# ChatIDS: Advancing 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*—An intrusion detection system (IDS) is a proven approach to securing networks. Network-based IDS solutions are typically installed on routers or Internet gateways. They can inspect all incoming and outgoing network traffic, compare network packet signatures against a database of suspicious signatures, or use artificial intelligence. If the IDS identifies a network connection as suspicious, it sends an alert to the user. However, on a home network, it is difficult for users without cybersecurity expertise to understand IDS alerts, distinguish cyberattacks from false alarms, and take appropriate action in a timely manner. This puts the security of home networks, smart home installations, home office workers, etc. at risk, even if an IDS is properly installed and configured. In this work, we propose ChatIDS, our approach to explain IDS alerts to non-experts using large language models. We evaluate the feasibility of ChatIDS using ChatGPT and identify open research questions with the help of interdisciplinary experts in artificial intelligence. Potential issues in areas such as trust, privacy, ethics, etc. need to be addressed before ChatIDS can be put into practice. Our results show that ChatIDS has the potential to improve network security by suggesting meaningful security measures from IDS alerts in an intuitive language.**

***Keywords*—*Intrusion Detection; ChatGPT; Smart Home.***

## I. Introduction

This paper is an extended version of our previous short paper [1]; the current paper extends that work by providing additional use cases, experiments, and analysis.

In recent years, private networks have come into the focus of cyberattacks. Reasons for this include the increased use of home-office work models [2], a shift to private areas during pandemics [3], or the proliferation of smart home devices [4]. IDSs are a well-established approach to detecting and mitigating cyberattacks [5], [6]. An IDS scans the network and/or network devices and sends alerts about suspicious network activity. This allows its users to detect cyberattacks at an early stage, possibly before any damage is done. On the other hand, an IDS might generate numerous false alarms.

In industry, business, and government, IDSs are a critical line of defense in the cybersecurity infrastructure. To this end, these sectors employ well-trained cybersecurity experts to configure, manage, and maintain IDSs, continuously improve the IDS rule set, distinguish false positives from real attacks, and design, prioritize, and implement appropriate countermeasures. It is possible to pre-configure a network-based IDS for home networks [7]. 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 develop appropriate and timely countermeasures. As cyberattacks, IDS configurations, and network traffic evolve rapidly, static explanations of known cyberattacks [8] cannot easily replace cybersecurity expertise.

In this paper, we outline ChatIDS, our approach to having 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 an LLM and allows the user to ask questions if the generated texts are not yet understandable. We make four contributions:

- We specify the requirements for an approach that increases network security in private networks by explaining IDS alerts to a non-expert.
- We describe ChatIDS, our approach to having ChatGPT [9] explain alerts from Snort [10], Suricata [11], and Zeek [12]. The explanations include cybersecurity actions and guidance on why/when to take the actions.
- We evaluate the feasibility of this approach through a small series of experiments with typical IDS alerts.
- To explore the design space of ChatIDS, we had interdisciplinary AI experts identify questions that need to be researched before ChatIDS can be put into practice.

Our experiments show that ChatIDS is easy to implement, although more work needs to be done on prompt engineering to ensure intuitive explanations on the first try. It is difficult to measure whether ChatIDS actually increases network security because it depends on the user. Our interdisciplinary experts have provided valuable insights. For example, from an ethical point of view, it is important to prevent the user from becoming too dependent on the technology, e.g., if ChatIDS allows security incidents to be repaired without the user having to acquire knowledge.

**Paper Structure:** Section II introduces related work. Section III outlines ChatIDS, our approach to explaining IDS messages to non-experts. Section IV describes a series of experiments to prove feasibility, and Section V identifies open research questions. Finally, Section VI concludes.


(a) Smart bulb (b) Smartphone application

Figure 1. Smart bulb and control application

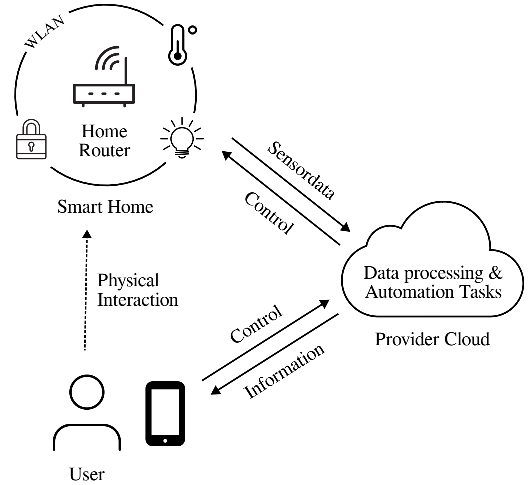


Figure 2. Smart Device Ecosystem

## II. Related Work

In this section, we present related work on the smart home, network security, general warning message design, and generative AI models.

