HyperGLLM: An Efficient Framework for Endpoint Threat Detection via Hypergraph-Enhanced Large Language Models

Hongyi Zhou[1], Jianfeng Pan[2], Min Peng[2*], Shaomang Huang[2], Hanzhong Zheng[2]

[1]Department of Computer Science and Technology, Tsinghua University. Beijing, China
[2]360 Security Technology Inc. Beijing, China
hy-zhou23@mails.tsinghua.edu.cn, {panjianfeng, pengmin1, huangshaomang, zhenghanzhong}@360.cn

Abstract

Endpoint Detection and Response (EDR) systems are a cornerstone of modern threat detection and endpoint protection. However, conventional heuristic- and learning-based approaches often fail to address sophisticated and continuously evolving attack patterns. Recent advances in large language models (LLMs) offer promising capabilities for behavioral analysis in EDR logs, yet their effectiveness is hindered by the massive volume of events and the interleaved nature of behavior sequences, where subtle and sporadic malicious actions are intricately interwoven with benign ones—posing significant challenges for long-context modeling and stealthy threat detection. To address these issues, we propose HyperGLLM, a novel detection framework that introduces hypergraph reasoning into LLMs. It first constructs an attribute-value level relation-aware graph to model low-order structural semantics while reducing textual redundancy. Then, it introduces a differential hypergraph module with multi-granularity clustering to capture high-order behavioral dependencies embedded in interleaved events and reinforce threat semantics. Finally, the hypergraph representations are aligned with an LLM for efficient contextual reasoning over potential malicious behaviors. To facilitate empirical evaluation, we curate EDR3.6B-63F, a large-scale EDR dataset containing 3.6 billion events across 63 distinct behavior families. Extensive experiments demonstrate that HyperGLLM significantly outperforms state-of-the-art methods by reducing the false alarm rate to 1.67%, achieving 94.65% accuracy across 63 behavior families, and improving the modeling efficiency of LLMs on long EDR logs. Our framework and dataset provide a solid foundation for future research and support the development of advanced detection solutions in endpoint security.

Introduction

The proliferation of digital infrastructure and endpoint devices has fueled increasingly sophisticated and stealthy cyber threats (Zhang et al. 2022; Chen et al. 2023). Traditional perimeter defenses, including firewalls and antivirus software, have shown limited effectiveness against persistent and evasive attacks (Kokulu et al. 2019). As a core component of modern cybersecurity architecture, Endpoint Detection and Response (EDR) systems continuously monitor endpoint activities and leverage data analytics to detect anomalous behavior and facilitate timely threat mitigation (Dong et al. 2023a). As a result, effective analysis of EDR logs is essential for uncovering malicious activities and enhancing automated threat detection capabilities.

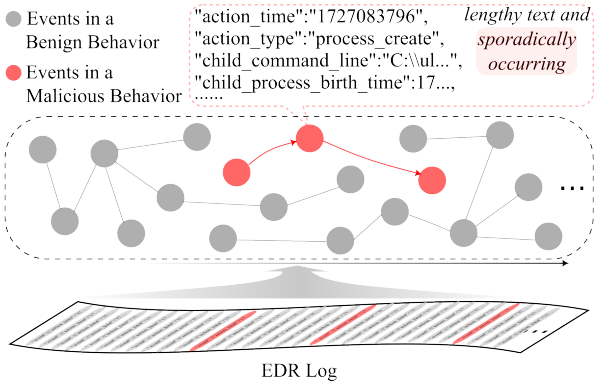

Figure 1: A typical EDR sample contains numerous interleaved, lengthy event descriptions, resulting in ultra-long logs. The sporadic occurrence and benign appearance of malicious events further complicate precise threat modeling.

Early approaches to EDR log analysis were primarily based on heuristic rules (Milajerdi et al. 2019; Hassan, Bates, and Marino 2020), which heavily rely on domain expertise and struggle to adapt to evolving or stealthy threats. To overcome these limitations, learning-based methods (Rosenberg et al. 2021; Kaur, Gabrijelčič, and Klobučar 2023; Macas, Wu, and Fuertes 2024) have been increasingly adopted to improve detection accuracy and generalization in security analytics. More recently, large language models (LLMs) (Achiam et al. 2023) have achieved impressive results in cybersecurity tasks, including malware analysis (Mohseni et al. 2025), intrusion detection (Li et al. 2024), phishing detection (Cao et al. 2025), and vulnerability management (Liu et al. 2024), due to their advanced capabilities in semantic understanding and contextual reasoning.

However, directly applying LLMs to EDR logs remains non-trivial due to the semantic stealthiness of malicious behaviors embedded within numerous interleaved event sequences. As illustrated in Fig.1. A typical EDR sample com-

*Corresponding author.

prises a large number of system events, each represented as a lengthy textual description. These ultra-long logs frequently exceed the context window limits of current LLMs (e.g., beyond 128K tokens) and incur substantial computational overheads. In addition, malicious behaviors are typically sporadically distributed, benign-appearing, and intricately interleaved with normal activities. This stealthiness poses significant challenges for accurate threat behavior modeling. Furthermore, the lack of large-scale EDR datasets with diverse and representative malware families (Alsaheel et al. 2021; Zengy et al. 2022; Sharif et al. 2024; Dong et al. 2023b) continue to hinder the development and evaluation of LLM-based solutions in this domain.

