Dawn of LLM4Cyber: Current Solutions, Challenges, and New Perspectives in Harnessing LLMs for Cybersecurity

Luca Caviglione[1], Carmela Comito[2], Erica Coppolillo[2,5], Daniela Gallo[2,3], Massimo Guarascio[2], Angelica Liguori[2,*], Giuseppe Manco[2], Marco Minici[2,6], Simone Mungari[2,5,7], Francesco Sergio Pisani[2], Ettore Ritacco[4], Antonino Rullo[2], Paolo Zicari[2] and Marco Zuppelli[1]

[1]*Institute for Applied Mathematics and Information Technologies, Via de Marini 6, Genova, 16149, Italy*
[2]*Institute for High Performance Computing and Networking, via P. Bucci 8-9/C, Rende, 87036, Italy*
[3]*University of Salento, Piazza Tancredi, 7, Lecce, 73100, Italy*
[4]*University of Udine, Via Palladio, 8, Udine, 33100, Italy*
[5]*University of Calabria, via P. Bucci, Rende, 87036, Italy*
[6]*University of Pisa, via Lungarno Pacinotti, Pisa, 56126, Italy*
[7]*Revelis s.r.l., Viale della Resistenza, Rende, 87036, Italy*

Abstract

Large Language Models (LLMs) are now a relevant part of the daily experience of many individuals. For instance, they can be used to generate text or to support working duties, such as programming tasks. However, LLMs can also lead to a multifaceted array of security issues. This paper discusses the research activity on LLMs carried out by the ICAR-IMATI group. Specifically, within the framework of three funded projects, it addresses our ideas on how to understand whether data has been generated by a human or a machine, track the use of information ingested by models, combat misinformation and disinformation, and boost cybersecurity via LLM-capable tools.

Keywords

Large Language Models, Watermarking, Cybersecurity, Fake news, Event log analysis

1. Introduction

Large Language Models (LLMs) allow to generate a wide array of contents. For instance, they can be used to create textual documents, pieces of music, as well as source code. A feature very relevant for their success is the ability of mimicking the human behavior. Unfortunately, this makes LLMs a double-edged sword since they can be exploited to generate realistic yet malicious content, such as fake news or text supporting misinformation campaigns. At the same time, LLMs have also proven to be effective in supporting various cyber-security duties, for instance, to analyze logs or network traffic [1].

In an attempt to fully understand the potential of LLMs in terms of offensive capabilities as well as the opportunities that should be seized to advance in the security of the Internet, researchers of the Institute for High Performance Computing and Networking - ICAR and of the Institute for Applied Mathematics and Information Technologies - IMATI of the National Research Council of Italy - CNR have intensified their efforts to investigate the pros and cons of LLMs. This research effort is established within the framework of three research projects. The first is funded by the Consortium named "SEcurity and RIghts In the CyberSpace - SERICS", and aims at using LLMs to increase the security posture of networking and computing systems. For instance, an LLM can be used to synthesize behaviors starting from logs of containerized microservices or to generate automatic textual replies to deceive e-mail scammers [2]. The second research action is funded by the project "Watermarking Hazards and novel perspectives in Adversarial Machine learning - WHAM!", and is devoted to quantifying the limits and opportunities of watermarking schemes when applied to AI artifacts. As an example, data can be hidden

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Corresponding author.
luca.caviglione@ge.imati.cnr.it (L. Caviglione); carmela.comito@icar.cnr.it (C. Comito); erica.coppolillo@icar.cnr.it (E. Coppolillo); daniela.gallo@icar.cnr.it (D. Gallo); massimo.guarascio@icar.cnr.it (M. Guarascio); angelica.liguori@icar.cnr.it (A. Liguori); giuseppe.manco@icar.cnr.it (G. Manco); marco.minici@icar.cnr.it (M. Minici); simone.mungari@icar.cnr.it (S. Mungari); francescosergio.pisani@icar.cnr.it (F. S. Pisani); ettore.ritacco@uniud.it (E. Ritacco); antonino.rullo@icar.cnr.it (A. Rullo); paolo.zicari@icar.cnr.it (P. Zicari); marco.zuppelli@ge.imati.cnr.it (M. Zuppelli)
0000-0001-6466-3354 (L. Caviglione); 0000-0001-9116-4323 (C. Comito); 0000-0002-4670-8157 (E. Coppolillo); 0009-0009-3245-7738 (D. Gallo); 0000-0001-7711-9833 (M. Guarascio); 0000-0001-9402-7375 (A. Liguori); 0000-0001-9672-3833 (G. Manco); 0000-0002-9641-8916 (M. Minici); 0000-0002-0961-4151 (S. Mungari); 0000-0003-2922-0835 (F. S. Pisani); 0000-0003-3978-9291 (E. Ritacco); 0000-0002-6030-0027 (A. Rullo); 0000-0002-9119-9865 (P. Zicari); 0000-0001-6932-3199 (M. Zuppelli)

to recognize deep fakes, to understand whether a model has been cloned, or to track usages in Machine-Learning-as-a-Service deployments [3]. Even worse, problem of exploiting unauthorized content during training or in deployment needs to be specifically addressed. The third research action is funded by the project "Limiting MIsinformation spRead in online environments through multimodal and cross-domain FAKe news detection - MIRFAK", which aims at developing an innovative content verification tool, delivering solutions for news verification on social media and online platforms. Within the project, we aim at exploring the potentials and risks of LLMs associated with misinformation.