### A. Smart Home

**Home Automation** transforms living spaces into dynamic, responsive environments for comfort and efficiency. Popular examples include climate control, home security, lighting or entertainment systems [13]. Typically, the installation of home automation systems is expensive and requires experts [14]. The **smart home** simplifies the process of home automation [13]. For example, let's consider a smart lighting system that can be controlled remotely, adjusting brightness and color tone, and automatically turning on in the morning or simulating a sunset at night. To install this system, all you need to do is replace the regular light bulbs with smart bulbs (Figure 1a), connect them to the Internet router, and install an app on the homeowner's smartphone. The smartphone becomes an remote control (fig. 1b) that can be used to control and configure all desired lighting. Typically, a smart home consists of several such smart devices [15].

Most smart devices depend on a sophisticated **IT ecosystem** that communicates with various external parties over the Internet, as shown in Figure 2. In particular, a smart device establishes Internet-based communication to enable functionalities such as remote control, automation tasks, multimedia services, interaction with other smart devices, cloud services, voice assistants, or software updates [15]. As a consequence, smart home devices are typically **accessible over the Internet**, perhaps through other devices. This distinguishes smart home devices from consumer devices such as laptops or smartphones, which use the Internet as a client, but do not listen for connections from the Internet. Because of this, smart home devices can be victims of unsolicited communication attempts that can be exploited for cyberattacks.

### B. Network Security

**IDSs** monitor a system for unauthorized or suspicious activity and can be distinguished by system type and detection type. The **system type** can be *host-based*, to monitor a single device, or *network-based* to monitor a network of devices. Host-based IDSs work by analyzing information in the local host's logs, which can be a combination of audit, system, and application and system call logs. Network-based IDSs are often centrally located to monitor multiple devices at the network level, they analyze network traffic by inspecting features of the traffic flow and network protocols [16].

**Detection types** can be *anomaly-detection* or *misuse-detection*. *Anomaly-detection* models the regular behavior of a system and detects activity that is significantly different from this behavior [16]. Anomaly detection systems calculate a score that is compared to an expert defined threshold, and if the score exceeds the threshold, an alert is sent. By adjusting the threshold, the sensitivity of the system can be set. There are detection engines based on Machine Learning [17], Deep Learning [18], Genetic Algorithms [19] and many more [20].

*Misuse-detection*, which models anomalous behavior and detects malicious patterns by comparing them to a predefined set of rules. Misuse-detection systems send an alert when a known pattern is identified [16]. Popular examples of network-based misuse-detection IDSs are Snort [10], Suricata [11] and Zeek [12]. A ruleset is required to use these IDSs. Popular pre-defined rulesets for networks are snort3-community-rules [21], suricata-rules [22], Yara [23] and Sigma [24].

### C. Warning Message Design

In general, **warning messages** are designed to warn people before a possible harm occurs. However, warning messages must be interpreted and understood. Thus, the effect of warning messages can be unreliable, and other measures should be exhausted before sending a warning message [25]. Non-experts do not always comply with the advice of warning messages. There are three key reasons for this behavior: (1) non-experts do not fully understand warning messages, (2) they do not always trust warning messagess, and (3) they think that compliance will cost them [26].

To increase the **effect of warning messages** in the field of cybersecurity, a developer might be tempted to design warning messages that try to make users fear cyberattacks. This approach has proven to be ineffective [27], [28]. Current recommendations say, that **good warning messages** should be *brief* [29], use *nontechnical language* [29], [30], describe the *risk* [29], describe the *consequences* of notcompliance [29], describe how the cyberattack will *affect* the user personally [31], [32], provide *instructions* on how to avoid the risk [29] and do so in a way that *aligns with how the user thinks* about cyberattacks [31].

### D. Generative AI

**Generative modeling** strives to create models that are capable of generating new data, such as sound, text, or images, that are similar to the data on which the model was trained [33]. Popular examples of generative models are WaveNet [34], which can generate speech and music, Pix2Pix, which can transform images into different styles [35], or GPT-3, an LLM that allows the generation of human-like text [36]. Another example of an LLM is ChatGPT [9]. Like a chatbot, ChatGPT is conversational and can generate detailed answers to questions. Bard [37] follows a similar approach. There are generative models trained for cybersecurity problems, such as Microsoft Security Copilot [38], 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, bias, and hallucinations [39].