To address these challenges, we propose HyperGLLM, a unified framework that introduces hypergraph neural networks into large language models for malicious behavior inference. HyperGLLM is designed to leverage the characteristics of EDR logs and the contextual reasoning abilities of LLMs. Concretely, we construct a relation-aware graph at the attribute-value level to model latent field dependencies and capture intra-event structure, generating low-order global semantic representations. To capture the complex, interleaved nature of behavioral sequences, we introduce a multi-granularity clustered differential hypergraph network. This module forms hyperedges at different cluster scales to capture high-order behavioral semantics, and incorporates global hypergraph differentials to enhance discriminative capacity. The resulting hypergraph embeddings are then aligned with the LLM representation space via a lightweight projection layer, enabling accurate and efficient inference of malicious behaviors. To support large-scale empirical analysis, we construct EDR3.6B-63F, a large-scale dataset with 3.6 billion EDR events and over 2 million labeled samples across 62 malicious behavior families in addition to benign ones. This dataset provides a high-quality benchmark for future research in endpoint security. Our main contributions are summarized as follows:

- We propose HyperGLLM, an efficient framework that introduces hypergraph reasoning into LLMs for malicious behavior detection in EDR logs, capturing both structural semantics and long-range temporal dependencies.
- We design an attribute-value level relation-aware graph and a differential hypergraph module with multi-granularity clustering to jointly model low- and high-order behavior semantics, thereby enhancing the semantic representation of threat behaviors.
- We construct EDR3.6B-63F, a large-scale EDR dataset that serves as a high-quality benchmark for advancing AI-driven research in endpoint security, offering diverse behavior types and detailed event records.
- Extensive experiments demonstrate that HyperGLLM consistently outperforms state-of-the-art baselines across multiple metrics while maintaining high inference efficiency. Ablation studies further validated the methodology and the contribution of each module.

Related Work

Traditional Learning-based Approaches. To address the limitations of early heuristic-based detection systems (Milajerdi et al. 2019; Hassan, Bates, and Marino 2020), researchers have increasingly turned to machine learning (ML) (Raff et al. 2021; Doan et al. 2023) and deep learning (DL) (Zhu et al. 2024; Park et al. 2025) approaches for endpoint threat detection. These models (Kumar, Janet, and Neelakantan 2022; Tsai et al. 2024; Sharif et al. 2024) either rely on manual feature engineering or adopt end-to-end representation learning to capture discriminative patterns from system logs and file metadata.

Despite these advances, most models fail to capture long-range dependencies and intricate semantics in EDR logs, hindering stealthy threat detection.

LLMs in Endpoint Security. Large Language Models (LLMs) have shown exceptional capabilities in contextual semantic modeling and cross-task generalization across diverse NLP tasks (Achiam et al. 2023). These properties have recently prompted increasing efforts to explore their applicability in endpoint security (Shao et al. 2024; Levi et al. 2025). For example, Zhao et al. (2025) proposed an LLM-assisted multi-view system that uses prompt engineering to enhance malware detection. Zhou et al. (2025) developed a framework leveraging task-adaptive pre-training to improve ransomware detection. Bitaab et al. (2025) introduced SCAMNET, a fine-tuned LLM-based model for detecting fraudulent shopping websites.

While these methods highlight the potential of LLMs for endpoint security, applying them directly to EDR log analysis remains challenging due to the ultra-long log texts and the stealthy nature of malicious behavior. Despite recent advances in extending the context windows of LLMs (Peng et al. 2024; Jin et al. 2024; Ding et al. 2024), they frequently result in significant computational overhead and do not fundamentally address these challenges. In this work, we propose a novel EDR log analysis framework that integrates graph semantic modeling with LLM-based contextual reasoning, substantially improving the detection of covert malicious behavior within ultra-long contexts.

Graph-Based Approaches in Endpoint Security. Graph-based methods are widely used in endpoint security for capturing semantic dependencies among system entities (Veličković et al. 2018). Egressy et al. (2024) introduced directed multigraphs for fraud and phishing detection. Gui et al. (2025) proposed a multi-stage APT detection framework that models abnormal event relationships to reconstruct attack paths. Zhang et al. (2025) introduced MalDetectFormer, a Transformer-based model with integrated subgraph convolutions for malicious traffic detection.

Building on these insights, we propose a relation-aware graph that models intra-event attribute-value semantics, alongside a differential hypergraph module with multi-granularity clustering to capture complex, high-order behavioral dependencies across events. Unlike prior methods that focus on entity-level interactions, our method captures both local structures and global high-order semantics in EDR logs, facilitating efficient long-sequence processing and more accurate detection of stealthy threats.

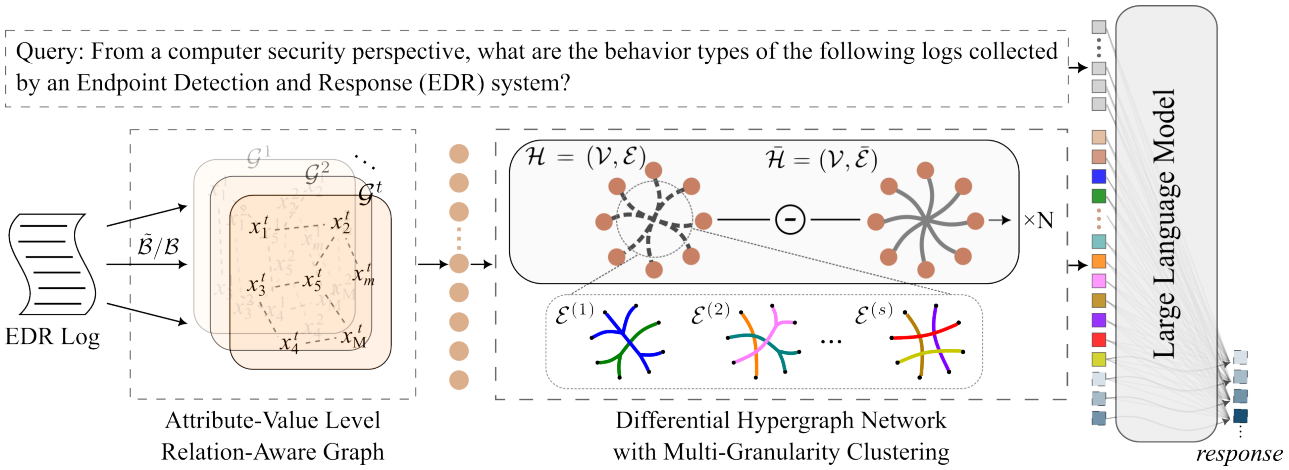

Figure 2: The HyperGLLM framework. The attribute-value level relation-aware graph encodes EDR logs into event-level representations. The differential hypergraph network performs multi-granularity clustering over event sequences, enhancing malicious semantics by capturing global hypergraph features. Finally, the hypergraph-enhanced representations are aligned with the LLM's semantic space to identify behavior through contextual reasoning.

