ShieldGPT: An LLM-based Framework for DDoS Mitigation

Tongze Wang
Tsinghua University

Xiaohui Xie*
Tsinghua University

Lei Zhang
Zhongguancun Laboratory

Chuyi Wang
Tsinghua University

Liang Zhang
Huawei Technologies Co., Ltd

Yong Cui
Tsinghua University

ABSTRACT

The constantly evolving Distributed Denial of Service (DDoS) attacks pose a significant threat to the cyber realm, which underscores the importance of DDoS mitigation as a pivotal area of research. While existing AI-driven approaches, including deep neural networks, show promise in detecting DDoS attacks, their inability to elucidate prediction rationales and provide actionable mitigation measures limits their practical utility. The advent of large language models (LLMs) offers a novel avenue to overcome these limitations. In this work, we introduce ShieldGPT, a comprehensive DDoS mitigation framework that harnesses the power of LLMs. ShieldGPT comprises four components: attack detection, traffic representation, domain-knowledge injection and role representation. To bridge the gap between the natural language processing capabilities of LLMs and the intricacies of network traffic, we develop a representation scheme that captures both global and local traffic features. Furthermore, we explore prompt engineering specific to the network domain and design two prompt templates that leverage LLMs to produce traffic-specific, comprehensible explanations and mitigation instructions. Our preliminary experiments and case studies validate the effectiveness and applicability of ShieldGPT, demonstrating its potential to enhance DDoS mitigation efforts with nuanced insights and tailored strategies.

CCS CONCEPTS

• **Networks** → **Network Security**; • **Machine Learning** → **Large Language Model**.

KEYWORDS

ShieldGPT, Distributed Denial of Service (DDoS), Large Language Model (LLM)

ACM Reference Format:
Tongze Wang, Xiaohui Xie, Lei Zhang, Chuyi Wang, Liang Zhang, and Yong Cui. 2024. ShieldGPT: An LLM-based Framework for DDoS Mitigation. In *The 8th Asia-Pacific Workshop on Networking (APNet 2024), August 3–4, 2024, Sydney, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3663408.3663424

*Corresponding Author: Xiaohui Xie (xiexiaohui@tsinghua.edu.cn)

APNet 2024, August 3–4, 2024, Sydney, Australia

ACM ISBN 979-8-4007-1758-1/24/08
https://doi.org/10.1145/3663408.3663424

1 INTRODUCTION

A Distributed Denial of Service (DDoS) attack disrupts a network, service, website, or online platform by overwhelming it with excessive internet traffic. These attacks significantly threaten the internet community, impacting critical infrastructure, public safety, and national security. A recent report [1] reveals that the global frequency of attacks in 2023 has escalated to 1.6 times that of 2022 and 1.8 times that of 2021, with the complexity of these attacks also on the rise.

The escalating prevalence and complexity of DDoS attacks underscore the urgent need for potent mitigation solutions. Numerous AI-based approaches, including machine learning [3, 6, 15] and deep learning [1, 2, 9], have been explored for DDoS attack detection. Machine learning techniques detect attacks by analyzing manually selected features to classify the data, whereas deep learning approaches autonomously extract features and perform classification. Despite promising detection capabilities, current AI-driven approaches to mitigating DDoS attacks face two key limitations: 1) limited explainability, and 2) absence of mitigation instructions, impeding their practical application. Recently, large language models (LLMs), exemplars of generative AI, have made significant strides in natural language processing (NLP), garnering global attention. Recent studies have demonstrated the application of LLMs in networking tasks, including network diagnosis [5, 19], network configuration [14], as well as network management [10, 12].

Intuitively, LLMs hold the promise of being effective in DDoS mitigation. However, two key challenges must be addressed: 1) LLMs are inherently designed for processing natural language text, so it is crucial to represent heterogeneous information in network scenarios, such as real-time binary traffic data and static domain-specific textual information, in a way that LLMs can understand. 2) While LLMs have strong general capabilities, for specific tasks, it is necessary to inform the model of its role in preventing hallucination issues and producing the desired outcomes.