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, it can be difficult to distinguish from real perception [40].

**Prompts** are the input to a generative model, they can be a text or an image that gives the model instructions for the requested output. Prompts provide an intuitive way to interact with generative models [41]. For image generation, a prompt can be another image or a text description. For LLMs, a prompt is text that provides context for the desired output, such as 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 construct prompts and optimize the output, analogous to software patterns [42]. For example, the *Persona Pattern* lets the LLM take on a specific role. This can be useful 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 allows the user to provide or remove context from a prompt.

## III. ChatIDS: Explainable Security

In this section we describe ChatIDS, our approach to explaining IDS messages to non-experts. Our goal is to integrate a network-based IDS into private networks to protect the network against cyberattacks from the Internet. Therefore, we want to replace the lack of cybersecurity expertise with an LLM. The LLM transforms and enhances the alerts of an IDS so that a private user can understand them and take appropriate action. 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 kind of expertise. A user may follow manuals written without technical vocabulary. It is difficult for a user to determine whether an IDS alert is due to a real attack or a false positive by the IDS, and to act accordingly.

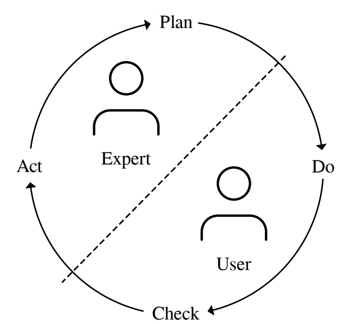


Figure 3. Adapted IT-Security Process

The IT-Security process follows a classic **Plan-Do-Check-Act** cycle [43]. A smart home IDS [7] can be integrated into such a process [44] as follows:

In the *Plan* phase, an expert preconfigures an IDS for typical smart homes. In the *Do* phase, the user installs the preconfigured IDS, which inspects network traffic for potential attacks in the *Check* phase. In the *Act* phase, an expert reviews logs to adapt the IDS for further attacks. Figure 3 illustrates this.

However, without knowledge of cybersecurity the user is left in the *Check* phase with only three possible actions: (a) do nothing, (b) turn off the device that may be under attack, 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.

### A. Requirements

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., whether or not immediate action is required.

**R3: (Actions)** The user must identify appropriate actions, such as performing a factory reset and installing a security patch.

To explore the solution space for a generative AI approach that fulfills these requirements for IDS, we use a constructive

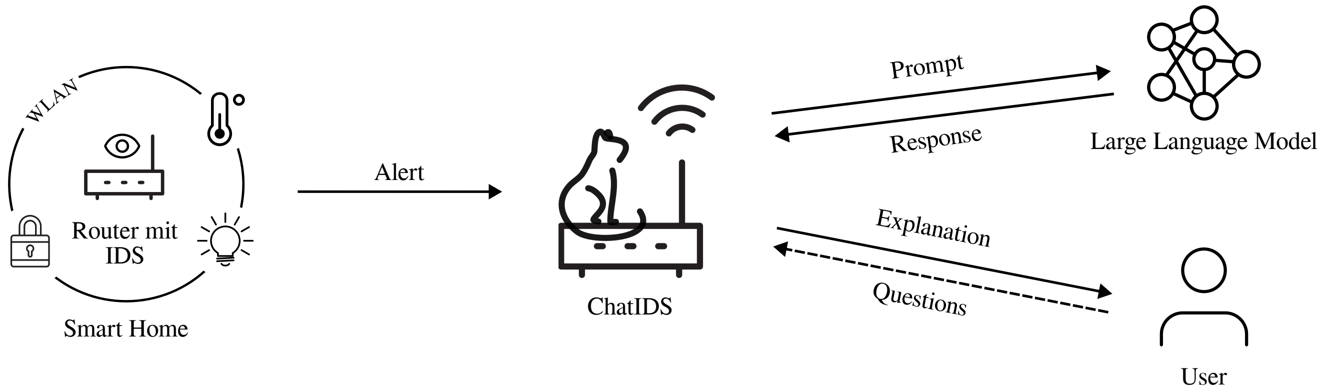


Figure 4. ChatIDS Workflow

research method. In particular, we (a) model ChatIDS, (b) use it to evaluate its technical feasibility, and (c) to discuss potential problems with interdisciplinary AI experts.