Method

An overview of our HyperGLLM model is shown in Fig.2. Specifically, the relation-aware graph encodes interactions among attribute-value pairs to extract structured semantics from EDR logs, producing event-level representations with fine-grained information. The differential hypergraph network constructs hyperedges over temporal event sequences via multi-granularity clustering, enabling the capture of latent malicious behaviors through global hypergraph differentials. Finally, the LLM performs behavior reasoning conditioned on task-specific prompts and hypergraph-enhanced semantics. By combining structured semantics with temporal dependencies, HyperGLLM enables accurate and efficient behavior inference over long and complex EDR logs.

Attribute-Value Level Relation-Aware Graph

Each event in EDR logs comprises multiple attribute-value pairs that encode both basic properties and complex intra-event dependencies. Capturing these internal relationships is essential for a precise modeling of event semantics. Here, we introduce a relation-aware graph at the attribute-value level, modeling each event as a graph where nodes correspond to attribute-value pairs, and edges capture latent relational dependencies. Fig. 3(a) shows this modeling schematic.

Formally, let $\mathcal{L} = \{E^t\}_{t=1}^{T}$ denote an EDR log with T events. Each event E^t is represented as a set of attribute-value pairs $\{a_m^t : v_m^t\}_{m=1}^{M}$, with M denoting the total number of attributes. We concatenate each attribute-value pair into a string $a_m^t \| v_m^t$, and convert it into a numeric vector using UTF-8 encoding for compactness and universality.

$$x_m^t = \bigcup_{l=1}^{l_m^t} \tilde{\mathcal{B}}\Big(\bigcup_{l=1}^{l_m^t} \mathcal{B}(a_m^t || v_m^t)\Big), \tag{1}$$

where l_m^t denotes the length of the string. $\mathcal{B}(\cdot)$ denotes the encoding function that maps text to binary and $\tilde{\mathcal{B}}(\cdot)$ decodes each 8-bit segment into a decimal value. We then construct a graph $\mathcal{G}^t$ for each event, where $\{x_m^t\}_{m=1}^{M}$ denotes the set of node features, and the adjacency matrix $A^t = \mathbb{I}(a_m^t \neq \emptyset) \cdot \mathbb{I}(a_m^t \neq \emptyset)^\top$ is determined by the presence of attributes. $\mathbb{I}(\cdot)$ denotes the indicator function. The node update in the graph follows a weighted message passing scheme:

$$\hat{x}_m^t = \sigma\Big(\sum_{m^* \in \mathcal{N}(m)} \alpha_{m,m^*}^t W \tilde{x}_{m^*}^t\Big), \tag{2}$$

where $\sigma(\cdot)$ is an activation function, $\mathcal{N}(m)$ is the set of nodes adjacent to the m-th node, and $W \in \mathbb{R}^{d \times d}$ is a learnable weight matrix. $\tilde{x}_{m^*}^t = f_{m^*}(x_{m^*}^t)$ maps each input vector into a d-dimensional space via a linear projection layer, providing a uniform representation for attribute-value pairs of varying lengths. α_{m,m^*}^t is the attention weight between the nodes m and m^*. We parameterize the edge weights as:

$$\alpha_{m,m^*}^t = \frac{exp(w_{m,m^*}^t)}{\sum_{m^* \in \mathcal{N}(m)} exp(w_{m,m^*}^t)} A_{m,m^*}^t, \tag{3}$$

$$w_{m,m^*}^t = \frac{\sigma(W' \tilde{x}_m^t + b') \cdot \sigma(W'' \tilde{x}_{m^*}^t + b'')^\top}{\sqrt{\tilde{d}}}, \tag{4}$$

where $W', W'' \in \mathbb{R}^{\tilde{d} \times d}$ are learnable matrices, $\tilde{d}$ is the reduced dimensionality, and b', b'' are biases. This design enables the model to infer latent dependencies between attribute-value pairs without requiring predefined schemas, improving its adaptability to diverse log events. We obtain an event-level embedding via residual-enhanced pooling:

$$v^t = \frac{1}{M} \sum_{m=1}^{M} (x_m^t + \hat{x}_m^t), \tag{5}$$

By modeling attribute-value dependencies through learned relational graphs, our method generates fine-grained event-level representations, enhancing semantic expressiveness, and significantly reducing the computational burden of processing lengthy log text.

Differential Hypergraph Network with Multi-Granularity Clustering

EDR logs typically exhibit interleaved behavioral processes, where sparse malicious events are often semantically indistinguishable from benign ones. This characteristic poses a significant challenge for effective threat discrimination. To tackle this, we propose a multi-granularity differential hypergraph network, which captures high-order behavioral semantics at multiple scales and enhances malicious pattern discrimination with a global differential mechanism. Fig.3(b) shows the operational flow of the n-th layer of the differential hypergraph network.

Given the semantic representation of each event, we construct a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v^t\}_{t=1}^{T}$ represents the set of event nodes and $\mathcal{E} = \{e_k\}_{k=1}^{K}$ denotes the set of hyperedges. Each hyperedge e_k connects a subset of semantically related events, enabling the modeling of high-order dependencies. Let $H \in \mathbb{R}^{T\times K}$ be the incidence matrix, where $H_{t,k} = 1$ if node $v_t \in e_k$, and 0 otherwise. The node features are iteratively updated via a multilayer hypergraph message passing scheme (Feng et al. 2019):

$$\begin{aligned} V^{n+1} &= \mathcal{F}^{(n)}(V^{(n)}, \mathcal{E}, W^{(n)}) \\ &= \sigma(D_v^{-1/2} H W_e D_e^{-1} H^\top D_v^{-1/2} V^{(n)} W^{(n)}), \end{aligned} \tag{6}$$

where $\mathcal{F}^{(n)}$ denotes the n-th hypergraph layer, $n \in [1, N]$, and $V^{(1)} = v^{[1:T]}$. N is the total number of layers of the hypergraph network. $D_v(t,t) = \sum_k (t,k)$ and $D_e(k,k) = \sum_t (t,k)$ denote the diagonal matrices of the vertex degrees and the edge degrees, respectively. W_e is a diagonal matrix of edge weights, and $W^{(n)} \in \mathbb{R}^{d\times d}$ is the learnable projection matrix at the n-th layer.