To tackle these challenges, we introduce ShieldGPT, an LLM-based framework designed for DDoS mitigation, comprising four core modules: attack detection, traffic representation, domain knowledge injection and role representation. Specifically, through the detection module, raw traffic will be tagged, with labels including benign or specific types of DDoS attacks. Inspired by the way GPT-4 processes traffic data presented in either plain textual formats or structured documents, we extract both global and local characteristics from the raw traffic, which includes statistical characteristics as well as selected content from the raw traffic. Upon identification of the attack label, domain-knowledge injection is used to gather pertinent information on DDoS attacks and mitigation devices.

[1] https://e.huawei.com/en/material/networking/security/0c561b8fd2d342999cd402bcecf6d452

Leveraging traffic characteristics, tags from the detection module and domain knowledge, we carefully craft prompt templates through role representation to enhance the LLMs' ability to generate explanations for predictions and offer mitigation suggestions. We implement a prototype system using GPT-4 as our backbone LLM. Experimental results on public DDoS datasets validate our framework's proficiency in providing clear explanatory analyses and practical mitigation guidance. We position ShieldGPT as a pioneering and significant advancement towards the development of an autonomous DDoS mitigation system in the future.

In summary, we make the following contributions:

- We introduce a novel LLM-based DDoS mitigation framework that leverages LLM for in-depth attack analysis and mitigation. A prototype system has been implemented and evaluated.
- We design a representation scheme for network traffic, including both global and local characteristics, tailored to meet the input constraints of LLMs while retaining rich informational content.
- We create two specialized role-based prompt templates aimed at facilitating the generation of explanatory analyses and actionable mitigation instructions. These templates are designed to ensure the LLM's accurate task comprehension, mitigate potential hallucination issues, and produce clear, detailed outputs.

2 MOTIVATION

Existing AI-based methods have shown excellent performance in DDoS attack detection. Leveraging the advanced, pre-trained transformer-based encoder, YaTC [18], as an example, we highlight its efficacy in traffic classification through self-supervised learning, facilitating accurate DDoS detection with fine granularity. We pre-train and fine-tune YaTC based on two public datasets depicted in § 4.1. Both datasets are randomly divided into training and test sets, each comprising 50% of the data. Evaluation results are listed in Table 1. YaTC demonstrates excellent detection performance, achieving an F1 score surpassing 95% across all 14 DDoS attacks and attaining a 100% F1 score for LDAP and SYN attacks.

Table 1: Detection performance of YaTC

Dataset	Attack	# flow	Precision	Recall	F1 Score
CIC-DoS2017	Benign	115,572	0.999	1.000	0.999
	Goldeneye	443	0.991	0.988	0.989
	Hulk	656	0.995	0.989	0.992
	RUDY	538	0.992	0.974	0.983
	Slowloris	1,027	0.998	0.994	0.996
	Slowbody	155	0.973	0.954	0.964
	Slowheaders	740	0.998	0.987	0.993
	Slowread	1,103	1.000	0.992	0.996
CIC-DDoS2019	Benign	1,578	0.995	1.000	0.997
	LDAP	36,052	1.000	1.000	1.000
	MSSQL	621	0.992	0.998	0.995
	NetBIOS	157	0.993	0.987	0.990
	PortMap	341	0.964	0.941	0.952
	SYN	14,560	1.000	1.000	1.000
	UDP	15,759	0.999	0.999	0.999
	UDP-Lag	517	0.965	0.959	0.962

However, the following two limitations hinder these AI-driven models from more practical applications:

(1) **Lack of explainability**. Current models fall short in offering detailed explanations for their predictions, specifically in delineating the rationale for identifying traffic as malicious. While shallow models like decision trees or linear regression can highlight influential traffic features through feature importance analysis, these features are manually curated, demanding substantial human effort and potentially lacking comprehensiveness. This deficiency in explainability hinders the broader adoption of existing AI models, often perceived as black boxes, in industrial settings.

(2) **Lack of mitigation instructions**. Existing models, designed primarily for classification or regression, lack the capability to generate actionable mitigation instructions, such as Access Control List (ACL) configurations, for DDoS attacks. Moreover, the capabilities, configuration methods, and commands accepted by different attack mitigation devices differ significantly, posing a challenge in generating tailored mitigation instructions for each device.

These limitations render AI models efficient yet impractical and unusable in real-world scenarios. The emergence of LLMs may address these challenges, which motivates the research presented in this work.

