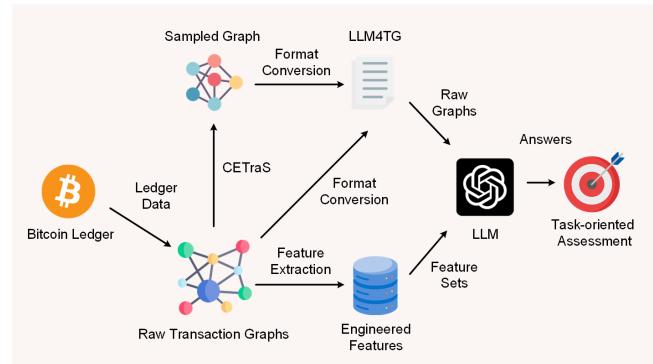


Figure 1: LLM Evaluation Framework for Bitcoin Transaction

### 3.2 LLM4TG

We introduce LLM4TG, a new format designed to optimize the analysis of transaction graphs using LLMs. This format is text-based and human-readable, minimizing syntactic noise/redundancy while reducing token usage and preserving data integrity.

In our approach, LLM4TG captures essential node information and integrates edge details directly within the nodes. It organizes nodes into layers based on their type, either address or transaction, thereby maintaining the structural integrity of the graph. This

hierarchical layering provides a segmented and clear overview of the network's dynamics. Each layer categorizes nodes which are further defined by properties such as degrees and token amounts, simplifying the analysis and enhancing readability.

We denote T as transaction and A as address.  $\langle \text{NodeID} \rangle$  represents the node's ID,  $\langle \text{Number} \rangle$  represents an integer.  $\langle \text{Float} \rangle$  represents a real number. The syntax is displayed as follows:

```

1 <LLM4TG> ::= <GraphLayer>
2 <GraphLayer> ::= "Layer" <Number> ":" <NodeCount> <
3   <NodeType> "nodes" <NewLine> <Node>+
4 <NodeCount> ::= <Number>
5 <NodeType> ::= "address" | "transaction"
6 <Node> ::= <NodeA> | <NodeT>
7 <NodeA> ::= <NodeID> "address" ":" <PropertiesA> <NewLine>
8   >
9 <NodeT> ::= <NodeID> "transaction" ":" <PropertiesT> <NewLine>
10 <PropertiesA> ::= "{" <PropertyA> ("," <PropertyA>)* "}"
11 <PropertiesT> ::= "{" <PropertyT> ("," <PropertyT>)* "}"
12 <Property> ::= <InDegree> | <OutDegree> | <InValue> | <OutValue>
13 <PropertyA> ::= <Property> | <TimeRange>
14 <PropertyT> ::= <Property> | <InNodes> | <OutNodes>
15 <InDegree> ::= "in_degree:" <Number>
16 <OutDegree> ::= "out_degree:" <Number>
17 <InValue> ::= "in_value:" <Float>
18 <OutValue> ::= "out_value:" <Float>
19 <TimeRange> ::= "time_range:" <Number>
20 <InNodes> ::= "in_nodes:" "[" <NodeIDList> "]"
21 <OutNodes> ::= "out_nodes:" "[" <NodeIDList> "]"
22 <NodeIDList> ::= <NodeID> ("," <NodeID>)* | <Empty>

```

**Listing 1: Graph Representation Definition**

This format provides three key advantages for representing and analyzing transaction graphs: 1) It organizes nodes into type-specific layers, closely mirroring the structure of original transaction graphs; 2) It efficiently utilizes the limited token budget of LLMs by allowing more data to be encoded; 3) It improves the interpretability of graph data for LLMs by organizing node attributes into closely associated key-value pairs.

To further demonstrate the effectiveness of LLM4TG, we compared its token consumption with other formats for the same graphs, as shown in Fig. 7. This comparison reveals that LLM4TG experiences a more gradual increase in token usage and consistently stays within the GPT-4/40 token limit across various graph sizes, making it a more efficient format, especially for larger graphs.

### 3.3 CETraS

Despite LLM4TG's efficiency, some transaction graphs are too large for tasks like classification that involve few-shot learning, which processes multiple graphs at once. To tackle this, we introduce CETraS, a method that condenses mid-sized transaction graphs while maintaining essential structures.

We denote  $I_{node}$  as the importance of the node.  $a_{in/out}$  is the input/output token amount.  $d_{in/out}$  is in/out-degree.  $L_s$  is the shortest distance from the node to  $n_0$ .  $\beta$  adjusts the relative significance of the node's degree. We set  $\beta = 2$  as our scheme prioritizes graph connectivity. CETraS establishes a metric of importance for each node (with logic in Algorithm 1), calculated as:

$$I_{node} = \frac{\log(a_{in} + a_{out} + 1) + \beta \cdot \log(d_{in} + d_{out} + 1)}{L_s + 1}$$

In CETraS, nodes with lower importance are prioritized for elimination to generate a subset of the nodes being preserved. The size of this retained subset is determined by a parameter that is flexible for specific demands. To maintain connectivity, paths connecting retained nodes are also preserved. Unlike other state-of-the-art graph summarization methods [21–24] focusing on keeping accuracy for structure-relevant queries or computations on large-scale graphs (billion node-level), CETraS concentrates on accurately conveying transaction-relevant information to LLMs for mid-scale graphs that typically contain thousands of nodes.

---

#### Algorithm 1: CETraS

---

**Input:** Original transaction graph  $G$ ; Target number of nodes to retain  $N_{target}$

**Output:** Sampled transaction graph  $G_{sampled}$

```

1 Function SampleGraph( $G, N_{target}$ ):
2    $I_{node} \leftarrow \{v : I_{node}[v] \text{ for each } v \in V(G)\};$ 
3    $P_{node} \leftarrow \frac{1}{I_{node}}$ , set  $n_0$  probability to 0;
4   Normalize  $P_{node}$  so that the sum equals 1;
5    $G_{subset} \leftarrow \text{Sample from } V(G) \text{ with } P_{node} \text{ until } N_{target}$ 
6   or fewer nodes are chosen;
7    $G_{sampled} \leftarrow \text{Initialize an empty graph};$ 
8   foreach node  $n \in G_{subset}$  do
9      $p \leftarrow \text{Compute shortest path from } n_0 \text{ to } n \text{ in } G;$ 
10    foreach node  $m \in p$  do
11      Add node  $m$  to  $G_{sampled}$ ;
12    end
13    Add all edges along  $p$  to  $G_{sampled}$ ;
14  end
15  return  $G_{sampled}$ ;
16 return

```

---

## 4 Evaluation and Analysis

### 4.1 Preparation

**Dataset.** We use two datasets for experiments. Both datasets are constructed from the Bitcoin transaction graph  $G_e$ . The data spans a 22-month period from 12 July 2019 to 26 May 2021.

