ChatIDS: Explainable Cybersecurity Using Generative AI

Victor Jüttner*, Martin Grimmer†, Erik Buchmann‡
Dept. of Computer Science, Leipzig University, Germany
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Germany
Email: *juettner@informatik.uni-leipzig.de,
†grimmer@informatik.uni-leipzig.de,
‡buchmann@informatik.uni-leipzig.de

***Abstract*—Intrusion Detection Systems (IDS) are a proven approach to secure networks. However, in a privately used network, it is difficult for users without cybersecurity expertise to understand IDS alerts, and to respond in time with adequate measures. This puts the security of home networks, smart home installations, home-office workers, etc. at risk, even if an IDS is correctly installed and configured. In this work, we propose ChatIDS, our approach to explain IDS alerts to non-experts by using large language models. We evaluate the feasibility of ChatIDS by using ChatGPT, and we identify open research issues with the help of interdisciplinary experts in artificial intelligence. Our results show that ChatIDS has the potential to increase network security by proposing meaningful security measures in an intuitive language from IDS alerts. Nevertheless, some potential issues in areas such as trust, privacy, ethics, etc. need to be resolved, before ChatIDS might be put into practice.**

***Index Terms*—Intrusion Detection; ChatGPT; Networks**

I. Introduction

In recent years, privately used networks have come into the focus of cyberattacks. Reasons for this include the increased use of home-office working models [1], a shift to private areas during pandemics [2] or the proliferation of smart-home devices [3]. Intrusion Detection Systems (IDS) are a well-established approach to detect and fight cyberattacks [4], [5]. IDS scan the network and/or network appliances and send alerts for suspicious network activity.

In industry, business and government, IDSs are an important line of defense in the cybersecurity infrastructure. To this end, these sectors employ well-trained cybersecurity experts to configure, manage and maintain IDS, continuously improve the IDS rule-set, distinguish false alarms from real attacks, and design, prioritize and implement appropriate countermeasures. It is feasible to pre-configure a network-based IDS for home networks [6]. However, without a solid background in cybersecurity, it is difficult for a home user to interpret IDS alerts such as "`MALWARE-CNC Harakit botnet traffic`", distinguish false alerts from real attacks, and devise appropriate and timely countermeasures. Static sets of explanations for well-known cyberattacks [7] do not solve this problem.

In this paper, we describe our work in progress on ChatIDS, our approach to let a large language model (LLM) – a generative artificial intelligence approach – explain IDS alerts and suggest countermeasures in an intuitive, non-technical way to users without cybersecurity knowledge. ChatIDS sends anonymized IDS alerts to a LLM, and allows the user to ask questions if the generated texts are not yet understandable enough. In particular, this paper makes four contributions:

- We specify the requirements for an approach that increases the network security in privately used networks by explaining the alerts of an IDS to a non-expert.
- We describe ChatIDS, our approach to let ChatGPT [8] explain alerts from Snort [9], Suricata [10] and Zeek [11]. The explanations include cybersecurity measures and hints on why/when the measures should be implemented.
- We evaluate the feasibility of this approach using a small series of experiments with typical IDS alerts.
- To explore ChatIDS' design space, we had interdisciplinary AI experts put together issues that must be researched, before ChatIDS can go into practice.

This paper is structured as follows: Section II introduces related work. In Section III, we outline ChatIDS, our approach to explain IDS messages to non-experts. Section IV describes a number of experiments to prove feasibility, and Section V contains open issues for interdisciplinary research.

II. Related Work

In this section, we introduce related work on network security approaches and generative AI models.

A. Network Security

Intrusion Detection Systems (IDS) monitor a system for unauthorized or suspicious activity. IDS can be distinguished by system type and detection type. The system type can be *host-based*, to control a single device, or *network-based* to control a network. Detection types can be *anomaly-detection*, which detect activities that significantly differ from the regular usage or *misuse-detection*, which uses signature rules to match known intrusions [12]. Popular examples for rule-based network-based IDS are Snort, Suricata and Zeek. To use these IDS it first needs a rule-set. Popular predefined rule-sets for networks are snort3-community-rules [13], suricata-rules [14], Yara [15] and Sigma [16].