3 SHIELDGPT

3.1 Framework Overview

ShieldGPT has two primary objectives: 1) To improve the explainability of black-box AI-based detection models, allowing network managers to better understand current threats; 2) To establish the groundwork for an autonomous DDoS mitigation system by creating actionable mitigation strategies, thus narrowing the divide between advanced detection algorithms in academia and the labor-intensive mitigation approaches common in the industry. To achieve these goals and tackle the challenges outlined in § 2, as depicted in Figure 1, ShieldGPT is structured around four core components: attack detection, traffic representation, domain-knowledge injection and role representation.

The flexible **attack detection** module serves to identify any potential attacks through a DDoS classifier employing pre-defined rules, machine learning, or deep learning techniques.

The **traffic representation** module is responsible for parsing raw traffic traces, segmenting distinct flows, and converting each flow into an appropriate textual representation suitable for processing by LLMs. The detailed transformation process is discussed in § 3.2.

The **domain-knowledge injection** module furnishes background knowledge pertaining to a specific DDoS attack and mitigation device. This aids the LLM in understanding the mechanisms of a DDoS attack and the functionalities of a mitigation device, thereby reducing hallucinations and enhancing response accuracy.

The **role representation** module is employed to define concrete roles by filling pre-defined prompt templates and soliciting responses from GPT-4 tailored to the input prompt. The prompt templates for attack explanation and mitigation are detailed in § 3.3.1 and § 3.3.2, respectively.

From a comprehensive standpoint, the workflow of ShieldGPT unfolds as follows: 1) The raw traffic underscores processing through a flow-level DDoS classifier, which categorizes each flow into benign or a specific type of DDoS attacks. 2) The raw traffic trace is