- BASD [25] includes eight types of subgraphs. Each subgraph, with a maximum of five hops and 3,000 nodes, is generated starting from a labeled Bitcoin address (denoted as  $n_0$ );
- BABD [13] contains labeled Bitcoin addresses, each associated with 148 engineered features.

**LLM selection.** We select three LLMs, GPT-3.5, GPT-4, and GPT-4o (i.e., gpt-3.5-turbo, gpt-4-turbo, and gpt-4o) due to their capacity to handle larger token inputs (e.g., only 8,192 for LLaMA-3) and their superior performance. Both LLMs are accessed through application programming interfaces (APIs). Users can send requests with concrete prompts and receive responses.

**Prompt engineering.** We employ the few-shot prompting [26] to interact with LLMs, where only a small number of examples are provided to guide the model in performing tasks effectively.

In our experiments, we format raw transaction graphs as LLM4TG. We use CETraS in level 2 and level 3 due to the token limit in few-shot prompts. We apply GPT-3.5 only in level 3 feature-based classification, given its particularly tight token limits (Fig.7). Experiments are conducted on BASD dataset subsets, including data from its corresponding BABD addresses.

## 4.2 Level 1 - Foundational Metrics

We randomly select 50 transaction graphs, each with 10 chosen nodes (at least one per layer) for experiments. Besides, we design 12 metrics from three perspectives that are response metrics, global metrics, and node metrics, described as follows:

- **Response metrics.** To evaluate if the responses from LLMs are correctly structured, `struct_correctness` is applied.
- **Global metrics.** To assess the ability of LLMs to basic metrics understanding for the entire transaction graphs, LLMs need to find the node with the largest in/out-degree (`global_in/out_degree`), the node with the largest in/out-value (`global_in/out_value`), and the node with the largest difference between input and output values/degrees (`global_diff_degree/value`).
- **Node metrics.** To investigate the capability of LLMs to understand foundational information of concrete nodes in transaction graphs, LLMs need to obtain the node's in/out-degree (`node_in/out_degree`), the node's in/out-value (`node_in/out_value`), and the node's special information (`node_special_info_a/t`). The special information for the address node is time interval; while for the transaction node is if a specific node exists in the input/output node sets.

**Our results.** Table 1 demonstrates that LLMs are excellent at node metrics. Accuracy for most metrics is between 98.50% and 100.00%. `node_special_info_t` is the exception. This may be due to the limited capability of LLMs to match many structurally similar data in transaction graphs. Compared with node metrics, however, for the global metrics, the accuracy significantly drops, ranging from 24.00% to 58.00%. Especially, we find that compared with the other metrics (35.00% to 58.00%) in global metrics, difference-related metrics are relatively low (24.00% to 34.00%). The reason for this may be that the capability of LLMs to calculate or compare is limited.

For different LLMs, i.e., GPT-4 and GPT4o, the most various points are `struct_correctness` in response metrics and `global_in/out_value` in global metrics. The enhancement of `struct_correctness` represents that the update and optimization of the LLMs improve the quality of the response format, which completely follows the requirements in prompts. Likewise, slightly improved `global_in/out_value` also illustrate the effectiveness of model upgrade. However, most of the metrics remain at similar levels, which shows the inherent flaws of LLMs for basic information understanding, especially global metrics, in transaction graphs.

**Level 1 findings.** For foundational metrics of transaction graphs, LLMs show high utility (98.50% to 100.00%) in node metrics and acceptable accuracy in several global metrics, especially for GPT-4o (44.00% to 58.00%). This illustrates that LLMs are excellent at obtaining specific node information in Bitcoin transaction graphs, but average at calculating or comparing data in the transaction graph.

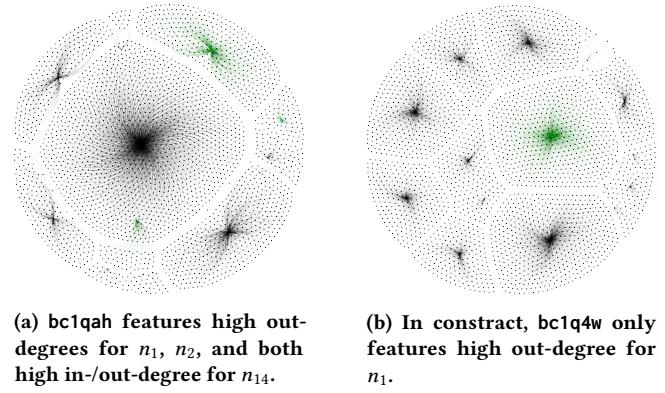


Figure 2: Examples with features

## 4.3 Level 2 - Characteristic Overview

To measure LLMs' capability in characteristic overview, we inquire about the top two most distinctive features of the subgraph. To achieve this, we randomly choose 16 subgraphs (two for each type) as knowledge pre-input to LLMs. Then, we randomly select 40 subgraphs (five for each type), and for each chosen subgraph, we combine it with the 16 reference subgraphs before inputting them into LLMs. Labels may affect the LLMs' output, so we removed them from subgraphs.

We divide the quality of LLMs' responses into three levels: high, average, and low. High-quality responses exclude invalid, inaccurate, or incorrect information. Average-quality responses include invalid or inaccuracies but no incorrect information. Low-quality responses contain incorrect information. In this context, an invalid response is one that, although accurate, fails to provide useful information. Apart from invalid/incorrect outputs, other responses generated by LLMs will be considered meaningful for analysis.

**Our results.** Based on our comprehensive empirical analysis of selected samples, the percentages of high-quality, average-quality, and low-quality cases for GPT-4 are 62.50%, 26.25% (inaccurate and invalid are 7.50% and 18.75%), and 11.25%; for GPT-4o are 82.50%, 13.75% (inaccurate and invalid are 12.50% and 1.25%), and 3.75%. That means the proportions of meaningful responses for GPT-4 and GPT-4o are 70.00% and 95.00%. We select two examples, i.e., `bc1qah`<sup>1</sup> and `bc1q4w`<sup>2</sup>, to ensure each category is included to further illustrate our results, combined with visualization by *Gephi*. We apply the ID of address node  $n_0$  to represent the corresponding subgraph.

**For bc1qah,** the responses provided by GPT-4 exhibit average or low quality (Appendix A). The first response correctly identifies that nodes  $n_1$  and  $n_2$  have high out-degrees (Fig. 2a), but it inaccurately describes their in-degrees. Although both nodes have higher in-degrees than many in the reference subgraph (typically in-degree is 1, while 4 and 3 for  $n_1$  and  $n_2$ ), their in-degrees are not particularly high within this graph. For example, node  $n_{14}$  has an in-degree of 198, significantly exceeding that of  $n_1$  and  $n_2$ . In the second response, most transactions have similar in-value and out-value;

<sup>1</sup>bc1qahe54yx133c1nwdt1ehuh4cw0fw4df62t0tnuk2

<sup>2</sup>bc1q4w090gzj7mhy3918e48gzz4uvdfn2x9y9g70h

while the differences between in-value and out-value of  $n_1$  and  $n_2$  are both about 0.003, which is trivial.