To adaptively capture the diverse and interleaved behavioral patterns in EDR logs, we design a multi-granularity clustering strategy for hyperedge construction. Specifically, we apply k-means clustering on node embeddings using a range of cluster sizes $\mathcal{C} = \{C^{(s)}\}_{s=1}^{S}$, where $C^{(s)} = \beta s$ and β denotes the initial cluster size. K-means is chosen for its simplicity and computational efficiency. S is the maximum clustering scale. At each granularity level s, the clustering partitions the node set into disjoint groups:

$$\mathcal{P}^{(s)} = \{P_1^{(s)}, P_2^{(s)}, ..., P_{C^{(s)}}^{(s)}\}, \quad \bigcup_{k=1}^{C^{(s)}} P_k^{(s)} = V^{(n)}, \tag{7}$$

Each cluster $P_k^{(s)}$ defines a hyperedge $e_k^{(s)}$, and the hyperedge set at scale s is given by:

$$\mathcal{E}^{(s)} = \{e_k^{(s)} = P_k^{(s)} | k \in [1, C^{(s)}]\}, \quad W_e(e_k^{(s)}) = 1, \tag{8}$$

The final hypergraph aggregates all scales $\mathcal{E} = \bigcup_{s=1}^{S} \mathcal{E}^{(s)}$. The above hyperedge construction strategy leverages multi-granularity clustering tailored to the characteristics of EDR data. It enables high-order semantic modeling across multiple behavioral scales, thus improving the model's ability to capture complex event patterns.

To further isolate subtle malicious patterns embedded within predominantly benign sequences, we introduce a

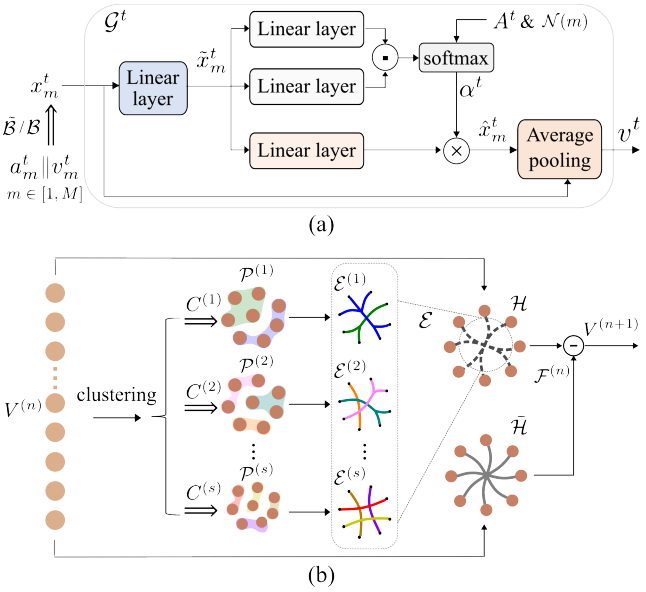

Figure 3: (a) Modeling of the t-th event in the relation-aware graph. (b) Computation flow of the n-th layer in the differential hypergraph network.

global differential mechanism. Given the sparsity of malicious events, we approximate the semantic center of benign behavior using a fully connected hyperedge $\bar{\mathcal{E}}$ that encompasses all nodes. A parallel hypergraph $\bar{\mathcal{H}} = (\mathcal{V}, \bar{\mathcal{E}})$ is constructed to serve as a baseline of benign semantics. At each layer, we compute the residual between the original and global hypergraph representations:

$$V^{(n+1)} \leftarrow V^{(n+1)} - \mathcal{F}^{(n)}(V^{(n)}, \bar{\mathcal{E}}, W^{(n)}), \tag{9}$$

This differential design progressively filters out global benign semantics, enabling the model to retain high-order behavioral semantics while enhancing the extraction of malicious behavior patterns. Finally, to obtain a compact and sequence-invariant representation, we perform average pooling over the nodes in each hyperedge $\mathcal{E}$ and incorporate residual connections from the initial input. The final hypergraph embedding is computed as:

$$Z = D_e^{-1} H^\top (V^{(N+1)} + V^{(1)}), \tag{10}$$

Optimization with LLMs

To fully leverage the contextual reasoning capabilities of LLMs, we apply a linear projection to map the hypergraph embedding $Z \in \mathbb{R}^{\sum_s C^{(s)}\times d}$ into the LLM's embedding space (Liu et al. 2023), and concatenate it with the query $\mathcal{Q}$ to form the final input sequence for behavior reasoning.

To train HyperGLLM, we adopt a two-stage curriculum optimization strategy to gradually integrate hypergraph-enhanced semantics into the LLM. Let θ_{LLM} denote the LLM parameters and θ represent the remaining parameters of HyperGLLM. The self-regression objective is defined as:

$$\mathcal{J} = -\sum_{r=1}^{R} log P(y_r | \mathcal{L}, \mathcal{Q}, y_1, ..., y_{r-1}; \theta, \theta_{LLM}), \tag{11}$$

where $y_{1:R}$ denotes the target behavior name. In the first stage, we freeze the LLM and update only θ to ensure stable semantic alignment. In the second stage, we fine-tune the entire model end-to-end, enabling the LLM to adapt to the hypergraph context and improve its reasoning capability.

Experiments

We first introduce the large-scale EDR dataset constructed for this study. Then, we evaluate our method against a series of state-of-the-art baselines and conduct ablation studies to fully assess the effectiveness of our method.