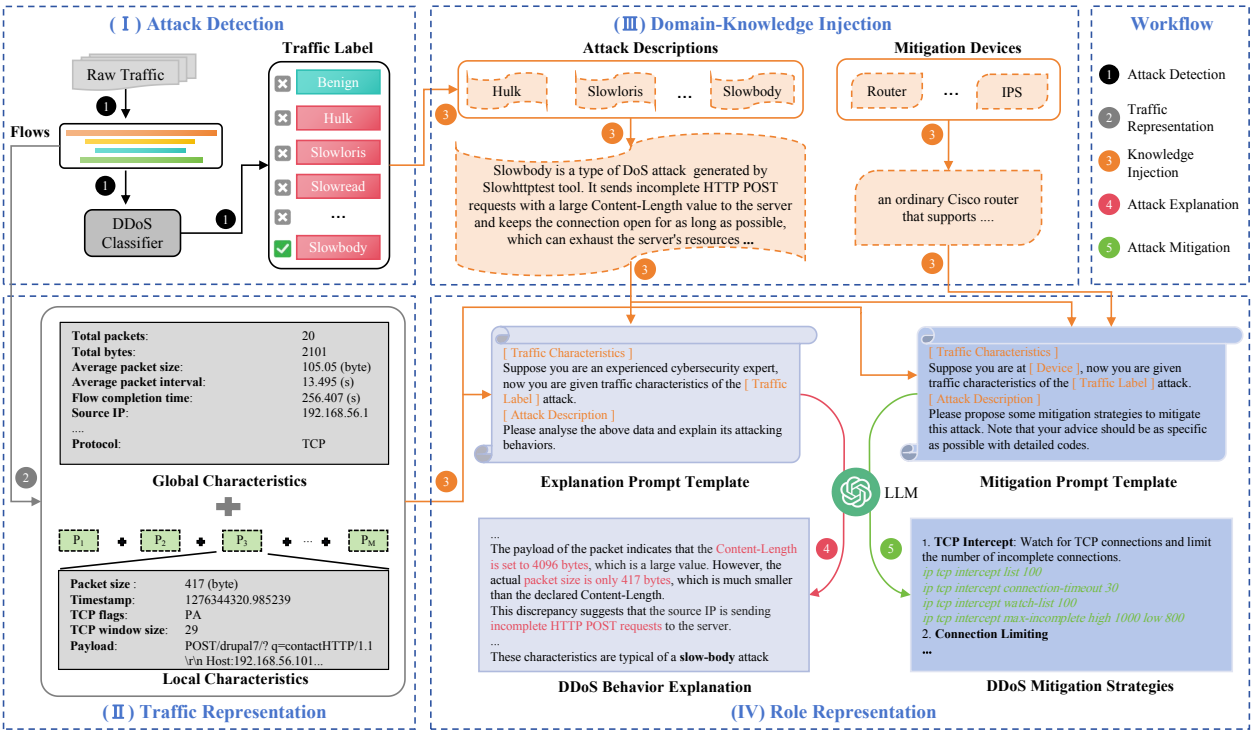

Figure 1: Overview of ShieldGPT framework

subsequently partitioned into individual flows, with each flow then being transformed into a textual representation reflecting its traffic characteristics. 3) External knowledge, comprising textual traffic characteristics, traffic labels, attack descriptions, and mitigation device specifications, is integrated into predefined prompt templates. 4) Leveraging a meticulously crafted explanation prompt, GPT-4 is employed to explain the attacking behaviors exhibited by a specific flow. 5) Similarly, through a detailed input prompt, GPT-4 generates actionable mitigation strategies for deployment on a designated device to counter the identified attack.

3.2 Traffic Representation

Given a raw traffic trace, we segment it into different flows based on their 5-tuple characteristics (Source IP, Source Port, Destination IP, Destination Port, Protocol).

To enable the analysis of traffic behaviors and the generation of mitigation instructions using Large Language Models (LLMs) like GPT-4, which are designed to process textual inputs and cannot directly handle raw traffic data, we introduce a novel method for creating a comprehensive textual representation of each traffic flow. This method integrates both global statistical characteristics and local individual attributes. This approach is inspired by the techniques utilized by GPT-4 in addressing tasks associated with the detection of attack traffic.

When processing traffic features formatted in document structures like JSON or CSV, GPT-4 autonomously generates scripts to parse these files, extracting and aggregating statistical metrics for key features, including packet size and packet intervals, thereby facilitating its decision-making process. This straightforward but effective approach underpins our endeavor to gather analogous global information systematically, as delineated in Table 2.

However, relying exclusively on statistical characteristics falls short by neglecting vital local information. To address this limitation, we undertake the serialization of individual traffic packets, encapsulating key fields shown in Table 3. Considering the limited context window of LLMs, we limit our extraction of local characteristics to the first M packets. These packets contain critical information related to connection establishment and early interactions, which are particularly vulnerable to DDoS attacks.

Table 2: Metrics for global flow characteristics

Type	Features
Categorical	Source IP, Source port, Destination IP, Destination port, Protocol
Numerical	Packet size(min, max, mean, std, sum), Packet interval(min, max, mean, std, sum), Packet count, Packet rate, Byte rate

Table 3: Metrics for local flow characteristics

Field	Interpretation
Packet size	The total number of bytes transmitted within a single packet.
Timestamp	For calculating the time interval between consecutive packet transmissions.
TCP flags	Indicators such as *SYN*, *ACK*, and *RST*, that govern the initiation, control, and termination of TCP connections.
TCP window size	The maximum number of bytes that the sender is prepared to accept.
Payload	The textual data pertaining to upper-layer protocols such as HTTP.

3.3 Role Representation

3.3.1 Attack Explanation. Leveraging prior knowledge of flow characteristics and associated attack labels, our objective is to formulate precise prompts that guide GPT-4 in analyzing attack behaviors and providing detailed explanations. We have empirically identified strategies that improve response quality.

Initially, it is crucial to elucidate the operational mechanisms of the attack. While GPT-4 possesses a foundational understanding of cybersecurity principles and comprehends the operational dynamics of various DDoS attacks, it is prone to generating inaccurate interpretations. This vulnerability stems from its incomplete internal knowledge or the use of non-standardized attack terminology, potentially leading to misunderstandings of the specific attack and introducing biases in its explanations. To tackle this challenge, we detail DDoS attack descriptions, exemplified in Figure 1, elucidate their attack mechanisms and characteristic patterns, and incorporate these descriptions into the LLM's input prompt.