In contrast, the responses by GPT-4o are both high-quality. In the first response,  $n_{14}$  does have high in-degree and out-degree (Fig.2a), while other address nodes in Layer 2, such as  $n_{13}$  and  $n_{19}$ , also have high in-degree and out-degree. Though the second response of GPT-4o focuses on similar characteristics as GPT-4, the description of GPT-4o is accurate and reveals the high transaction values.

**For bc1q4w**, the responses by GPT-4 are high-quality or average-quality. The first response is accurate and illustrates high out-degree of  $n_1$  in Layer 1 (Fig.2b). Also, transaction nodes, including  $n_{886}$  and  $n_{1228}$  in Layer 3, have high out-degree that exceed 400. The second response is correct and demonstrates the difference between in-value and out-value. However, it is invalid since many transactions follow this pattern, which is ineffectual for transaction analysis.

In comparison, the responses by GPT-4o both have high quality. The first response is not only accurate and nearly identical to that of GPT-4, but also introduces the meaning of high out-degree for transaction nodes. The second response focuses on the same feature in the transaction node. It is accurate while meaningful for transaction analysis since the values are significantly high.

**Level 2 findings.** Despite some inaccuracies responses exist, LLMs extract numerous beneficial features for transaction analysis, achieving 70.00% and 95.00% for GPT-4 and GPT4o, respectively. This indicates that LLMs have an exceptional characteristic overview understanding capability, especially GPT-4o, highlighting their utility in characteristics identification within Bitcoin transaction graphs.

#### 4.4 Level 3 - Contextual Interpretation

We implement two experiments to evaluate the models' contextual interpretation capabilities. The first is based on *graph features*, while the second utilizes *raw graphs*. In both experiments, we apply the few-shot prompting strategy. This involves giving the LLMs with labeled subgraphs (or their features) as references and one unlabeled subgraph (or its features) for explainable classification in each iteration of LLM query.

**Graph feature-based classification.** For the graph feature-based classification tasks, we employ five labeled subgraphs per category, where each subgraph is represented by its starting Bitcoin address (i.e.,  $n_0$ ) and its corresponding features, along with 500 randomly chosen unlabeled subgraphs with features. Except for inputting the values of these features to LLMs, we also input the corresponding descriptions. We select the ten most important features in BABD [13] (Table 2). Our analysis focuses on the accuracy, macro precision, recall, and F1 scores of LLMs and their performance in each class.

The **overall accuracy** of LLMs (Fig.3a) is between 39.83% to 46.07%, which is not impressive, as well as the macro precision, macro recall, and macro F1 score. However, this is acceptable since the reference samples are significantly few. In contrast, the top-3 accuracy achieves high, ranging from 67.20% to 71.02%. This suggests that while achieving exact classification might be challenging due to the limited data, LLMs can achieve relatively high accuracy in identifying the correct classes within their top three predictions.

**Table 1: Capability of LLMs in Foundational Metrics**

Metrics	GPT-4	GPT-4o
<i>struct_correctness</i>	80.00% (40/50)	100.00% (50/50)
<i>global_in_degree</i>	50.00% (20/40)	44.00% (22/50)
<i>global_out_degree</i>	50.00% (20/40)	58.00% (29/50)
<i>global_in_value</i>	37.50% (15/40)	56.00% (28/50)
<i>global_out_value</i>	35.00% (14/40)	48.00% (24/50)
<i>global_diff_degree</i>	27.50% (11/40)	34.00% (17/50)
<i>global_diff_value</i>	27.50% (11/40)	24.00% (12/50)
<i>node_in_degree</i>	99.25% (397/400)	99.40% (497/500)
<i>node_out_degree</i>	100.00% (400/400)	99.80% (499/500)
<i>node_in_value</i>	98.50% (394/400)	99.00% (495/500)
<i>node_out_value</i>	98.50% (394/400)	99.00% (495/500)
<i>node_special_info_a</i>	99.08% (216/218)	100.00% (259/259)
<i>node_special_info_t</i>	70.88% (129/182)	67.63% (163/241)

Beyond overall metrics, **class-specific** results are displayed in Fig.4. GPT-4 and GPT-4o show strong performance in mining pool, achieving 80.00% and 95.00% precisions; They also reach extremely high recalls for darknet market which are 98.46% and 96.92%. For the majority of classes, GPT-4 and GPT-4o show better performance, though in some classes GPT-3 has similar or higher performance (e.g., recall for money laundering and precision for darknet market).

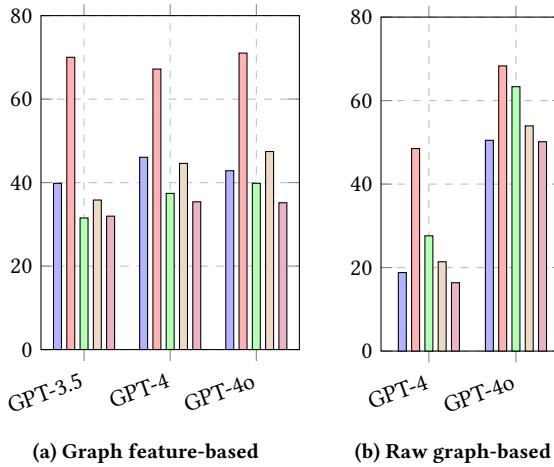
**Table 2: Summary of important Metrics in analysis**

Label	Metric Description
<i>S2-2</i>	Maximum out-degree in subgraphs
<i>S1-6</i>	Standard deviation of in- and out-degree in subgraphs
<i>S1-2</i>	Standard deviation of in-degree in subgraphs
<i>S3</i>	Degree correlation of subgraphs
<i>PA1a21-1</i>	Ratio of the minimum input token amount of an address node to the total input token amount of the address node
<i>PT1a41-2</i>	Minimum transaction time interval of an address node
<i>S6</i>	Longest distance between any two nodes in the subgraph
<i>S5</i>	Closeness centrality of the subgraph
<i>CI3a32-2</i>	Maximum change ratio in in-degree to each transaction time interval for the address node in chronological order
<i>S7</i>	Density of the subgraph

Overall, GPT-4o achieves the highest top-3 accuracy (71.02%) and recall (47.45%), while also outperforming GPT-3.5 and GPT-4 in macro precision. However, its accuracy remains lower than GPT-4, and its F1 score, while surpassing GPT-3.5, is slightly lower than GPT-4. Nevertheless, GPT-4o demonstrates better stability, with fewer categories having macro precision, recall, or F1 scores below 10%. These results suggest that despite updates leading to marginal improvements in graph feature-based classification, GPT-4o still offers mild but limited advantages over GPT-3.5 and GPT-4.