### B. Our ChatIDS Approach

The ChatIDS workflow is illustrated in Figure 4. A **network-based IDS component** inspects network packets passing through a router for suspicious traffic and generates alerts. The IDS should be a signature-based IDS so that its alerts are specific enough for the LLM.

The **LLM component** is responsible for translating the alerts from the IDS into a language that a non-expert can understand. Furthermore, this component can be used in an interactive way: If users do not understand the explanation or the suggested actions, they can ask for details. Similar to the IDS, the LLM is an external component of 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 predefined 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 to the LLM component along with a set of dummy alerts, so that the LLM component cannot learn the real alert with certainty. The explanations from the LLM component are cached so that the same explanation does not have to be requested twice.

## IV. Experimental Evaluation

We evaluate ChatIDS with 20 selected cyberattack alerts to find out how well it meets our requirements R1-R3.

### A. Experimental Setup

In line with Figure 4, we assume a home network with several smart home devices. A router connects the network to the Internet and can monitor all network packets. We assume that a Philips Hue Bridge [45] is being attacked. To implement the network-based IDS it is installed on the router and runs either the Snort [21], Suricata [22], Yara [23], or Sigma [24] ruleset. From each IDS implementation, we experiment with 20 alerts, as shown in the first column of Table I. All alerts are classified as important, and a user intervention is required.

The alerts generated are influenced by the target device, the user, and the alert. In our evaluation, we only set the alert as a variable to demonstrate variation and allow for comparability of our proof of concept.

> 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 [USER] 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.

Figure 5. Template for a ChatGPT Prompt

We realized ChatIDS using ChatGPT (gpt-3.5-turbo) [9]. To generate an explanation, ChatIDS embeds each alert into a ChatGPT prompt, as shown in Figure 5. This prompt implements the "Template", "Persona", and "Context Manager" patterns, 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 ChatIDS to produce hallucinations or erroneous output.

### B. Evaluation

Figure 6 shows an example of the output generated by ChatIDS. For better understanding, the figure shows the non-anonymized output, which includes names and devices. We

TABLE I. EVALUATION OF ALERTS AND RESPONSES

| Alert | Description | Intuition | Consequences | Urgency | Countermeasures | Correctness |
|---|---|---|---|---|---|---|
| MALWARE-CNC Harakit botnet traffic | 2 | 1 | 2 | 1 | 1 | -1 |
| SERVER-WEBAPP NetGear router default password login attempt admin/password | 2 | 2 | 1 | 1 | -2 | 1 |
| PROTOCOL-ICMP TFN Probe2 | 2 | 2 | 2 | 2 | 1 | 0 |
| PROTOCOL-FTP Bad login | 2 | 2 | 2 | 1 | -1 | 1 |
| SERVER-OTHER SSH server banner overflow | 1 | 2 | 1 | 2 | 0 | -1 |
| SURICATA MQTT unassigned message type (0 or >15) | 0 | 2 | 2 | 2 | 0 | 1 |
| SURICATA HTTP Response abnormal chunked for transfer-encoding | 0 | -1 | 2 | 0 | 0 | -1 |
| SURICATA SSH too long banner | 2 | 2 | 2 | 2 | 2 | 0 |
| SURICATA FTP Request command too long | 1 | 2 | 2 | 2 | 1 | 1 |
| SURICATA HTTP invalid content length field in request | 0 | -1 | 2 | 2 | 2 | 1 |
| Mirai Botnet TR-069 Worm - Generic Architecture | 1 | 0 | 2 | 2 | -1 | -1 |
| Linux.IotReaper | 2 | 2 | 2 | 2 | -1 | -1 |
| BleedingLife2 Exploit Kit Detection | 1 | 2 | 2 | 2 | 0 | 2 |
| Weevely Webshell - Generic Rule - heavily scrambled tiny web shell | 2 | 2 | 1 | 1 | 1 | 1 |
| Mirage Identifying Strings | 1 | 2 | 2 | 1 | 0 | 0 |
| (Zeek) Identifies IPs performing DNS lookups associated with common Tor proxies. | 0 | 1 | 0 | 1 | -1 | -1 |
| Ensure that all account usernames and authentication credentials are transmitted across networks using encrypted channels. | -2 | 1 | 1 | 2 | -1 | 1 |
| Identifies clients that may be performing DNS lookups associated with common currency mining pools | 1 | 1 | 2 | 1 | -1 | 0 |
| Detects URL pattern used by iOS Implant | 2 | -1 | 2 | 1 | -2 | -1 |
| Detects a bash contecting to a remote IP address | 2 | -1 | 2 | 1 | -2 | -1 |

have tested the prompts from the first column of the table I with ChatIDS, one by one. To evaluate their output according to our three requirements **R1: (Errors)**, **R2: (Urgency)**, and **R3: (Actions)**, we used a qualitative approach. We rated the degree of fulfillment of each requirement on a 5-point Likert [46] scale from strongly agree (2) to strongly disagree (-2).