B. Generative AI

Generative modeling strives to create models capable of creating new data, like sound, text or images that are similar

 ISBN: 978-1-68558-092-6

to the data the model was trained on [17]. Popular examples for generative models are WaveNet [18] that can generate speech and music, Pix2Pix that can transform images into different styles [19] or GPT-3, a large language model (LLM), that allows for the generation of human like text [20]. Another example for a LLM is ChatGPT [8]. Like a chatbot, ChatGPT engages in a conversational manner and can generate detailed responses to questions. Bard [21] follows a similar approach. There are generative models trained for cybersecurity problems like Microsoft Security Copilot [22] but these are aimed at experts and therefore not suitable for our purpose.

ChatGPT's reliability varies across domains, it shows high levels of accuracy in recreation and technology domains but struggles with science and law. Problems that reduce the accuracy of ChatGPT are false information, biases and hallucinations [23]. ChatGPT and LLMs in general are capable of generating text that appears natural and to be grounded in the real context, but is unfaithful and nonsensical. This is called *hallucinated text* and much like psychological hallucinations, they can be difficult to distinguish from real perception [24].

Prompts are the input for a generative model, they can be a text or image that give the model instructions for the requested output. Prompts provide an intuitive way to engage with generative models [25]. For image generation a prompt could be a different image or a text description. For LLMs a prompt is a text that provides context for the desired output e.g., a question or a command to summarize information.

Prompt Engineering deals with optimizing prompts to achieve better responses from LLMs. For recurring problems design patterns can be used to form prompts and optimize the output, analogous to software patterns [26]. For example, the *Persona Pattern* lets the LLM assume a certain role. This can help if the LLM should respond in a special way. If the output must follow a structure, a *template* can be given in the prompt. The *Context Manager* Pattern enables the user to provide or remove context from a prompt.

III. ChatIDS: Explainable Security

We aim at integrating a network-based IDS in privately used networks, to protect the network against cyberattacks from the Internet. For this purpose, we distinguish two roles:

An ***expert*** has the cybersecurity expertise necessary to operate and maintain an IDS, to understand its alarms, and respond to alarms with appropriate and timely actions.

A ***user*** lacks this type of expertise. A user may follow manuals written without technical vocabulary. It is difficult for a user to figure out if an IDS alert is from a real attack or due to false detection of the IDS, and to act accordingly.

An IDS [6] can be preconfigured for home networks, and integrated into a security process [27]. However, without knowledge of cybersecurity the user is left with only three possible actions: (a) do nothing, (b) turn off the device, or (c) ask an expert for help. Our ChatIDS approach strives to provide intuitive and understandable explanations of IDS alerts to give users a wider range of appropriate security measures. Therefore, ChatIDS must meet three requirements:

R1: (Errors) The user must assess the probability that the IDS has sent a false alert. For example, the IDS might have detected by mistake an attack that is impossible on the device.

R2: (Urgency) The user must assess the urgency of the alert, i.e., if it calls for immediate action, or not.

R3: (Actions) The user must identify appropriate measures, e.g., to execute a factory reset and install a security patch.

To explore the solution space for a generative AI approach fulfils these requirements for IDS, we use a constructive research method. In particular, we (a) model ChatIDS, we use it (b) to evaluate its technical feasibility, and (c) to discuss potential issues with interdisciplinary AI experts.

A. Our ChatIDS Approach

The information flow of ChatIDS is illustrated in Figure 1.

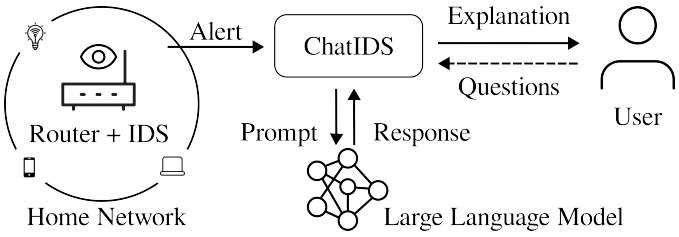

Fig. 1. Information Flow of ChatIDS

A **network-based IDS component** inspects the network packets that pass a router for suspicious traffic, and generates alerts. The IDS should be a signature-based IDS so that its alert messages are specific enough for the LLM.

The **LLM component** contains a large language model that is responsible for translating the alerts from the IDS into a language a non-expert can understand. Furthermore, this component can be used in an interactive way: If the users do not understand the explanation or the suggested measures, they can ask for details. Similarly, to the IDS the LLM is also an external component to ChatIDS.