**Raw graph-based classification.** For the raw graph-based classification, we utilized three labeled subgraphs per category and 101 unlabeled subgraphs, due to higher computational costs.

The **overall accuracy** of GPT-4 (Fig.3b) is significantly lower than desired. In contrast, GPT-4o surpasses all feature-based classification tasks, achieving an accuracy of 50.49%. Additionally, the macro precision, recall, and F1 scores of GPT-4 on raw graph-based classification tasks fall below those of GPT-3.5, GPT-4, and GPT-4o



**Figure 3: Classification via different pipelines (x for LLM models, y for percentage (%); bars 1st-5th represents accuracy, top-3 accuracy, precision, recall and F1 score, respectively)**

on graph feature-based classification tasks. Conversely, the metrics for GPT-4o on graph-based classification tasks are notably higher than those on graph feature-based classification tasks, all exceeding 50.14%. As for top-3 accuracy, GPT-4 is still lower than average, while GPT-4o achieves similar performance on raw graph-based classification compared with those on feature-based.

For the **specific class** as shown in Figure 5, GPT-4o performs better than GPT-4 in almost all metrics, in some classes such as dark-net market, the differences of F1 scores are even close to 80%. One remarkable exception is that both GPT-4 and GPT-4o recorded zero scores in all metrics on blackmail; However, for GPT-4, classes including gambling, and Ponzi scheme recorded zero scores across all metrics, reflecting a huge failure in identifying relevant instances.

To sum up, GPT-4o achieves significantly higher performance than GPT-4 across all metrics in raw graph-based tasks and outperforms it in most specific classes. This demonstrates that optimizing LLMs may effectively enhance raw graph-based classification tasks.

**Extended with more models.** As illustrated in Fig.6, in the context of a very small number of reference samples, LLMs based on graph features perform significantly better than support vector machine (SVM) and MLP. However, they are inferior compared to DT, RF, CatBoost, and GNN. For LLMs based on raw graphs, GPT-4 slightly outperforms SVM and MLP, with noticeable advantages in accuracy and precision, although it is at a low level.

Notably, GPT-4o, also using raw graphs, achieves much higher accuracy than other LLMs using graph features and is very close to DT, RF, CatBoost, and GNN. Specifically, its precision greatly exceeds that of these tree models. This is crucial for illegal address detection because misidentifying legitimate addresses as illegal can have severe consequences, such as account suspension and even fund freezing. Overall, the performance of GPT-4o analyzing sampled graphs is comparable to the performance of tree models processing features selected through crafty feature engineering.

**Analysis of classification explanation.** Unlike traditional results of classification tasks, LLM-based outcomes can include detailed explanations that may be valuable for further analysis. However, our initial investigation indicates that these explanations are not always accurate. We choose two near-accurate explanations for 1Bsjsa<sup>3</sup> and 124mpe<sup>4</sup> (Appendix A) from GPT-4o to show the potential of LLMs and how they utilize different types of data.

As demonstrated above, GPT-4o can almost accurately categorize 1Bsjsa from two perspectives. From a graph feature perspective, *CI3a32-2* and *S7* are the primary factors influencing GPT-4o to classify it as a Ponzi address. *S5* and *S6* also play significant roles, where the values of these features compared with the reference graph features are crucial in this determination. From a raw graph perspective, the high in-degree and out-degree with massive small-value transactions is the key reason for the conclusion.

Similarly, GPT-4o is capable of accurately identifying 124mpe from two perspectives. From a graph feature perspective, *PTIIa41-2* is the primary factor influencing GPT-4o to classify it as a gambling address. *S2-2*, *S1-6*, *S1-2*, and *S3* also play significant roles. The values of these features compared with reference graph features are crucial in determining its classification. From a raw graph perspective, the high in-degree transactions and multiple out-degree addresses with significant transaction values are key reasons for this conclusion, strongly indicating gambling, pool, or exchange activities.

These examples illustrate that for feature-based classification, adequate quantity and quality samples with labels are required to improve classification consequences since the selected features are already processed. In contrast, for graph-based classification, the quality of samples themselves seems to be also essential. The unsuitable compression of the graphs leading to much information loss may negatively affect the results. Also, the chain of thought (CoT) prompting [26] could be used for further explanations, with detailed reasoning and better classification interpretability.

**Level 3 findings.** LLMs exhibit strong capabilities in contextual interpretation, achieving high top-3 accuracy rates even with limited data sets. They provide plausible explanations, although these are not always completely accurate. Despite these strengths, the overall accuracy of LLMs in classification tasks is moderate, indicating significant room for improvement in their performance.

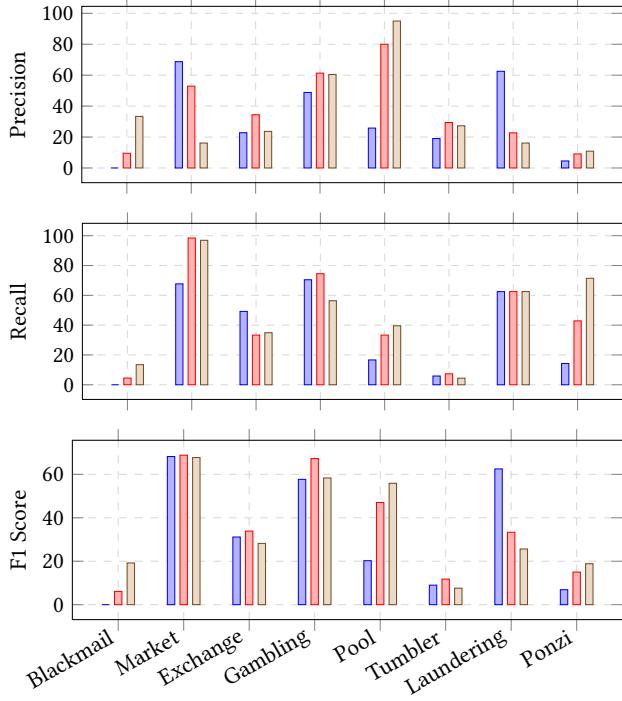
## 5 Discussion

**Addressing RQs.** We summarise our answers to RQs.

- **Answer to RQ1.** We proposed LLM4TG format. The format demonstrates significant superiority (Fig.7). When combined with CETraS, it transforms tasks that were previously almost impossible into achievable outcomes, which means the combination enables experiments at various levels of understanding.
- **Answer to RQ2.** We proposed a three-level framework for better understanding transaction graphs. For each level, we have established clear standards and conducted experiments to validate