Datasets

A large-scale and diverse EDR dataset is essential for evaluating performance and improving robustness in threat detection. However, the scarcity of open-source EDR datasets (Alsaheel et al. 2021; Zhu et al. 2023), particularly those with diverse behaviors, complete behavioral chains, and fine-grained event-level records, significantly impedes the development of LLM-based solutions. To this end, we developed a program to monitor and capture real-time system activities, including process creation, network connections, file access, and registry modifications. The resulting dataset, EDR3.6B-63F, comprises over 3.6 billion events and 2 million labeled samples, covering 62 distinct malicious behavior types alongside benign activities. Specifically, the number of tokens per sample after tokenization (with over 80% of EDR samples containing more than 1 million tokens), along with the proportion and positional distribution of malicious events within each sample (average proportion below 5% and an interquartile range of positions of approximately 0.42 on a normalized [0, 1] scale), reflect the extremely long nature of EDR logs and the sparsity and interleaved pattern of malicious activities.

Implementation Details

Data Preprocessing. To mitigate the imbalance in sample size across threat families, we sample up to 2,000 instances per family for training and apply resampling strategies to ensure a more uniform distribution. For evaluation, we randomly select 24,800 representative samples from the remaining 1.9 million instances to reduce computational overhead. This subset is balanced between benign and malicious instances, with equal samples per malicious family. During training, we apply two data augmentation strategies to enhance model robustness. The first is event-level augmentation, which randomly perturbs attribute-value pairs (e.g., changing case, replacing characters with whitespace or underscores). The second is sequence-level augmentation, involving random insertion or deletion of benign events. Each strategy is applied independently with a 0.5 probability and is carefully designed to preserve the original semantics of the attack chain.

Experimental Setup. For the HyperGLLM framework, the feature dimension is set to $d = 512$. In the differential hypergraph module, the initial cluster size β is 8 and the maximum clustering scale is $S = 6$. The total number of hypergraph layers is $N = 4$. We adopt Qwen2.5-3B-Instruct as the backbone LLM for HyperGLLM. Training is performed for 1 and 8 epochs in the first and second stages, respectively, with a batch size of 32. AdamW is used as the optimizer, with a cosine annealing learning rate schedule initialized at $1e-5$. All experiments are conducted on a single node equipped with eight NVIDIA H100 (80GB) GPUs.

Evaluation Metrics. For binary detection, we report *Binary Accuracy*, *Recall*, and *False Alarm*, which respectively evaluate the correct detection of benign and malicious samples, the sensitivity to malicious behaviors, and the proportion of benign samples incorrectly identified as malicious. For multi-family detection, we use *Overall Accuracy*, which measures the proportion of correctly identified samples across all behavior families.

Comparison with State-of-the-arts

We compare our method with conventional and LLM-based baselines to assess its effectiveness. Conventional baselines include LGTF (Kumar et al. 2022), ADE (Tsai et al. 2024), DrSec (Sharif et al. 2024), and MalDetectFormer (Zhang et al. 2025). LLM-based baselines include DeepSeek-R1 (Guo et al. 2025) in the zero-shot setting, as well as Qwen2.5-3B (Yang et al. 2024), LongRoPE (Ding et al. 2024), LLaMA3-8B (Grattafiori et al. 2024), and ScamNet (Bitaab et al. 2025) in the fine-tuned setting.

Table 1 summarizes the comparative results. Conventional learning-based methods exhibit high false alarm rates and limited overall accuracy. While the LLM DeepSeek-R1 exhibits strong general NLP capabilities, its performance on EDR tasks remains suboptimal, primarily due to the lack of domain-specific training. Notably, the low false alarm rate observed in DeepSeek-R1 can be attributed to its tendency to classify all inputs as benign. Moreover, even after fine-tuning, models like Qwen2.5-3B (32K tokens) and LLaMA3-8B (8K tokens) fail to consistently outperform traditional deep learning methods (e.g., DrSec), as they are constrained by limited context windows that hinder the modeling of complete long-range dependencies.

To assess the impact of extended context, we conduct experiments with longer context windows on baseline LLMs. Specifically, we apply LongRoPE to Qwen2.5-3B (up to 1024K tokens) and fine-tune LLaMA3-8B with a 128K window on ScamNet. Although both show performance improvements, these gains come at the cost of substantial computational overhead and still lag behind the performance of our method. This reveals a key limitation of applying LLMs directly: the stealthy and sporadically distributed nature of malicious behaviors constrains the effectiveness of pure long-context modeling, impeding the capture of subtle but critical discriminative semantics.

In contrast, our method encodes fine-grained event-level semantics and captures high-order behavioral dependencies across EDR logs through a differential hypergraph. This design enables efficient compression of ultra-long inputs while preserving critical contextual signals. As a result, HyperGLLM improves both accuracy and efficiency in behavior-level reasoning by integrating LLMs with hypergraph-enhanced modeling. This demonstrates the effectiveness and practical potential of our approach.

Method	Binary Acc ↑	Recall ↑	False Alarm ↓	Overall Acc ↑
LGTF (Kumar et al. 2022)	81.75	99.01	35.50	61.71
ADE (Tsai et al. 2024)	77.10	99.05	44.84	53.11
DrSec (Sharif et al. 2024)	90.14	99.75	19.48	84.55
MalDetectFormer (Zhang et al. 2025)	69.06	99.20	61.08	38.73
DeepSeek-R1 (Guo et al. 2025)	61.50	24.25	**1.24**	50.66
Qwen2.5-3B (Yang et al. 2024)	88.70	98.98	21.57	76.00
LongRoPE (Ding et al. 2024)	92.33	99.54	14.89	86.31
LLaMA3-8B (Grattafiori et al. 2024)	84.98	98.46	28.50	60.16
ScamNet (Bitaab et al. 2025)	85.44	99.67	28.78	77.42
HyperGLLM (Ours)	**99.11**	**99.89**	1.67	**94.65**

Table 1: Comparison between our HyperGLLM framework and existing state-of-the-art methods. The first four rows correspond to conventional learning-based approaches, while the remaining methods, including ours, are LLM-based. LongRoPE and ScamNet are based on Qwen2.5-3B and Llama3-8B, respectively, with context window sizes of 1024K and 128K.