Furthermore, it has been observed that variations in variable M (representing the packet number of local characteristics) within the range [5, 10] do not exert a significant influence on the response. To maintain consistency with the detection module, which is designed to process only the initial 5 packets as input, we opt to set M to 5. Ultimately, the prompt template with three slots designed to elicit detailed attack analyses is structured as follows:

The prompt template for the explanation generation

{Traffic Characterisitcs}
Please role-play as a cybersecurity expert. You've gathered traffic statistical characteristics above and the first 5 packets' original information of the {Attack Description}.
You are required to analyze the traffic and explain why it is a {Attack Name} attack step by step.
Please provide a professional yet easy-to-understand explanation for attacking behaviors.

3.3.2 Attack Mitigation. A wide range of network devices and security systems, including routers, firewalls, Intrusion Prevention Systems (IPSs), and Content Delivery Networks (CDNs), are equipped for DDoS mitigation. However, there is significant variability in the mitigation effectiveness and the commands supported across different types of devices and manufacturers.

Fortunately, LLMs are capable of generating detailed and actionable instructions if provided with relevant information about the mitigation devices. Presently, our primary focus for DDoS attack mitigation encompasses the Cisco router and Snort IPS.

Moreover, we have observed that GPT-4 exhibits a thorough understanding of the functionalities available in Cisco IOS, Snort or iptables, thereby alleviating the need for explicit specifications. Consequently, the prompt template with four slots devised for generating appropriate mitigation strategies is presented herein.

The prompt template for mitigation strategy generation

{Traffic Characterisitcs}
The above are traffic statistical characteristics and the first 5 packets' original data of the {Attack Description}.
Suppose you are at {Device}. Please propose some defense strategies to mitigate this {Attack Name} attack.
Note that your advice should be as specific as possible, and detailed configuration codes are preferred.

4 EVALUATION

4.1 Experiments Setup

We conduct experiments for ShieldGPT on two public datasets:

- *CIC-DoS2017* [7]: This dataset contains 8 different application layer DoS attack traces, totaling 4.6 GB in size. Following annotation based on destination IP address and time offset, it was observed that traces for the ddossim attack were absent. Consequently, only traces of the remaining 7 DoS attacks were utilized for training and evaluation.
- *CIC-DDoS2019* [16]: We utilize traffic traces from the testing day, comprising 7 DDoS attacks with a total size of 29 GB. Since the raw PCAP file is unlabeled, we annotate each flow based on the corresponding CSV file and discard flows with conflicting labels.

Both datasets are randomly divided into training and test sets, each comprising 50% of the data. We pre-train and fine-tune ShieldGPT using the training set and then evaluate it on the test set.

The ShieldGPT prototype is implemented using Python, where the detection module primarily relies on YaTC [18]. All experiments are conducted on a Ubuntu 22.04 server equipped with CPU of Intel(R) Xeon(R) Gold 6240C CPU @ 2.60GHz, GPU of NVIDIA A100 (40GB × 4).

To evaluate the detection performance, we employ three metrics: precision, which emphasizes the accuracy of positive predictions; recall, which underscores the capacity to capture positive instances among all actual positives; and the F1 score, which provides a balanced measure of both precision and recall.

4.2 Attack Behavior Explanation

Constructed upon the designed representation, our research findings indicate that ShieldGPT proficiently elucidates the behaviors of various DDoS attacks when provided with a well-defined description of the corresponding attack mechanism. Typically, when analyzing attack behaviors, ShieldGPT adopts a multifaceted approach, considering key factors such as packet size, flow completion time, TCP flags, packet rate, payload, etc.

Besides, we have conducted experiments with several parameters of the prompt, including the number of packets provided, scenario changes, and alterations in roles. Across diverse environments and roles, ShieldGPT consistently delivers precise and insightful analyses.

Take the Slowloris attack for example, which aims to overload a web server by sending HTTP requests with incomplete headers and then keep the connection open for as long as possible. The generated analysis would reveal a relatively low average byte rate alongside an excessively long flow completion time. Notably, ShieldGPT identifies common attack patterns, such as the sequential transmission of two consecutive '\r\n' within the HTTP header across separate requests.