For each requirement, we consider two different features: For **R1**, we evaluate the quality of the generated *Description* and how *Intuitive* it is. The *Description* should provide an accurate description of the attack, while the *Intuition* is considered good if the message is written in accessible language without technical vocabulary. For **R2** we evaluate the description of the *Consequences* for ignoring the alert and the *Urgency* of the description. The *Consequences* of an attack should be explained objectively, taking into account the attacked device, the home network, and the user's privacy. To convey a sense of *Urgency* and motivate the user to act, the message should be clear and concise. For **R3**, we look for a *Correct* and meaningful description of *Countermeasures*. A user should be able to counter the attack with the countermeasures provided by the message and not need further instructions. We also evaluate the correctness of the explanation. The countermeasures provided by the message should resolve all problems on the compromised device and on the home network.

### C. Results

Columns 2-7 in Table I list the results of the evaluation of these features. The table shows that ChatIDS fulfilled R1 (Errors) rather well. With only one exception, ChatIDS produced a good description of the security issue that caused the alert. The terminology used in the generated messages is almost always intuitive, with only a few exceptions. R2 (Urgency) was fulfilled even better. ChatIDS described the consequences of inaction and conveyed a sense of urgency almost always to our complete satisfaction. R3 (Errors) leaves room for improvement. We were surprised to find that the description of countermeasures was often too general to be followed by a user without further instruction. In addition, the recommended countermeasures wouldn't completely eliminate the threat described by the IDS alert and are therefore incorrect.

***Subject: Important Alert - Your Philips Hue Bridge is at Risk***

Dear Jon,

I hope you're well. I have some concerning news about your Smart Home, specifically the Philips Hue Bridge. Our system detected something unusual, and I want to make sure you understand the situation and how to address it.

**1. What Happened?**
The alert we received mentions "SURICATA SSH too long banner." This means someone is trying to access your Philips Hue Bridge in a way that's not typical or authorized. It's like someone trying to open a window to your house without your permission.

**2. Potential Consequences:**
If we don't address this intrusion, the person trying to access your Smart Home could potentially:
- Change your lighting settings or turn off lights when you don't want them to.
- Access personal data or settings related to your lights and Smart Home devices.
- Create a security risk for your home network, potentially leading to other issues.

Now, here's what you can do to stop this intrusion and secure your Smart Home:

**Instructions:**

**A. Update Your Philips Hue Bridge:**
- Go to the Philips Hue app on your smartphone.
- Check for any available updates for your Hue Bridge and follow the on-screen instructions to install them. Updates often contain important security fixes.

**B. Change Your Password:**
- Open the Philips Hue app.
- Find the settings or account section.
- Look for the password or security settings and change your password to something strong and unique.
- Avoid using common passwords, like "123456" or "password."

**C. Check for Suspicious Devices:**
- Open the Philips Hue app.
- Look for the "Connected Devices" or "Authorized Devices" section.
- Review the list and make sure you recognize all the devices connected to your Hue Bridge. If you see any unfamiliar devices, remove them.

**D. Monitor Activity:**
- Keep an eye on the activity of your Philips Hue Bridge in the app.
- If you notice any unexpected changes or activities, such as lights turning on or off by themselves, take it seriously.

These steps should help you secure your Philips Hue Bridge and prevent any further unauthorized access to your Smart Home. If you have any questions or need further assistance, please don't hesitate to reach out.

Your Smart Home security is important, so let's work together to keep it safe. Stay vigilant, and don't hesitate to take action. Thank you for your attention to this matter.

Best regards,

ChatIDS

Figure 6. Example for an Explanation

## V. Open Research Issues

In this section, we outline promising directions for future research. We are interested in warnings that address different audiences, we are interested in practical implementation, and we want to compile open questions for interdisciplinary research.

### A. *Prompt Design for Different Target Groups*

We know from related work (cf. Subsection. II-C) that warning messages are more effective when they are tailored to the recipient's information needs and mindset. For example, a suspicious person may ignore a message that emphasizes the potential damage, while a confident person needs such a warning to spur action. An experienced user may be frustrated by a message that oversimplifies technical details.