Comparison on Computational Efficiency

To evaluate the end-to-end inference efficiency of our framework, we compare GPU memory usage (GPU-MU) and Time-to-First-Token (TTFT) against LLM-based baselines. Specifically, we benchmark HyperGLLM against Qwen2.5-3B and LLaMA3-8B using EDR inputs of 32K, 64K, 128K, 512K, and 1M tokens. For each setting, we report peak metrics averaged over 100 runs, under identical hardware and software environments. As shown in Fig.4, HyperGLLM achieves substantial efficiency improvements, reducing GPU memory usage and inference latency to less than 1/15 and 1/1000 of the baselines, respectively, on a 1M-token input. Moreover, as the input length increases, the baselines exhibit a steep rise in all metrics, whereas HyperGLLM shows a significantly slower growth, indicating better stability. These results demonstrate the scalability and efficiency of our method in long-context EDR scenarios.

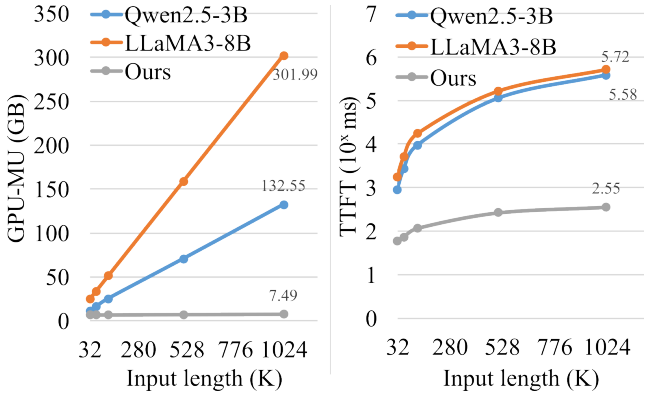

Figure 4: Comparison on computational efficiency. All methods are tested end-to-end.

Ablation Study

Impact of the Maximum Clustering Scale. In the differential hypergraph network, the maximum clustering scale S determines both the number of hyperedges and the granularity of behavioral semantics. Specifically, larger values of S enable the inclusion of finer-grained behavioral hyperedges but may also introduce redundant or noisy hyperedges. Here, we conduct experiments varying S from 3 to 7 to evaluate its effect on model performance.

Method	Binary Acc ↑	Overall Acc ↑
HyperGLLM (S=3)	94.53	87.97
HyperGLLM (S=4)	93.71	88.44
HyperGLLM (S=5)	94.18	88.60
HyperGLLM (S=6)	**99.11**	**94.65**
HyperGLLM (S=7)	94.14	89.27

Table 2: Ablation study on the maximum clustering scale.

As shown in Table 2, the overall performance of the model improves as S increases from 3 to 6, indicating that incorporating finer-grained hyperedges enhances the model's ability to capture complex behavioral relationships. However, at $S = 7$, performance deteriorates, suggesting that an excessively large clustering scale may introduce redundant information and impair the model's discriminative power. Overall, setting $S = 6$ strikes a favorable balance between inference accuracy and capturing rich contextual semantics.

Analysis of the Multi-Granularity Differential Hypergraph Network. To assess the contribution of each component in the proposed differential hypergraph network, we conduct ablation studies in a controlled setting by selectively removing key modules and observing the resulting performance changes. First, we remove the entire differential hypergraph module (w/o DHGNN) and replace it with fully connected layers to maintain parameter parity and semantic alignment with the LLM. Second, to isolate the effect of the differential structure, we construct a variant by removing the global hypergraph $\bar{\mathcal{H}}$ while retaining the base hypergraph $\mathcal{H}$, denoted as (w/o Differential). Third, to assess the necessity of the multi-granularity design, we construct single-granularity baselines using either the coarsest ($s = 3$) or the finest ($s = 7$) clustering level.

As shown in Table 3, removing the full DHGNN module results in a notable performance drop across all metrics, underscoring its essential role in modeling behavioral evolution and capturing discriminative semantic patterns within event sequences. Removing only the differential structure also significantly reduces performance, demon-

Method	Binary Acc ↑	Overall Acc ↑
w/o DHGNN	93.90	88.12
w/o Differential	94.94	89.58
single-granularity (s=3)	92.05	84.34
single-granularity (s=7)	92.71	85.29
HyperGLLM (default)	**99.11**	**94.65**

Table 3: Ablation study on the multi-granularity differential hypergraph network.

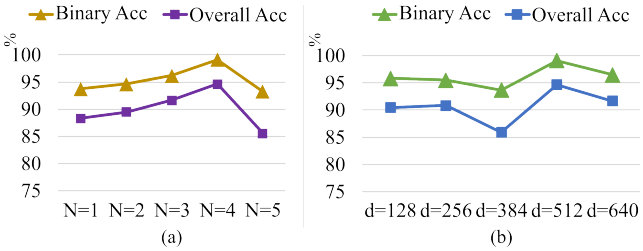

Figure 5: The impact of hyperparameters N and d.

strating the effectiveness of differential hyperedges in distinguishing malicious from benign patterns. Finally, the single-granularity variants consistently underperform the default model, confirming that the multi-granularity design provides complementary advantages for modeling hierarchical semantics and enhances the hypergraph's capacity to capture high-level behavioral patterns.

Impact of Tunable Hyperparameters N and d. Two noteworthy hyperparameters in our framework are the number of layers N and the feature dimension d. We conduct ablation studies to analyze their impact on HyperGLLM. The results are shown in Fig.5. Specifically, we vary the number of layers N from 1 to 5 to evaluate its impact on model performance. The results show a consistent performance gain as N increases, peaking at $N = 4$. Fewer layers limit the model's ability to capture high-order behavioral relationships, while excessive depth may introduce overfitting or redundant representations, ultimately impairing performance. For the feature dimension d, we evaluate values ranging from 128 to 640. The results show that performance fluctuates with increasing d due to varying representational capacity, with the best result at $d = 512$, suggesting a trade-off between expressiveness and model complexity.