The **ChatIDS component** is the core of our approach. ChatIDS accepts alerts from the IDS component, sends them to the LLM component for a translation into an intuitive explanation, and presents a user interface with the explanations to the user. If the user requires further support, they can use the interface to send follow up questions to the LLM. To translate alerts into intuitive explanations, the ChatIDS component contains pre-defined templates for LLM prompts.

For privacy reasons, the alerts are anonymized in three ways, before being sent to the LLM component: First, ChatIDS removes any device identifiers or network information from the alert. Second, ChatIDS sends the anonymized alert together with a set of dummy alerts to the LLM component, so that this component does not learn the real alert with certainty. The explanations from the LLM component are stored in a cache, so that the same explanation must not be requested twice.

IV. Experimental Evaluation

Since this paper contains work in progress, we exemplarily evaluate ChatIDS with selected use cases.

 ISBN: 978-1-68558-092-6

TABLE I
EVALUATION OF ALERTS AND RESPONSES

Alert	Corr.	Desc.	Cons.	Meas.	Urg.	Int.
MALWARE-CNC Harakit botnet traffic	✓	✓	✓	x	x	x
SERVER-WEBAPP NetGear router default password login attempt admin/password	✓	✓	✓	x	✓	x
SURICATA MQTT unassigned message type (0 or >15)	✓	✓	✓	✓	x	✓
SURICATA HTTP Response abnormal chunked for transfer-encoding	✓	x	✓	✓	✓	x
Mirai Botnet TR-069 Worm - Generic Architecture	✓	✓	✓	x	✓	x
Linux.IotReaper	✓	✓	✓	x	✓	✓
Identifies IPs performing DNS lookups associated with common Tor proxies.	✓	x	✓	x	✓	x
Detects remote task creation via at.exe or API interacting with ATSVC namedpipe	✓	✓	✓	x	✓	x

A. Experimental Setup

In line with Figure 1, we assume a home network with several smart-home devices. A router connects the network to the Internet and can observe any network packets. We assume a Philips Hue Bridge [28] is attacked. To implement the network-based IDS they are installed on the router and execute either the Snort [13], Suricata [14], Yara [15] or Sigma [16] ruleset. From each IDS implementation, we experiment with two alerts, as shown in the first column of Table I. The alerts are classified as important, and a user intervention is required.

The Intrusion Detection System in a home network has detected an intrusion and sent out the alert [ALERT MSG]. Your job is to inform [USER] about the alert in a warning message. You're in the role of a cybersecurity expert that interprets the alert and explains the alert in a warning message to [USER]. Your goal is to inform Jon about the intrusion in a way he understands and motivate the user to take steps to stop the intrusion. [USER] has no cybersecurity expertise and won't understand technical instructions, you need to provide clear, easy and non-technical instructions to follow. Don't use technical terms like "two-factor-authentication", "Intrusion Detection System", "intrusion" or "unassigned message", use simple non-technical terms instead. Don't use the term "Intrusion Detection System". Your explanation and instructions have to align with how [USER] thinks about cyberattacks. The Smart Home consists of several Smart Home Devices, the Intrusion Detection System has detected the intrusion on [DEVICE]. The warning message has to follow this order: Explain the intrusion, explain the potential consequences for the user if he won't comply with the warning message and give instructions on how to stop the intrusion in an itemized list.

Fig. 2. Template for a ChatGPT Prompt

We realized ChatIDS using ChatGPT (gpt-3.5-turbo) [8]. To generate an explanation, ChatIDS embeds each alert into a ChatGPT prompt, as shown in Figure 2. This prompt implements the patterns "Template", "Persona" and "Context Manager", as explained in Section II. Since the training data for ChatGPT comes from the Internet and much of the text on the Internet deals with security issues, we do not expect ChatGPT to produce hallucinations.

B. Results

Figure 3 shows an example of the output produced by ChatGPT. For a better understanding, the figure shows the non-anonymized output that includes names and devices. We have tested the prompts from the first column of Table I, one by one, and we have evaluated the output of ChatGPT according to our three requirements *R1: (Errors)*, *R2: (Urgency)* and *R3: (Actions)*. To evaluate R1, we test how well the generated explanation describes the problem (Column "Desc." in Table I),

Dear Jon,

We have detected an unauthorized access attempt to your Philips Hue Bridge, which is a device in your Smart Home. This access attempt was classified as "MALWARE-CNC Harakit botnet traffic" by our system, which means that a group of computers infected with malware is trying to control your device without your permission.