One promising way to address this issue is to customize the LLM prompts for different cultural backgrounds, skill levels, and protection requirements. Our ChatIDS approach makes it easy to integrate multiple pre-defined templates for LLM prompts, so that warning messages can be targeted to different audiences. We ran a series of preliminary experiments with different templates for the same IDS alert. For illustration, Figures 8 and 9 in the Appendix contain two different warning messages. Both of them translate the IDS alert “SURICATA SSH too long banner”. Figure 8 addresses users with some technical expertise, while Figure 9 is aimed at reducing anxiety.

### *B. Implementation and Design*

As ChatIDS addresses a practical cybersecurity challenge, implementation aspects such as system architectures or user interfaces need to be considered. We have implemented an interface from ChatIDS to the open source home automation software Home Assistant [47]. Figure 7 shows what a warning message from Home Assistant looks like on a Raspberry Pi. To generate the warning message, we used a ChatIDS template for a very short and simple message.

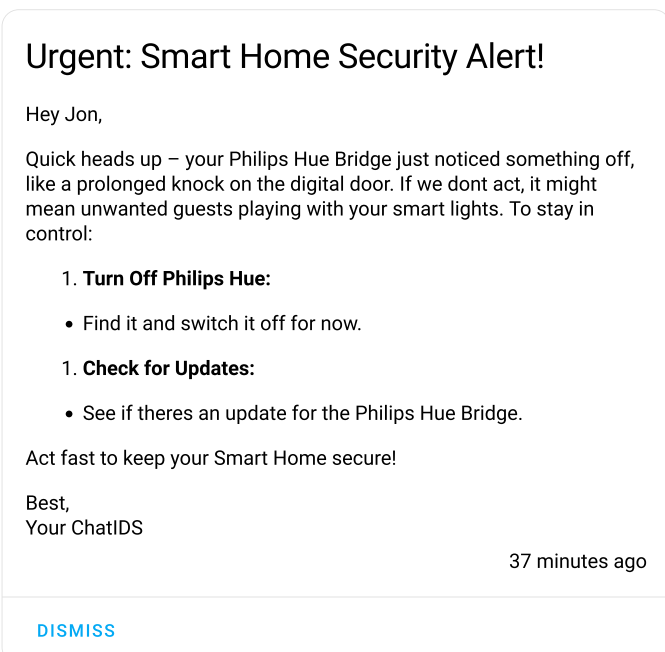


Figure 7. Short Warning Message displayed by Home Assistant

### *C. Interdisciplinary Research*

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. Furthermore, we demonstrate different showcases on how to vary warning messages and aim them at specific user groups. Our experts identified challenges from 6 areas:

**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 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 [48].

**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

This paper describes our ongoing work on ChatIDS, our approach to explaining alerts from an IDS to non-experts. ChatIDS sends anonymized alerts to ChatGPT, an LLM, to intuitively explain the alert and suggest meaningful countermeasures to cyberattacks. Our experiments show that ChatIDS can be easily implemented, although more work is needed on prompt engineering to ensure intuitive explanations on the first try. Furthermore, it needs to be analyzed whether anonymizing the data could remove relevant context or influence the report. It is difficult to measure whether ChatIDS actually increases network security, as it depends on the user. Our interdisciplinary experts have provided valuable insights. In the future, we will continue to improve ChatIDS in terms of security and privacy, as well as 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] V. Jüttner, M. Grimmer, and E. Buchmann, "ChatIDS: Explainable cybersecurity using generative AI," in *Proceedings of the 17th International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'23)*, 2023.

[2] N. Vakakis, O. Nikolis, D. Ioannidis, K. Votis, and D. Tzovaras, "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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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, pp. 1341–1346.

[9] Introducing ChatGPT. OpenAI. (Accessed on: 18.12.2023). [Online]. Available: https://openai.com/blog/chatgpt

[10] What is Snort? Cisco. (Accessed on: 18.12.2023). [Online]. Available: https://www.snort.org/

[11] Suricata. Open Information Security Foundation. (Accessed on: 18.12.2023). [Online]. Available: https://suricata.io/

[12] The Zeek Project. An open source network security monitoring tool. (Accessed on: 18.12.2023). [Online]. Available: https://zeek.org/

[13] W. A. Jabbar, T. K. Kian, R. M. Ramli, S. N. Zubir, N. S. M. Zamrizaman, M. Balfaqih, V. Shepelev, and S. Alharbi, "Design and fabrication of smart home with internet of things enabled automation system," *IEEE Access*, vol. 7, 2019.