Generalization to Unseen Attacks. To evaluate the model's generalization ability to unseen malicious behaviors, we construct an independent test set of 10,000 samples with a benign-to-malicious ratio of 9:1—a simplified setting, as benign samples typically dominate in real-world scenarios. All malicious families are excluded from the training set. This setup emulates real-world scenarios in which novel attacks emerge, and the model must detect malicious activity without prior exposure to those specific threats. As shown in Table 4, baseline methods achieve high recall for malicious behaviors, primarily due to their tendency to over-predict samples as malicious, rather than effectively distinguishing between benign and malicious activities. This bias is further reflected in their notably higher false alarm rates. In

Method	Binary Acc ↑	Recall ↑	False Alarm ↓
LongRoPE	86.20	97.00	15.00
ScamNet	73.70	**99.80**	29.20
HyperGLLM	**97.98**	93.90	**1.57**

Table 4: Performance comparison of different methods on unseen attacks.

Method	Binary Acc ↑	Recall ↑	False Alarm ↓
LongRoPE	69.4	25.0	**4.0**
ScamNet	58.8	96.7	64.0
HyperGLLM	**88.1**	**100.0**	19.0

Table 5: Performance comparison of different methods on the ATLASv2 dataset.

contrast, our method achieves overall superior performance, with greater accuracy in distinguishing malicious from benign samples. These results demonstrate the strong generalization and robustness of our framework and highlight its potential for detecting previously unknown threats.

Evaluation on Public Datasets. To further evaluate the generalization capability of our method on a different EDR dataset, we conduct additional experiments on the publicly available ATLASv2 dataset (Alsaheel et al. 2021), which contains host activity logs from two Windows 7 machines over a five-day period, collected from multiple audit sources. In our setting, we use only the logs recorded by Carbon Black Cloud, characterized by long event sequences and complex temporal dependencies, albeit on a small dataset. The model is trained on data from host 1 using the oversampling strategy provided in ATLASv2 to augment the log sequences, and tested on host 2 to ensure no data leakage. As shown in Table 5, LongRoPE predominantly outputs benign decisions, whereas ScamNet over-flags malicious cases, indicating limited generalization to novel datasets. In contrast, our approach achieves a 100% recall rate on malicious samples while keeping the false alarm rate on benign ones at a moderate level. This demonstrates both robust generalization to previously unseen environments and the effectiveness of our framework.

Conclusion

In this work, we present HyperGLLM, a novel detection framework that integrates hypergraph reasoning with large language models to advance threat detection in ultra-long and complex EDR logs. By modeling both low-order structural semantics and high-order behavioral dependencies through a relation-aware graph and a differential hypergraph module, HyperGLLM effectively models interleaved event sequences and long-range dependencies that underlie stealthy and evasive threats. We also construct EDR3.6B-63F, a large-scale dataset to drive research in EDR log-based threat detection. Extensive experiments show that HyperGLLM significantly outperforms state-of-the-art baselines while maintaining high inference efficiency. We believe our framework and data offer a robust and scalable foundation for advancing AI-driven endpoint security.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alsaheel, A.; Nan, Y.; Ma, S.; Yu, L.; Walkup, G.; Celik, Z. B.; Zhang, X.; and Xu, D. 2021. {ATLAS}: A sequence-based learning approach for attack investigation. In *30th USENIX security symposium (USENIX security 21)*, 3005–3022.

Bitaab, M.; Karimi, A.; Lyu, Z.; Mosallanezhad, A.; Oest, A.; Wang, R.; Bao, T.; Shoshitaishvili, Y.; and Doupé, A. 2025. ScamNet: Toward Explainable Large Language Model-Based Fraudulent Shopping Website Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 27841–27848.

Cao, T.; Huang, C.; Li, Y.; Huilin, W.; He, A.; Oo, N.; and Hooi, B. 2025. Phishagent: a robust multimodal agent for phishing webpage detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 27869–27877.

Chen, T.; Zheng, C.; Zhu, T.; Xiong, C.; Ying, J.; Yuan, Q.; Cheng, W.; and Lv, M. 2023. System-level data management for endpoint advanced persistent threat detection: Issues, challenges and trends. *Computers & Security*, 135: 103485.

Ding, Y.; Zhang, L. L.; Zhang, C.; Xu, Y.; Shang, N.; Xu, J.; Yang, F.; and Yang, M. 2024. LongRoPE: extending LLM context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, 11091–11104.

Doan, B. G.; Yang, S.; Montague, P.; De Vel, O.; Abraham, T.; Camtepe, S.; Kanhere, S. S.; Abbasnejad, E.; and Ranashinghe, D. C. 2023. Feature-space bayesian adversarial learning improved malware detector robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14783–14791.

Dong, F.; Li, S.; Jiang, P.; Li, D.; Wang, H.; Huang, L.; Xiao, X.; Chen, J.; Luo, X.; Guo, Y.; et al. 2023a. Are we there yet? an industrial viewpoint on provenance-based endpoint detection and response tools. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2396–2410.

Dong, F.; Wang, L.; Nie, X.; Shao, F.; Wang, H.; Li, D.; Luo, X.; and Xiao, X. 2023b. {DISTDET}: A {Cost-Effective} distributed cyber threat detection system. In *32nd USENIX Security Symposium (USENIX Security 23)*, 6575–6592.

Egressy, B.; Von Niederhäusern, L.; Blanuša, J.; Altman, E.; Wattenhofer, R.; and Atasu, K. 2024. Provably powerful graph neural networks for directed multigraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11838–11846.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3558–3565.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gui, J.; Nie, M.; Guo, J.; Zou, F.; Rehman, M. U.; and Hassan, W. U. 2025. A Principled Approach for Detecting APTs in Massive Networks via Multi-Stage Causal Analytics. In *Proceedings of the 44th IEEE International Conference on Computer Communications (INFOCOM)*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hassan, W. U.; Bates, A.; and Marino, D. 2020. Tactical provenance analysis for endpoint detection and response systems. In *2020 IEEE symposium on security and privacy (SP)*, 1172–1189. IEEE.