In the case of the Slowbody attack, which sends incomplete HTTP POST requests with a large Content-Length value within the HTTP header and keeps the connection open for a long time, the explanation generated by ShieldGPT based on the provided flow information is outlined below:

Traffic-dependent explanation for a Slowbody attack

1. Large Content-Length Value: The fifth packet has a payload with a Content-Length of 4096, but the actual payload is much smaller. This indicates that the attacker sends less data than expected.
2. Low Packet Rate: The packet rate is 0.644 packets per second, which is relatively low. The attacker declares a large content length but sends very slowly to keep the connection open for as long as possible.
3. Incomplete HTTP POST Requests: The payload includes the header of the POST request but the body of the request is not fully sent.

The significant alignment of ShieldGPT's analyses with the unique characteristics of various attacks enhances the explainability and dependability of the detection model employed in ShieldGPT.

4.3 Attack Mitigation Strategies

Based on existing DDoS datasets, our analysis reveals that ShieldGPT is capable of generating pragmatic mitigation directives, accompanied by actionable commands, when provided with appropriately structured prompts.

Using the Cisco router as a defense device, we list common mitigation instructions generated by ShieldGPT against various DDoS attacks in Table 4. As shown, in addition to typical but stringent defense methods such as IP address blocking through ACL, less aggressive mitigation approaches like rate limiting, connection limiting, timeout adjustment, and address verification are also considered. This indicates that ShieldGPT has the potential to generate complex and hierarchical mitigation strategies to block malicious traffic while allowing legitimate traffic to pass through unaffected.

Table 4: DDoS mitigation instructions formulated for Cisco IOS by ShieldGPT

Instruction	Example Command	Interpretation
ACL	*access-list 101 deny ip <malicious-IP> any*	Block traffic from an identified malicious source IP address
Rate Limiting	*rate-limit input <rate> <normal-size> <burst-size> conform-action transmit exceed-action drop*	Constrain the number of requests a single IP can make within a certain period.
Connection Limiting	*ip http max-connections <number>*	Restrict the number of simultaneous HTTP connections from a single IP address.
TCP Intercept	*ip tcp intercept max-incomplete high <maximum> low <minimum>*	Limit the number of incomplete TCP connections.
Connection Timeout	*ip tcp synwait-time <time>*	Set a short timeout for incomplete TCP connection.
Address Verification	*ip verify unicast source reachable-via rx*	Validate the accessibility of the source IP address in an incoming packet.

For instance, in the case of a Slowheaders attack, which sends HTTP requests in pieces as slowly as possible to keep the server waiting and thus wasting its resources, the mitigation strategy is outlined below:

The mitigation strategy for a Slowheaders attack

1. **Set Connection Timeout**: Configure the router to close the connection if it does not receive the full HTTP header within a certain time limit.
ip http timeout-policy idle 60 life 86400 requests 10000
2. **Limit the Number of Connections**: Limit the number of simultaneous connections from a single IP address.
ip http max-connections 100

In more detail, the command *ip http timeout-policy idle 60 life 86400 requests 10000* configures the HTTP server to close idle connections after 60 seconds of inacitvity, terminate any connection after it has been open for 24 hours, and close any connection after it has processed 10,000 requests. By configuring *ip http max-connections 100*, the router is instructed to allow up to 100 HTTP connections simultaneously.

Futhermore, to counteract a Hulk attack, which endeavors to exhaust server resources by generating a high volume of unique and seemingly legitimate HTTP requests, the following configuration will be devised for a Snort IPS:

The mitigation strategy for a Hulk attack

1. **Rate Limiting**: Limit the number of requests that a single IP address can make in a certain period.
rate_filter track by_src, count 100, seconds 60, new_action drop, timeout 300
2. **IP Blacklisting**: If a source IP is identified as a Hulk attacker, it can be added to a blacklist to drop all its future packets.
ipvar BLACKLIST [192.168.56.1]
drop tcp $BLACKLIST any -> any any (msg:"Blacklisted IP Detected"; sid:1000002; rev:1;)

The first rule will discard all packets from an IP address if it sends more than 100 packets in 60 seconds, and this rule will remain in effect for 300 seconds. The second rule will block all TCP packets from the blacklisted IP address.