If you don't take any action, your Smart Home devices, including your Philips Hue Bridge, could be used to conduct malicious actions without your knowledge. For example, attackers could use your devices to perform Distributed Denial of Service (DDoS) attacks, steal your personal data or control your devices to cause harm.

To stop the intrusion and protect your Smart Home, please follow these simple instructions:

1. Isolate the Philips Hue Bridge by disconnecting it from the internet.
2. Reset the Philips Hue Bridge to its default factory settings.
3. Configure the Philips Hue Bridge with a new username and password.
4. Check for any suspicious activity in your other Smart Home Devices.

Please don't hesitate to contact us in case you need further assistance or have any questions.

Best regards, your ChatIDS

Fig. 3. Example for an Explanation

and how intuitively understandable it is (Column "Int."). For R2, we assess the description of the consequences of ignoring the alert (Column "Cons.") and how urgent the description is (Column "Urg."). For R3 we search for a meaningful description of countermeasures (Column "Meas."). We also evaluate the correctness of the explanation (Column "Corr.").

As expected, each generated explanation was correct. Furthermore, the table shows that ChatGPT was able to explain the consequences of not taking actions in any case. In the majority of cases, ChatGPT produced a good description of the security issue that has caused the alert, and transported a sense of urgency. The description of countermeasures and the use of non-intuitive terminology leave room for improvement.

V. OPEN RESEARCH ISSUES

As a part of our ongoing work, we have conducted a pre-study with experts from the Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig to compile open issues for interdisciplinary research. Our AI experts cover the topics applications, cybersecurity, ethics, jurisprudence and privacy. We presented our ChatIDS approach, asked for potential problems, and consolidated the answers:

Security: ChatIDS potentially increases network security, compared to a scenario where a non-expert is left alone with the alert. However, an external LLM can be a new attack

 ISBN: 978-1-68558-092-6

surface, and incorrect or incomprehensible explanations might lead to inappropriate actions.

Privacy: With ChatIDS, the LLM learns that a cyberattack may have occurred on a particular network. Anonymizing device IDs and sending dummy alerts still allows the LLM to infer some information, e.g., if none of the (dummy) alerts sent to the LLM is possible for a particular type of device.

Compliance: ChatIDS has an impact on cybersecurity. However, it is unclear yet, how to conduct a risk analysis on LLMs and on components building upon these, how to evaluate and mitigate associated risks, and to integrate ChatIDS into security frameworks such as the Common Criteria [29].

Jurisprudence: If an alert is not explained well enough, the network could be successfully attacked. Conversely, ChatIDS could convince the user to take action upon false alerts. This creates legal issues. Do special liabilities exist, e.g. from user expectations into a superior AI? How to prove that a harm was caused by a misconducting or negligent AI engineer?

Trust: Users might have a non-rational view on AI approaches, and could fear that a persuasive, non-human intelligence plots against their interests. Conversely, if a user trusts ChatIDS too much, false alerts might result in false actions.

Ethics: ChatIDS could provide explanations that are not only convincing, but manipulative, even if this is in the interest of the user. This raises ethical and moral questions. How drastic can explanations be formulated to induce them to take action (which may even be harmful due to a false positive)? At what point does this limit the autonomy of the user?

VI. Conclusion and Future Work

Comprehensibility is important for any security approach in privately used networks. This paper outlines our work in progress on ChatIDS, our approach to explain alerts from an intrusion detection system to non-experts.

ChatIDS sends anonymized alerts to ChatGPT, a large language model, to explain the alert in an intuitive way and suggest meaningful countermeasures for cyberattacks. Our experiments show that ChatIDS can be implemented easily, although more work is needed on prompt engineering to ensure intuitive explanations in the first attempt. Furthermore it needs to be analyzed if the anonymization of the data could remove relevant context or affect the report. It is difficult to measure if ChatIDS actually increases network security, because this depends on the user. Our interdisciplinary experts have provided valuable insights. In the future, we will improve ChatIDS regarding security and privacy, and consider interdisciplinary aspects such as compliance, ethics and trust.

Acknowledgement

We would like to thank Prof. Dr. Birte Platow, Dr. Hermann Diebel-Fischer, and Prof. Dr. Johannes Eichenhofer for their valuable contributions on ethical and legal questions.