<sup>3</sup>1BsjsaHST2Qohs8ZHxNHeZ1UfWhtxoKHEN

<sup>4</sup>124mpePGM2vEqcHUK96wQXur9vS7Vn7Kdj



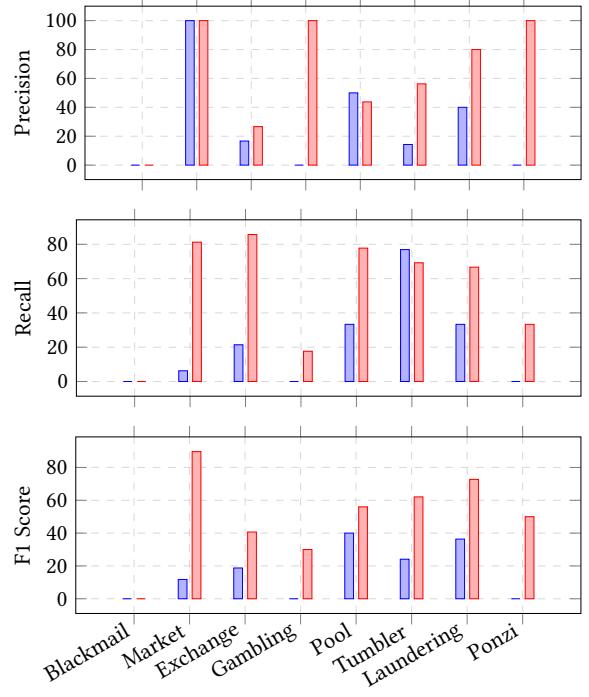
**Figure 4: LLMs’ performance in contextual interpretation using graph features (x axis for category, y for rate (%); GPT-3.5 in blue bar, GPT-4 in red, GPT-4o in brown).**

their practical application. The results underscore the framework’s effectiveness in enhancing the analytical capabilities of LLMs in this domain.

- **Answer to RQ3.** We experimentally demonstrate that the engineered graph feature-based analyses focus on pre-computed metrics, simplifying complexity. In contrast, raw graph data analysis directly interprets structural relationships, providing deeper but more challenging insights. GPT-4o performs better with raw graph data compared to other LLMs using engineered features, whereas GPT-4 performs the worst with raw graph data (Fig.6).

**Token consumption for different graph formats.** We studied the differences in token consumption among various graph representation formats under the byte pair encoding-based tokenizer `c1100k_base`, the default GPT-3.5 and GPT-4 tokenizer. We chose three well-defined formats, i.e., GEXF, GML, and GraphML, for their efficiency and flexibility in representing transaction graphs. These widely used formats offer diverse encoding styles while supporting complex attributes of nodes and edges.

As in Fig.7, token usages of these formats exhibit significantly steeper curves, i.e., they consume a substantially larger number of tokens with an increasing number of nodes. These graph formats have extreme syntactic noise (i.e., great redundancy exists when representing graph data), which would result in numerous unnecessary tokens for LLMs to consume. The maximum token limits for GPT-3.5 and GPT-4/4o (16,385 and 128,000) are indicated by horizontal lines, where token usages of these formats surpass the



**Figure 5: LLMs’ performance in contextual interpretation using raw graphs (x axis for category, y for rate (%); GPT-3.5 in blue bar, GPT-4 in red).**

GPT-4/4o limit when the number of nodes of a single graph exceeds approximately 500 to 750. Therefore, these graph representation formats are insufficient for LLMs to analyze transaction graphs.

We can also observe a counterintuitive phenomenon, i.e., at some points of the graphic, the number of necessary tokens decreases while the number of nodes increases. The explanation is that the total size of a graph is determined not just by the number of nodes, but also by the edges and their corresponding attributes (e.g., a graph may have more nodes but fewer edges than another).

**How LLMs benefit transaction graph analysis.** We summarise three advantages based on our experience applying LLMs.

- **High accuracy with minimal data.** Utilizing LLMs for cryptocurrency transaction analysis and monitoring only requires minimal data. LLMs can provide high top-3 accuracy results with explanations as well as excellent precision when using raw graph data. Many addresses play multiple roles, such as exchanges may be involved in money laundering. LLMs explanations could be efficient for experts to further analyze, especially for some illicit transaction types with only a few samples.
- **Contextual aids.** The underlying human motivations behind different types of transactions vary significantly and can reveal behavioral patterns often overlooked by purely data-driven approaches. Traditional models typically focus on identifying data patterns, neglecting the subtle reasoning behind user actions. In contrast, LLMs offer more context-aware interpretations, leveraging their extensive training in human language and behavior

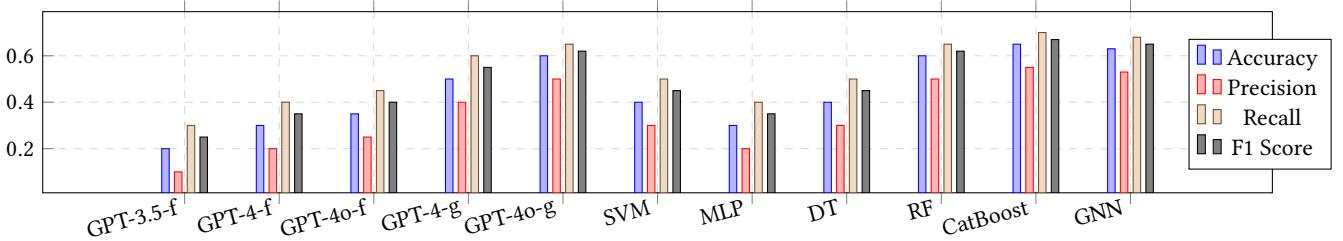


Figure 6: Evaluation on different models (x axis for category, y for corresponding rates (%))

to infer intent or reasoning that conventional models fail to capture. While the explanations provided by LLMs may not always be entirely accurate, they still offer insights by interpreting the intent behind transactions.

- *Graph understanding*: LLMs can extract node information and provide meaningful overviews of transaction graphs, identifying complex patterns and relationships effectively. Unlike traditional algorithms, which often struggle with complex, high-dimensional data and contextual information embedded within graphs, LLMs leverage their capacity for processing raw data and capturing intricate patterns, relationships, and anomalies.

**Remaining challenges.** We stress three challenges when applying LLMs to cryptocurrency transaction graphs.

- *Token limits* restrict the amount of graph data that can be processed at once. This constraint impacts the analysis of large cryptocurrency transaction graphs, as LLMs often lack sufficient contextual information to deliver accurate insights. When only a limited portion of the graph is available, it becomes challenging to capture the complexity and nuances of data.
- The *selection of reference graphs*, i.e., the identification of the most representative labeled samples, is another issue. Different graphs may emphasize varying features, introducing bias. The choices can significantly influence classification outcomes and their explanations. Token limits further complicate this process. They restrict the number of samples that can be used as references, particularly for raw graph data, limiting the model's ability to generalize its findings.
- Improving the *accuracy of LLM-generated explanations* is crucial yet hard. Developing rigorous methods to enhance the precision of these explanations is necessary. Enhancing the precision of these explanations could lead to deeper analytical capabilities, potentially advancing both the interpretability and effectiveness of cryptocurrency transaction analysis.