To mitigate a Slowbody attack on a Linux server using *iptables*, ShieldGPT can generate a variety of commands, including those for connection limiting and timeout filtering, as illustrated below:

The mitigation strategy for a Slowbody attack

1. **Limit the number of connections per IP**: The following command will limit the number of connections to 10 per client IP.
iptables -A INPUT -p tcp –syn –dport 80 -m connlimit –connlimit-above 10 -j REJECT
2. **Set the connection timeout**: The following command will set the connection timeout to 30 seconds to allow the server to close idle connections more quickly.
echo 30 > /proc/sys/net/ipv4/tcp_fin_timeout

5 RELATED WORK

The integration of Large Language Models (LLMs) into networking, particularly for enhancing network security, is garnering significant interest.

For traffic classification tasks, transformer-based models, e.g., BERT, GPT, and T5, demonstrate their promising performance. For instance, Lin et al. [11] propose ET-BERT for encrypted traffic classification, which pre-trains deep contextualized datagram-level representation from large-scale unlabeled data. To better encode and represent traffic data, the NetGPT [13] and Lens [17] have been proposed based on pre-training and fine-tuning the GPT-2 and T5 large models, respectively. These models employ network-specific pre-training tasks, demonstrating efficacy in downstream applications such as traffic classification and attack detection. Notably, SecurityBERT [4] has been introduced for cyber threat detection within IoT networks.

Beyond traffic classification, efforts are being made to harness LLMs for generating explanations in Network Intrusion Detection Systems (NIDS). Ziems et al. [20] propose LLM-DTE which explores the use of LLMs to provide explanations and additional background knowledge for decisions made by tree-based NIDS. They also conduct human evaluation studies to show the correlation between LLM-generated explanations and human understanding. Additionally, ChatIDS [8] is designed to explain alerts from NIDSs to non-experts. Specifically, by sending anonymized alerts to ChatGPT, ChatIDS can intuitively explain the alert and suggest meaningful countermeasures for cyber threats.

Current research demonstrates the potential of leveraging LLMs in network security, particularly for DDoS mitigation efforts. Despite these advancements, the absence of comprehensive explanations for attack classification and the need for actionable recommendations highlight the importance of continued exploration in this area.

6 CONCLUSION AND FUTURE WORK

In this paper, we present ShieldGPT, an innovative DDoS mitigation framework leveraging large language models (LLMs). We develop a specialized traffic representation scheme and tailored prompt templates to optimize LLMs for DDoS mitigation tasks. Initial experiments demonstrate ShieldGPT's superior capability in providing detailed explanatory analyses and suggesting effective mitigation strategies. We believe that ShieldGPT represents a meaningful starting point for applying LLMs to the DDoS mitigation area.

Our prototype and early findings highlight the feasibility of creating an autonomous system for DDoS mitigation. This research opens up several avenues for further investigation:

Validity and Safety. While ShieldGPT exhibits proficiency in generating tailored mitigation instructions for various DDoS attacks, further checks are necessary given that these instructions are intended for deployment on actual network devices. First, we need to verify that the generated commands are syntactically and semantically correct, as this is a prerequisite for effective attack mitigation. Second, it is essential to implement safety checks to prevent unexpected risks, such as improper route configuration causing congestion or loops, and unauthorized commands leading to data breaches or service disruptions. Consequently, establishing a robust validation mechanism is critical for future research to ensure the validity and safety of automated mitigation strategies.

Automatic Execution. Currently, ShieldGPT generates text-based mitigation instructions necessitating manual implementation by network administrators. For a truly autonomous DDoS defense system, these instructions must be automatically executed via mature technical stacks and extensive application programming interfaces (APIs). This advancement will enable rapid response to threats, significantly improving cybersecurity. Achieving this objective requires concerted efforts in both research and industry sectors.

Broader Applications. In ShieldGPT, we construct corresponding prompt templates for generating desired outputs through role-based representations. This approach can be generalized to other network tasks, such as generating diagnostic analysis in network diagnosis or generating control commands in network management.

ACKNOWLEDGMENTS

We would like to express our gratitude to the anonymous reviewers for their insightful and constructive suggestions. This work is supported by National Key R&D Program of China (No. 2023YFF0717602).

REFERENCES

[1] Gabriel Chukwunonso Amaizu, Cosmas Ifeanyi Nwakanma, Sanjay Bhardwaj, Jae-Min Lee, and Dong-Seong Kim. 2021. Composite and efficient DDoS attack detection framework for B5G networks. *Computer Networks* 188 (2021), 107871.