References

[1] N. Vakakis *et al.*, "Cybersecurity in smes: The smart-home/office use case," in *2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2019, pp. 1–7.

[2] R. O. Andrade, I. Ortiz-Garcés, and M. Cazares, "Cybersecurity attacks on smart home during covid-19 pandemic," in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 2020, pp. 398–404.

[3] Y. Shah and S. Sengupta, "A survey on classification of cyber-attacks on iot and iiot devices," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE, 2020, pp. 0406–0413.

[4] I. Butun, S. Morgera, and R. Sankar, "A survey of intrusion detection systems in wireless sensor networks," *IEEE Communications Surveys & Tutorials*, vol. PP, pp. 266 – 282, 05 2013.

[5] R. Mitchell and I.-R. Chen, "A survey of intrusion detection techniques for cyber-physical systems," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, pp. 1–29, 2014.

[6] C. Haar and E. Buchmann, "Securing smart homes using intrusion detection systems," in *Proceedings of the 14th International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'20)*, 2020.

[7] P. Datta *et al.*, "A user-centric threat model and repository for cyber attacks," in *37th ACM/SIGAPP Symposium on Applied Computing*. Association for Computing Machinery, 2022, p. 1341–1346.

[8] Introducing ChatGPT. OpenAI. (Accessed on: 8.6.2023). [Online]. Available: https://openai.com/blog/chatgpt

[9] What is Snort? Cisco. (Accessed on: 8.6.2023). [Online]. Available: https://www.snort.org/

[10] Suricata. Open Information Security Foundation. (Accessed on: 8.6.2023). [Online]. Available: https://suricata.io/

[11] The Zeek Project. An open source network security monitoring tool. (Accessed on: 8.6.2023). [Online]. Available: https://zeek.org/

[12] T. S. Sobh, "Wired and wireless intrusion detection system: Classifications, good characteristics and state-of-the-art," *Computer Standards & Interfaces*, vol. 28, no. 6, pp. 670–694, 2006.

[13] Snort community rules. Cisco. (Accessed on: 8.6.2023). [Online]. Available: https://www.snort.org/downloads/

[14] Open Information Security Foundation. (Accessed on: 8.6.2023). [Online]. Available: https://github.com/OISF/suricata/tree/master/rules

[15] Yara-rules. YaraRules Project. (Accessed on: 8.6.2023). [Online]. Available: https://github.com/Yara-Rules/rules

[16] Sigma rules. SigmaHQ. (Accessed on: 8.6.2023). [Online]. Available: https://github.com/SigmaHQ/sigma/tree/master/rules

[17] A. Lamb, "A brief introduction to generative models," *Computing Research Repository*, vol. abs/2103.00265, 2021.

[18] Wavenet. Alphabet Inc. (Accessed on: 8.6.2023). [Online]. Available: https://www.deepmind.com/research/highlighted-research/wavenet

[19] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Computing Research Repository*, vol. abs/1611.07004, 2016.

[20] GPT-3 powers the next generation of apps. OpenAI. (Accessed on: 8.6.2023). [Online]. Available: https://openai.com/blog/gpt-3-apps

[21] Meet Bard. Google Ireland Limited. (Accessed on: 8.6.2023). [Online]. Available: https://bard.google.com/

[22] Introducing Microsoft Security Copilot. Microsoft. (Accessed on: 9.8.2023). [Online]. Available: https://www.microsoft.com

[23] X. Shen, Z. J. Chen, M. Backes, and Y. Zhang, "In ChatGPT we trust? measuring and characterizing the reliability of ChatGPT," *Computing Research Repository*, vol. abs/2304.08979, 2023.

[24] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, mar 2023.

[25] Y. Zhou *et al.*, "Large language models are human-level prompt engineers," *Computing Research Repository*, vol. abs/2211.01910, 2022.

[26] J. White *et al.*, "A prompt pattern catalog to enhance prompt engineering with ChatGPT," 2023.

[27] C. Haar and E. Buchmann, "It-security compliance for home offices," in *Proceedings of the 15th International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'21)*, 2021.

[28] Philips Hue Bridge. Signify Holding. (Accessed on: 8.6.2023). [Online]. Available: https://www.philips-hue.com

[29] Management Committee of the CC Recognition Arrangement, "CC:2022 Release 1," https://www.commoncriteriaportal.org/cc/, 2022.

 ISBN: 978-1-68558-092-6