[14] J. Greichen, "Value based home automation for todays' market," *IEEE Transactions on Consumer Electronics*, vol. 38, 1992.

[15] B. K. Sovacool and D. D. Furszyfer Del Rio, "Smart home technologies in Europe: A critical review of concepts, benefits, risks and policies," *Renewable and Sustainable Energy Reviews*, vol. 120, p. 109663, 2020.

[16] T. Verwoerd and R. Hunt, "Intrusion detection techniques and approaches," *Computer communications*, vol. 25, no. 15, pp. 1356–1365, 2002.

[17] H. Xu, Z. Sun, Y. Cao, and H. Bilal, "A data-driven approach for intrusion and anomaly detection using automated machine learning for the internet of things," *Soft Computing*, 2023.

[18] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Anomaly based network intrusion detection for IoT attacks using deep learning technique," *Computers and Electrical Engineering*, vol. 107, 2023.

[19] A. H. Hamamoto, L. F. Carvalho, L. D. H. Sampaio, T. Abrão, and M. L. Proença, "Network anomaly detection system using genetic algorithm and fuzzy logic," *Expert Systems with Applications*, vol. 92, pp. 390–402, 2018.

[20] M. Alkasassbeh and S. Baddar, "Intrusion detection systems: A state-of-the-art taxonomy and survey," *Arabian Journal for Science and Engineering*, vol. 48, 2022.

[21] Snort community rules. Cisco. (Accessed on: 18.12.2023). [Online]. Available: https://www.snort.org/downloads/

[22] Open Information Security Foundation. (Accessed on: 18.12.2023). [Online]. Available: https://github.com/OISF/suricata/tree/master/rules

[23] Yara-rules. YaraRules Project. (Accessed on: 18.12.2023). [Online]. Available: https://github.com/Yara-Rules/rules

[24] Sigma rules. SigmaHQ. (Accessed on: 18.12.2023). [Online]. Available: https://github.com/SigmaHQ/sigma/tree/master/rules

[25] M. Wogalter, "Purposes and scope of warnings," *Handbook of Warnings*, pp. 3–9, 01 2006.

[26] K. S. Jones, N. R. Lodinger, B. P. Widlus, A. Siami Namin, E. Maw, and M. E. Armstrong, "How do non experts think about cyber attack consequences?" *Information & Computer Security*, vol. 30, no. 4, pp. 473–489, 2022.

[27] A. Sasse, "Scaring and bullying people into security won't work," *IEEE Security & Privacy*, vol. 13, no. 3, pp. 80–83, 2015.

[28] M. Dupuis, A. Jennings, and K. Renaud, "Scaring people is not enough: An examination of fear appeals within the context of promoting good password hygiene," in *Proceedings of the 22nd Annual Conference on Information Technology Education*. Association for Computing Machinery, 2021, pp. 35–40.

[29] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri, "Bridging the gap in computer security warnings: A mental model approach," *IEEE Security & Privacy*, vol. 9, pp. 18–26, 2011.

[30] L. Bauer, C. Bravo-Lillo, L. Cranor, and E. Fragkaki, "Warning design guidelines," CyLab, Carnegie Mellon University, Tech. Rep., 2013.

[31] S. Bartsch, M. Volkamer, H. Theuerling, and F. Karayumak, "Contextualized web warnings, and how they cause distrust," in *Trust and Trustworthy Computing: 6th International Conference*. Springer, 2013, pp. 205–222.

[32] M. Kauer, T. Pfeiffer, M. Volkamer, H. Theuerling, and R. Bruder, "It is not about the design - it is about the content! making warnings more efficient by communicating risks appropriately," in *SICHERHEIT 2012 – Sicherheit, Schutz und Zuverlässigkeit*, 2012.

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

[34] Wavenet. Alphabet Inc. (Accessed on: 18.12.2023). [Online]. Available: https://www.deepmind.com/research/highlighted-research/wavenet

[35] 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.

[36] GPT-3 powers the next generation of apps. OpenAI. (Accessed on: 18.12.2023). [Online]. Available: https://openai.com/blog/gpt-3-apps

[37] Meet Bard. Google Ireland Limited. (Accessed on: 18.12.2023). [Online]. Available: https://bard.google.com/

[38] Introducing Microsoft Security Copilot. Microsoft. (Accessed on: 18.12.2023). [Online]. Available: https://www.microsoft.com

[39] 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.