[2] Abdullah Emir Cil, Kazim Yildiz, and Ali Buldu. 2021. Detection of DDoS attacks with feed forward based deep neural network model. *Expert Systems with Applications* 169 (2021), 114520.

[3] Rohan Doshi, Noah Apthorpe, and Nick Feamster. 2018. Machine learning ddos detection for consumer internet of things devices. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 29–35.

[4] Mohamed Amine Ferrag, Mthandazo Ndhlovu, Norbert Tihanyi, Lucas C Cordeiro, Merouane Debbah, Thierry Lestable, and Narinderjit Singh Thandi. 2024. Revolutionizing Cyber Threat Detection with Large Language Models: A privacy-preserving BERT-based Lightweight Model for IoT/IIoT Devices. *IEEE Access* (2024).

[5] Pouya Hamadanian, Behnaz Arzani, Sadjad Fouladi, Siva Kesava Reddy Kakarla, Rodrigo Fonseca, Denizcan Billor, Ahmad Cheema, Edet Nkposong, and Ranveer Chandra. 2023. A Holistic View of AI-driven Network Incident Management. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. 180–188.

[6] Mohamed Idhammad, Karim Afdel, and Mustapha Belouch. 2018. Semi-supervised machine learning approach for DDoS detection. *Applied Intelligence* 48 (2018), 3193–3208.

[7] Hossein Hadian Jazi, Hugo Gonzalez, Natalia Stakhanova, and Ali A Ghorbani. 2017. Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling. *Computer Networks* 121 (2017), 25–36.

[8] Victor Jüttner, Martin Grimmer, and Erik Buchmann. 2023. Chatids: Explainable cybersecurity using generative ai. *arXiv preprint arXiv:2306.14504* (2023).

[9] Ömer Kasim. 2020. An efficient and robust deep learning based network anomaly detection against distributed denial of service attacks. *Computer Networks* 180 (2020), 107390.

[10] Manikanta Kotaru. 2023. Adapting Foundation Models for Operator Data Analytics. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. 172–179.

[11] Xinjie Lin, Gang Xiong, Gaopeng Gou, Zhen Li, Junzheng Shi, and Jing Yu. 2022. Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In *Proceedings of the ACM Web Conference 2022*. 633–642.

[12] Sathiya Kumaran Mani, Yajie Zhou, Kevin Hsieh, Santiago Segarra, Trevor Eberl, Eliran Azulai, Ido Frizler, Ranveer Chandra, and Srikanth Kandula. 2023. Enhancing network management using code generated by large language models. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. 196–204.

[13] Xuying Meng, Chungang Lin, Yequan Wang, and Yujun Zhang. 2023. Netgpt: Generative pretrained transformer for network traffic. *arXiv preprint arXiv:2304.09513* (2023).

[14] Rajdeep Mondal, Alan Tang, Ryan Beckett, Todd Millstein, and George Varghese. 2023. What do LLMs need to Synthesize Correct Router Configurations?. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. 189–195.

[15] Sagar Pande, Aditya Khamparia, Deepak Gupta, and Dang NH Thanh. 2021. DDOS detection using machine learning technique. In *Recent Studies on Computational Intelligence: Doctoral Symposium on Computational Intelligence (DoSCI 2020)*. Springer, 59–68.

[16] Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali A Ghorbani. 2019. Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In *2019 international carnahan conference on security technology (ICCST)*. IEEE, 1–8.

[17] Qineng Wang, Chen Qian, Xiaochang Li, Ziyu Yao, and Huajie Shao. 2024. Lens: A Foundation Model for Network Traffic in Cybersecurity. *arXiv e-prints* (2024), arXiv–2402.

[18] Ruijie Zhao, Mingwei Zhan, Xianwen Deng, Yanhao Wang, Yijun Wang, Guan Gui, and Zhi Xue. 2023. Yet another traffic classifier: A masked autoencoder based traffic transformer with multi-level flow representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5420–5427.

[19] Yajie Zhou, Nengneng Yu, and Zaoxing Liu. 2023. Towards Interactive Research Agents for Internet Incident Investigation. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*. 33–40.

[20] Noah Ziems, Gang Liu, John Flanagan, and Meng Jiang. 2023. Explaining tree model decisions in natural language for network intrusion detection. *arXiv preprint arXiv:2310.19658* (2023).