Jin, H.; Han, X.; Yang, J.; Jiang, Z.; Liu, Z.; Chang, C. Y.; Chen, H.; and Hu, X. 2024. LLM Maybe LongLM: SelfExtend LLM Context Window Without Tuning. *Proceedings of Machine Learning Research*, 22099–22114.

Kaur, R.; Gabrijelčič, D.; and Klobučar, T. 2023. Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 97: 101804.

Kokulu, F. B.; Soneji, A.; Bao, T.; Shoshitaishvili, Y.; Zhao, Z.; Doupé, A.; and Ahn, G.-J. 2019. Matched and mismatched SOCs: A qualitative study on security operations center issues. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 1955–1970.

Kumar, S.; Janet, B.; and Neelakantan, S. 2022. Identification of malware families using stacking of textural features and machine learning. *Expert Systems with Applications*, 208: 118073.

Levi, M.; Allouche, Y.; Ohayon, D.; and Puzanov, A. 2025. Cyberpal. ai: Empowering llms with expert-driven cybersecurity instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24402–24412.

Li, Y.; Xiang, Z.; Bastian, N. D.; Song, D.; and Li, B. 2024. IDS-Agent: An LLM Agent for Explainable Intrusion Detection in IoT Networks. In *NeurIPS 2024 Workshop on Open-World Agents*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, P.; Liu, J.; Fu, L.; Lu, K.; Xia, Y.; Zhang, X.; Chen, W.; Weng, H.; Ji, S.; and Wang, W. 2024. Exploring {ChatGPT's} capabilities on vulnerability management. In *33rd USENIX Security Symposium (USENIX Security 24)*, 811–828.

Macas, M.; Wu, C.; and Fuertes, W. 2024. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. *Expert Systems with Applications*, 238: 122223.

Milajerdi, S. M.; Gjomemo, R.; Eshete, B.; Sekar, R.; and Venkatakrishnan, V. 2019. Holmes: real-time apt detection through correlation of suspicious information flows. In *2019 IEEE symposium on security and privacy (SP)*, 1137–1152. IEEE.

Mohseni, S.; Mohammadi, S.; Tilwani, D.; Saxena, Y.; Ndawula, G. K.; Vema, S.; Raff, E.; and Gaur, M. 2025. Can LLMs Obfuscate Code? A Systematic Analysis of Large Language Models into Assembly Code Obfuscation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24893–24901.

Park, J.; Ji, A.; Park, M.; Rahman, M. S.; and Eun Oh, S. 2025. MalCL: Leveraging GAN-Based Generative Replay to Combat Catastrophic Forgetting in Malware Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 658–666.

Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2024. YaRN: Efficient Context Window Extension of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Raff, E.; Fleshman, W.; Zak, R.; Anderson, H. S.; Filar, B.; and McLean, M. 2021. Classifying sequences of extreme length with constant memory applied to malware detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9386–9394.

Rosenberg, I.; Shabtai, A.; Elovici, Y.; and Rokach, L. 2021. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5): 1–36.

Shao, M.; Jancheska, S.; Udeshi, M.; Dolan-Gavitt, B.; Milner, K.; Chen, B.; Yin, M.; Garg, S.; Krishnamurthy, P.; Khorrami, F.; et al. 2024. Nyu ctf bench: A scalable open-source benchmark dataset for evaluating llms in offensive security. *Advances in Neural Information Processing Systems*, 37: 57472–57498.

Sharif, M.; Datta, P.; Riddle, A.; Westfall, K.; Bates, A.; Ganti, V.; Lentzk, M.; and Ott, D. 2024. DrSec: Flexible distributed representations for efficient endpoint security. In *2024 IEEE Symposium on Security and Privacy (SP)*, 3609–3624. IEEE.

Tsai, Y.-D.; Liow, C.; Siang, Y. S.; and Lin, S.-D. 2024. Toward more generalized malicious url detection models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21628–21636.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zengy, J.; Wang, X.; Liu, J.; Chen, Y.; Liang, Z.; Chua, T.-S.; and Chua, Z. L. 2022. Shadewatcher: Recommendation-guided cyber threat analysis using system audit records. In *2022 IEEE symposium on security and privacy (SP)*, 489–506. IEEE.

Zhang, S.; Fan, Y.; Zhou, H.; and Li, B. 2025. MalDetectFormer: Leveraging Sparse SpatioTemporal Information for Effective Malicious Traffic Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22533–22541.

Zhang, Z.; Ning, H.; Shi, F.; Farha, F.; Xu, Y.; Xu, J.; Zhang, F.; and Choo, K.-K. R. 2022. Artificial intelligence in cyber security: research advances, challenges, and opportunities. *Artificial Intelligence Review*, 1–25.

Zhao, W.; Wu, J.; and Meng, Z. 2025. Apppoet: Large language model based android malware detection via multi-view prompt engineering. *Expert Systems with Applications*, 262: 125546.

Zhou, C.; Liu, Y.; Meng, W.; Tao, S.; Tian, W.; Yao, F.; Li, X.; Han, T.; Chen, B.; and Yang, H. 2025. SRDC: Semantics-based Ransomware Detection and Classification with LLM-assisted Pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28566–28574.

Zhu, J.; He, S.; He, P.; Liu, J.; and Lyu, M. R. 2023. Loghub: A large collection of system log datasets for ai-driven log analytics. In *2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE)*, 355–366. IEEE.

Zhu, T.; Ying, J.; Chen, T.; Xiong, C.; Cheng, W.; Yuan, Q.; Zheng, A.; Lv, M.; and Chen, Y. 2024. Nip in the Bud: Forecasting and Interpreting Post-exploitation Attacks in Real-time through Cyber Threat Intelligence Reports. *IEEE Transactions on Dependable and Secure Computing*, 1–18.