**Factors that may affect LLM performance?** The performance is influenced by both the models used and the input data.

- *Model bottleneck*. Though the major update and optimization could enhance the capability of LLMs in some aspects, it may be unable to solve the inherent problem of LLMs on transaction graph understanding. For example, using the same graph feature data for LLM-based classification, compared with GPT-3.5, the effectiveness of GPT-4 and GPT-40 improves slightly. Currently, a more effective method to boost LLMs' performance might involve using more labeled samples and enriching the feature set in graph

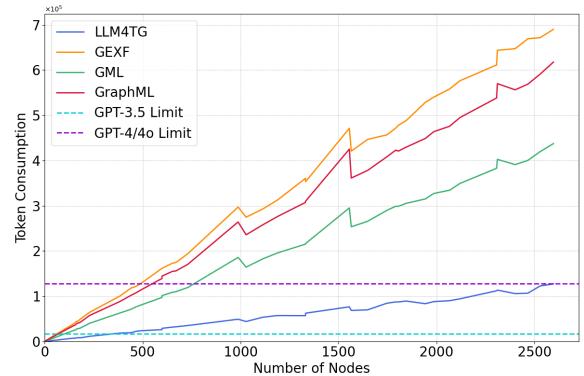


Figure 7: Token consumption in different graph formats

feature-based classifications. Feature data requires fewer LLM tokens compared to processing raw graphs.

- *Graph complexity/size*. The size and complexity of transaction graphs impact LLM performance. Smaller graphs yield better foundational metric performance, while larger graphs may reduce effectiveness. At the contextual interpretation level, we found that even when using the same data, the accuracy results may differ slightly, by approximately 5%. Finally, due to the application of LLM4TG, there is a slight loss of temporal information, which may affect the results using raw graphs.
- *Data representation and features*. The type of data used (e.g., graph features versus raw graphs) and the availability of labeled samples affect performance. Graph feature-based data is more efficient in terms of LLM token consumption and may improve classification effectiveness with additional features and labels.

## 6 Conclusion

This work evaluates LLMs' capabilities in analyzing Bitcoin transaction graphs. We introduced a three-level framework along with two key innovations: LLM4TG format to enhance readability and reduce graph sizes, and CETraS algorithm to optimize graph simplification. Our experiments demonstrate satisfactory accuracy in foundational metrics, effectiveness in obtaining useful overview characteristics, and solid top-3 accuracy in classification tasks. These findings highlight the significant potential of LLMs and establish a foundation for their broader application in cryptocurrency analysis.

## Acknowledgment

This research is supported by OpenAI Researcher Access Program.