[40] 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, 2023.

[41] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *Computing Research Repository*, vol. abs/2211.01910, 2022.

[42] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with ChatGPT," 2023.

[43] J. Eloff and M. Eloff, "Information security architecture," *Computer Fraud & Security*, 2005.

[44] 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.

[45] Philips Hue Bridge. Signify Holding. (Accessed on: 18.12.2023). [Online]. Available: https://www.philips-hue.com

[46] R. Likert, "A technique for the measurement of attitudes," *Archives of psychology*, 1932.

[47] Home Assistant. Home Assistant Core Team and Community. (Accessed on: 18.12.2023). [Online]. Available: https://www.home-assistant.io/

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


## APPENDIX A
### WARNING MESSAGE FOR EXPERTS OR ANXIOUS USERS

Figures 8 and 9 show warning messages that were generated with templates tailored for expert users or anxious users.

***Subject: Urgent: Security Alert - Intrusion Detected on Philips Hue Bridge***

Dear Jon,

I hope this message finds you well. I need to bring an urgent matter to your attention regarding your Smart Home network. Our Intrusion Detection System has flagged an alert with the message "SURICATA SSH too long banner," indicating a potential security breach on your Philips Hue Bridge.

**Explanation of the Intrusion:**

The alert suggests that an unauthorized attempt has been made to access your Philips Hue Bridge using an SSH connection with an excessively long banner. This could be a sign of a malicious actor attempting to gain unauthorized access to your Smart Home network.

**Potential Consequences:**

If this intrusion is not addressed promptly, it could lead to the following consequences:

1) Unauthorized access to your Smart Home devices, compromising your privacy.
2) Manipulation or control of your Philips Hue lights and other connected devices.
3) Potential exposure of sensitive information within your network.

**Instructions to Stop the Intrusion:**

To secure your Smart Home and mitigate the potential risks, please follow these steps:

1) **Isolate Philips Hue Bridge:**
   - Disconnect the Philips Hue Bridge from your network immediately.
   - Power it off to prevent further communication.

2) **Review Device Logs:**
   - Examine the logs on the Philips Hue Bridge for any suspicious activities or login attempts.

3) **Update Firmware:**
   - Check for and apply the latest firmware updates for your Philips Hue Bridge.
   - This ensures that known vulnerabilities are patched.

4) **Change SSH Credentials:**
   - If applicable, change the SSH credentials for your Philips Hue Bridge.
   - Use a strong, unique password to enhance security.

5) **Network Security Check:**
   - Conduct a thorough review of your home network's security settings.
   - Ensure that all devices are protected with strong passwords and up-to-date security measures.

6) **Monitor Activity:**
   - Keep a close eye on your Smart Home devices for any unusual behavior.
   - Report any further suspicious activity to our IT support team.

Please take these steps urgently to safeguard your Smart Home from potential threats. If you have any questions or need assistance, feel free to reach out to our IT support team immediately.

Thank you for your prompt attention to this matter.

Best regards,
ChatIDS

Figure 8. Warning message aimed at expert users

***Subject: Important: Action Required to Secure Your Smart Home***

Dear Jon,

I hope this message finds you well. I wanted to bring to your attention a recent notification from your Smart Home setup, specifically concerning the "Philips Hue Bridge." No need to worry, but it's essential to take a few simple steps to ensure everything stays secure.

**What Happened:**
Our Smart Home system noticed something unusual with the "Philips Hue Bridge," which helps control your smart lights. It's like a little detective that caught a small hiccup.

**Potential Consequences (But Nothing to Panic About!):**
If we don't address this now, it could affect the smooth functioning of your smart lights. Think of it as a minor hiccup that might cause some inconvenience, but nothing serious. We want to make sure everything runs seamlessly for you.

**Here's What You Can Do (Super Easy, I Promise!):**

- **Power Cycle the Philips Hue Bridge:**
  1) Unplug it from the power source.
  2) Wait for 10 seconds.
  3) Plug it back in.
- **Check for Software Updates:**
  1) Open the app you use for your smart lights.
  2) Look for any updates available for the Philips Hue system.
  3) If there are any, go ahead and update.
- **Secure Your Wi-Fi:**
  1) Ensure your Wi-Fi password is strong and known only to you.
  2) If you haven't changed it recently, now might be a good time to do so.

Remember, this is just a precaution to keep everything running smoothly. If you have any questions or need further assistance, feel free to reach out.

Best regards,

ChatIDS

Figure 9. Warning message aimed to reduce anxiety