## References

- [1] Yupeng Chang, Xu Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 15(3):1–45, 2024.
- [2] Bonan Min, Hayley Ross, et al. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 2023.
- [3] Robert Timm, Hao Cheng, et al. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4), 2023.
- [4] Wenhui Wang, Zhe Chen, et al. VisionLlm: Large language model is also an open-ended decoder for vision-centric tasks. *NeurIPS*, 36, 2024.
- [5] Deyao Zhu, Jun Chen, et al. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.
- [6] Jingxuan He and Martin Vechev. Large language models for code: Security hardening and adversarial testing. In *CCS*, pages 1865–1879, 2023.
- [7] Gustavo Sandoval, Hammond Pearce, et al. Lost at C: A user study on the security implications of large language model code assistants. In *USENIX Sec*, 2023.
- [8] Karan Singhal, Shekoofeh Azizi, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [9] Jiayian Guo, Lun Du, and Hengyu Liu. GPT4Graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*, 2023.
- [10] Yuntong Hu, Zheng Zhang, and Liang Zhao. Beyond text: A deep dive into large language models' ability on understanding graph data. *arXiv preprint arXiv:2310.04944*, 2023.
- [11] Heng Wang, Shangbin Feng, et al. Can language models solve graph problems in natural language? In *NeurIPS*, volume 36, pages 30840–30861, 2023.
- [12] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Yijian Qin, and Wenwu Zhu. LLM4DyG: Can large language models solve spatial-temporal problems on dynamic graphs? In *KDD*, page 4350–4361, 2024.
- [13] Yuexin Xiang, Yuchen Lei, et al. BABD: A Bitcoin address behavior dataset for pattern analysis. *TIFS*, 19:2171–2185, 2024.
- [14] Ting Chen, Zihao Li, Yuxiao Zhu, Jiachi Chen, Xiapu Luo, John Chi-Shing Lui, Xiaodong Lin, and Xiaosong Zhang. Understanding Ethereum via graph analysis. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–32, 2020.
- [15] Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I Weidele, Claudio Bellei, Tom Robinson, and Charles E Leiserson. Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv preprint arXiv:1908.02591*, 2019.
- [16] Dongchao Guo, Jiaqing Dong, and Kai Wang. Graph structure and statistical properties of Ethereum transaction relationships. *Information Sciences*, 2019.
- [17] Guangsheng Yu, Qin Wang, Tanzela Altaf, Xu Wang, Xiwei Xu, and Shiping Chen. Predicting nft classification with GNN: A recommender system for web3 assets. In *ICBC*, 2023.
- [18] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [19] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. Large language models on graphs: A comprehensive survey. *TKDE*, 2024.
- [20] Yuhang Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. A survey of graph meets large language model: Progress and future directions. In *IJCAI*, 2024.
- [21] Mahdi Hajibabai, Jasbir Singh, Venkatesh Srinivasan, and Alex Thomo. Graph summarization with controlled utility loss. In *KDD*, pages 536–546, 2021.
- [22] Shinhyun Kang, Kyuhan Lee, and Kijung Shin. Personalized graph summarization: formulation, scalable algorithms, and applications. In *ICDE*, 2022.
- [23] Kyuhan Lee, Hyeonsoo Jo, Jihoon Ko, Sungsu Lim, and Kijung Shin. Ssumm: Sparse summarization of massive graphs. In *KDD*, pages 144–154, 2020.
- [24] Meiquan Lai, Yaqi Huang, et al. An optimized lossless graph summarization for large-scale graphs. In *ICPADS*, 2023.
- [25] Yuexin Xiang et al. Leveraging subgraph structure for exploration and analysis of Bitcoin address. In *IEEE BigData*, 2022.
- [26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, volume 35, pages 24824–24837, 2022.
- [27] Qin Wang, Guangsheng Yu, and Shiping Chen. Cryptocurrency in the aftermath: Unveiling the impact of the SVB collapse. *TCSS*, 2024.
- [28] Jintao Huang, Ningyu He, Kai Ma, Jiang Xiao, and Haoyu Wang. Miracle or mirage? a measurement study of NFT rug pulls. *SIGMETRICS*, 7(3):1–25, 2023.
- [29] Qin Wang, Guangsheng Yu, Yilin Sai, Caijun Sun, Lam Duc Nguyen, Sherry Xu, and Shiping Chen. An empirical study on snapshot DAOs. *arXiv preprint arXiv:2211.15993*, 2022.
- [30] Bingyu Gao, Haoyu Wang, et al. Tracking counterfeit cryptocurrency end-to-end. *SIGMETRICS*, 4(3):1–28, 2020.
- [31] Natkamon Tovanich, Nicolas Soulié, et al. An empirical analysis of pool hopping behavior in the Bitcoin blockchain. In *ICBC*, 2021.
- [32] Binbing Hou and Feng Chen. A study on nine years of Bitcoin transactions: Understanding real-world behaviors of bitcoin miners and users. In *ICDCS*, 2020.
- [33] Xigao Li, Anurag Yepuri, and Nick Nikiforakis. Double and nothing: Understanding and detecting cryptocurrency giveaway scams. In *NDSS*, 2023.
- [34] Bowen He, Yuan Chen, et al. TxPhishScope: Towards detecting and understanding transaction-based phishing on Ethereum. In *CCS*, pages 120–134, 2023.
- [35] Gibran Gomez, Pedro Moreno-Sánchez, and Juan Caballero. Watch your back: Identifying cybercrime financial relationships in Bitcoin through back-and-forth exploration. In *CCS*, 2022.
- [36] Jiajing Wu, Dan Lin, et al. Toward understanding asset flows in crypto money laundering through the lenses of ethereum heists. *TIFS*, 2024.
- [37] Jialan Chen, Dan Lin, and Jiajing Wu. Do cryptocurrency exchanges fake trading volumes? an empirical analysis of wash trading based on data mining. *Physica A: Statistical Mechanics and its Applications*, 586:126405, 2022.
- [38] Pranav Nerurkar, Dhiren Patel, Yann Busnel, et al. Dissecting Bitcoin blockchain: Empirical analysis of Bitcoin network (2009–2020). *Journal of Network and Computer Applications*, 2021.
- [39] Luca Serena, Stefano Ferretti, and Gabriele D'Angelo. Cryptocurrencies activity as a complex network: Analysis of transactions graphs. *Peer-to-Peer Networking and Applications*, 2022.
- [40] Bishenghui Tao, Hong-Ning Dai, Jiajing Wu, Ivan Wang-Hei Ho, Zibin Zheng, and Chak Fong Cheang. Complex network analysis of the Bitcoin transaction network. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2021.
- [41] Dan Lin, Jiajing Wu, Yunmei Yu, Qishuang Fu, Zibin Zheng, and Changlin Yang. DenseFlow: Spotting cryptocurrency money laundering in Ethereum transaction graphs. In *WWW*, 2024.
- [42] Deepesh Chaudhari, Rachit Agarwal, and Sandeep Kumar Shukla. Towards malicious address identification in Bitcoin. In *IEEE International Conference on Blockchain (Blockchain)*, 2021.
- [43] M Mazhar Rathore, Sushil Chaurasia, and Dhirendra Shukla. Mixers detection in Bitcoin network: a step towards detecting money laundering in cryptocurrencies. In *IEEE BigData*, 2022.
- [44] Anton Wahrstätter et al. Improving cryptocurrency crime detection: Coinjoin community detection approach. *TDSC*, 2023.
- [45] Yu Gai, Liyi Zhou, Kaihua Qin, Dawn Song, and Arthur Gervais. Blockchain large language models. *arXiv preprint arXiv:2304.12749*, 2023.
- [46] Chang Liu and Bo Wu. Evaluating large language models on graphs: Performance insights and comparative analysis. *arXiv preprint arXiv:2308.11224*, 2023.
- [47] Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. Table meets LLM: Can large language models understand structured table data? a benchmark and empirical study. In *WSDM*, 2024.
- [48] Jinhao Jiang, Kun Zhou, zican Dong, KeMing Ye, Xin Zhao, and Ji-Rong Wen. StructGPT: A general framework for large language model to reason over structured data. In *EMNLP*, 2023.
- [49] Debarati Das, Ishaa Gupta, Jaideep Srivastava, and Dongyeop Kang. Which modality should i use—text, motif, or image?: Understanding graphs with large language models. *arXiv preprint arXiv:2311.09862*, 2023.
- [50] Zhikai Chen, Haitao Mao, et al. Exploring the potential of large language models (LLMs) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 2024.
- [51] Zhikai Chen, Haitao Mao, et al. Label-free node classification on graphs with large language models (LLMs). In *ICLR*, 2024.
- [52] Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (LLMs)? a.k.a. will LLMs replace knowledge graphs? In *Proceedings of the NAACL-HLT*, 2024.

## A LLMs Responses

### Level 2 Example - bc1qah.

#### GPT-4:

- High in-degree and out-degree of transactions; Transactions like  $n_1$  and  $n_2$  have high in-degrees and out-degrees. [Average - Inaccurate]
- Significant value differences; The in- and out-value of certain transactions show significant volume and discrepancies, such as in  $n_1$  and  $n_2$ . [Low]

#### GPT-4o:

- High in-degree and out-degree of addresses in Layer 2; The addresses in Layer 2 have exceptionally high in-degree and out-degree values, indicating a large number of transactions. For example, address  $n_{14}$  has an in-degree of 198 and out-degree of 188. [High]
- High in-value and out-value in transactions; The transactions in this graph involve significant amounts of value. For instance, transaction  $n_1$  has an in-value of 77.29740945 and out-value of 77.29452845. [High]

### Level 2 Example - bc1q4w.

#### GPT-4:

- High out-degree in transaction nodes; A noticeable feature in this graph is the significant out-degree in some transaction nodes. For instance,  $n_1$  has an out-degree of 600. [High]
- Value discrepancy between in-value and out-value in transactions; The transaction node  $n_1$  shows a discrepancy between the total incoming value (in-value: 27.69691553) and the outgoing value (out-value: 27.69184153). [Average - Invalid]

#### GPT-4o:

- High out-degree transactions; The transaction node  $n_1$  has an extremely high out-degree of 600, meaning it distributes funds to 600 different address nodes. [High]
- Large volume of transactions; The transaction node  $n_1$  also handles a substantial volume of in-value and out-value (27.69691553). [High]

### Level 3 Example - 1Bsjsa.

**Ground Truth:** Ponzi

**Feature-based Result:** Ponzi, blackmail, tumbler

#### Reason:

- $C13a32-2$  value is extremely high, which is a distinctive characteristic of Ponzi graphs.
- $S7$  is also higher than average, another indicative feature for Ponzi. Additionally,  $S5$  and  $S6$  align well with Ponzi characteristics.
- Blackmail and tumbler categories are also considered due to similar  $S2-2$  and  $S6$  values.

**Graph-based Result:** Ponzi, tumbler, pool

#### Reason:

- The graph has a structure where a single address node ( $n_0$ ) has a high degree of both incoming and outgoing transactions (300 in-degree and 197 out-degree). This pattern is indicative of schemes like Ponzi, tumbler, and pool, where a central node is involved in numerous transactions with other nodes.
- The high number of small-value transactions and the presence of multiple transaction nodes with varying degrees further support these categories.

### Level 3 Example - 124mpe.

**Ground Truth:** gambling

**Feature-based Result:** gambling, darknet market, exchange

#### Reason:

- The high value of  $PT1a41-2$  is a strong indicator of gambling as similar high values are observed in reference graphs.
- The  $S2-2$ ,  $S1-6$ ,  $S1-2$ , and  $S3$  values are within ranges observed in darknet market and exchange categories, making them the next most probable categories.

**Graph-based Result:** gambling, pool, exchange

#### Reason:

- The structure and value distributions of the graph suggest that it is most likely a gambling, pool, or exchange category. The presence of large in-degree transactions and multiple out-degree addresses is characteristic of gambling and pool activities, as seen in the reference graphs.
- The significant transaction values and multiple address connections also align with the patterns found in exchange-related graphs

## B Related Work

**Empirical analysis in cryptocurrency.** Empirical analysis plays an important role in understanding the dynamics of the cryptocurrency ecosystem [27–29] and the behaviors of addresses/entities [14, 30]. For instance, Tovanich et al. [31] and Hou et al. [32] reveal that factors such as payout schemes and pool fees influence miners’ behaviors in Bitcoin mining pools, and then impact the overall system performance.

Empirical analysis can also aid in identifying cryptocurrency scams. Li et al. [33] used *CryptoScamTracker* to analyze cryptocurrency giveaway scams. He et al. [34] developed *TxPhish* to detect Ethereum scams in which users lured by high profits to fake websites are tricked into signing transactions that allow scammers to steal their crypto assets. Gomez et al. [35] explored bidirectional transactions to map cybercrime networks. Wu et al. [36] focused on identifying Ethereum-based money laundering via asset flows. Chen et al. [37] combined on-chain and off-chain data to detect fake trading volumes of famous exchanges Huobi and Binance.

**Cryptocurrency analysis with advanced techniques.** We introduce two methods as below.

- *Complex networks* use graph theory, centrality measures, and network topology analysis to examine patterns within networks. This method explores the structures of cryptocurrency transaction networks from a macro perspective, highlighting the interconnections between nodes and the network’s overall structure.

Nerurkar et al. [38] and Serena et al. [39] drew parallels between cryptocurrency systems and other complex systems and identified characteristics such as small-world property, indicating most nodes in the graph are not neighbors but most of them can be reached by every other within a few hops. Moreover, Tao et al. [40] employed an innovative random walk with a flying-back sampling method on Bitcoin transaction graphs, uncovering phenomena such as the non-rich-club effect, i.e., that high-degree nodes are not more interconnected among themselves than with lower-degree nodes. Guo et al. [16] analyzed Ethereum transaction graphs, revealing heavy-tailed property in transaction networks, i.e., the majority of nodes have a relatively low degree while a small number of nodes have a very high degree. To mitigate the issue of money laundering on blockchain networks, *DenseFlow* framework, proposed by Lin et al. [41], uses dense subgraphs and the maximum flow algorithm to trace laundering activities. This approach improves precision compared to existing methods on Ethereum, demonstrating the effectiveness of network analysis in combating money laundering.

- *Machine learning* is used to achieve node- or graph-level classification and prediction tasks for concrete addresses/entities.

Chaudhari et al. [42] studied utilizing temporal features to detect Bitcoin address behavioral changes and identify money laundering activities. The proposed approaches based on the decision tree (DT) by Rathore et al. [43] show high accuracy rates in detecting illicit activities and phishing scams in cryptocurrencies. Wahrstatter et al. [44] also contributed by enhancing the detection of criminal activities in Bitcoin transactions using unsupervised learning. Additionally, various machine learning methods, such as random forest (RF), multilayer perceptron (MLP), and graph neural network (GNN), are applied to benchmark cryptocurrency datasets, including the *Elliptic Data Set* by Weber et al. [15] and datasets by Xiang

et al. [13, 25]. Besides, Gai et al. [45] proposed a transformer-based anomaly detection model *BlockGPT* for the Ethereum network that demonstrated acceptable utility.

**LLMs in graph analysis.** Several studies introduced LLMs to graph analysis, using LLM as a classifier and GNN enhancement [20]. Wang et al. [11, 17] evaluated the basic capabilities of LLMs in natural language graph problem-solving. Both studies showed limitations of LLMs, particularly in solving complex graph structures and tasks. Complementing these insights, Liu and Wu [46] and Hu et al. [10] assessed the performance of LLMs in graph data analysis and prediction, compared them with specialized GNNs. Likewise, Sui et al. [47] and Jiang et al. [48] explored the effectiveness of LLMs including GPT-3.5 and GPT-4 in processing structured data, such as tables and various structured data types, introducing innovative prompting methods for performance enhancement.

Besides, Das et al. [49], Chen et al. [50, 51], and Guo et al. [9] adopted a different approach by integrating LLMs with graph data, focusing on graph structure analysis, node classification, and a range of graph processing tasks. These studies investigated the potential and limitations of LLMs in more specialized and advanced

graph analysis applications, offering new insights and directions for future research in LLMs and graph data analysis. Moreover, Sun et al. [52] uniquely studied the factual knowledge of LLMs, providing a broader perspective on their comprehension capabilities, especially for lesser-known entities and facts.

**Gaps in existing studies.** We conclude three primary research limitations according to the abovementioned work:

- Existing research mainly focuses on knowledge graphs and randomly generated graphs [10, 11, 46, 49]. However, how to measure the LLMs' ability to understand and analyze real-world cryptocurrency transaction graphs is still unresolved.
- Common graph representation formats, such as GEXF and GraphML, are not ideally suited for LLMs due to their inherent space constraints. This limitation explains why recent studies have focused exclusively on testing LLMs with smaller graphs [11, 49] (e.g., graphs containing ten nodes or marginally more [11]).
- In addition to applying raw graph data to LLMs, the effect of using engineered graph features for the cryptocurrency transaction graph analysis remains insufficiently studied.