In this work, we outline our research agenda on these topics, which is devised in three directions: *i*) we present mid-term challenges for using LLMs to solve security-related issues; *ii*) we discuss how watermarks can be applied to LLMs to mitigate attacks aiming at stealing information or disseminating fake news; *iii*) we showcase the gaps to be filled to make LLMs a real asset for the Internet.

The rest of the paper is structured as follows. Section 2 deals with the problems of understanding whether the output has been generated by an LLM and of tracking its provenance, while Section 3 considers usage violations, such as unauthorized harvesting of data for training models. Section 4 discusses challenges and opportunities relative to the adoption of LLMs in the context of online social platforms and debates. Section 5 discusses the adoption of LLMs in assessing cybersecurity risks related to systems and infrastructures in containerized environments. Lastly, Section 6 concludes the work and portrays some prospected action points.

2. Are the Data Generated?

One of the main goals of our research is to investigate challenges and solutions for protecting the Intellectual Property (IP) of the Machine/Deep Learning (ML/DL) models as well as of the dataset used for the training phase [4]. Moreover, we also aim at considering techniques to mark the output produced by ML/DL services, for instance, to understand whether an attacker "cloned" the model through multiple remote invocations. Specifically, we are interested in techniques that allow the cloaking of secret information within the contents we want to protect. In this respect, an emerging research line considers watermarking techniques, i.e., arbitrary pieces of data that are embedded within the item to deliver and that are difficult to recognize besides proprietary decryption schemes. Such mechanisms are common with images and multimedia objects [5] and can be used to embed control data within ML/DL models.

Techniques used to prevent unwanted/unfair usages or to enforce IP can also be envisioned for generative models, with a particular focus on large language models. There are essentially two scenarios that are relevant in this respect. The first scenario is relative to the opportunity to mark generated text in a way that it can be easily recognized. Watermarking can be employed in this context to embed the watermark within the output of the LLM and, thus, distinguish between the data generated by a human and those produced by a machine. The objective here is to enforce IP protection as well as to claim ownership on the generated data. The second scenario is relative to the problem that such generative models can deliver malicious content. To mitigate potential harm caused by such generated data, it is crucial to develop methods to identify content generated by a machine, when a watermark is not embedded. It is worth noting that the generation of malicious content can be both unintentional or intentional. Unintentional generation may happen due to the stochastic nature of such generative models, which causes the phenomenon of *hallucinations* (i.e., unrealistic or imaginary content). By contrast, intentional generation is typically done by a malicious threat actor, who pushes the generative model to obtain mischievous data. In both cases, the generated data could be of high quality, infusing trust among readers eventually forcing them to fall into error or forward the content, e.g., through sharing functionalities of online social networks. Our research in this context aims at developing methods to identify contents generated by a machine through a language model. We are interested both in devising watermarking schemes and in the more general challenge relative to the problem of devising predictive methods for discriminating generated data. Besides, this research activity is aligned with the current requirements enforced by the recently released European AI Act[1]. The latter in fact introduces specific transparency obligations to ensure that humans are informed when necessary, to ensure trust, and in particular, that AI-generated content is identifiable.

The research approaches to this topic are quite recent. To the best of our knowledge, the first LLM watermarking technique for distinguishing human-generated from machine-generated texts was proposed by Kirchenbauer et al. [6]. In text generation, language-based models produce a probability distribution over a vocabulary, i.e., the set of words or word fragments (i.e., *tokens*), used for predicting the most likely next word based on the previous ones. The authors propose to alter such distribution, in order to promote sampling of specific tokens. The occurrence within a given statistical significance of such tokens characterizes the watermark within the text. One of the main limitations of this approach is the gen-

[1] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

eration of low-quality texts in contexts characterized by relatively deterministic content, such as code snippets or structured text. Lee et al. [7] refine the approach by ensuring that sampling is only focused on high-entropy tokens.

One of our research objectives is to generalize these approaches to other generative models, such as Diffusion Models or Generative Adversarial Networks (GANs). In addition, the analysis of the distribution of generated data, and its comparison with that of real (not synthetic) data can also be exploited for devising predictive models aimed at automatically detecting the reliability and authenticity of data.

3. Have You Stolen My Data?

Membership Inference attacks (MIAs) [8] aim to predict whether a data sample was included in the training dataset of a machine learning model. These attacks serve to evaluate the privacy vulnerabilities present in machine learning models, like in Neural Networks [9], GANs [10] and Diffusion Models [11]. Formally, the goal of a MIA is to infer whether a given data point x was part of the training dataset D for model M by computing a membership score $s(x; M)$. This score is then thresholded to determine a target sample's membership.

Membership inference attacks exploit the tendency of the models to overfit their training data and hence exhibit lower loss values for these elements. A first and widely used attack is the LOSS attack [12], in which samples are classified as training members if their loss values are lower than a fixed threshold (that is, $s(x; M)$ is defined in terms of $\mathcal{L}(x; M)$).

Recent works aim to design and improve MIAs for LLMs. In this case, MIAs consider a target model M which gives as output a probability distribution of the next token given a prefix as input, $\mathbb{P}(x_t|x_0 \dots x_{t-1}; M)$. The goal of MIA is hence to infer whether the target sample $x = x_1 \dots x_n$ of n tokens has been considered in the training set. Duan et al. [13] consider several membership inference attacks and show that they just outperform random guessing for most settings across different LLM size and domains. They also argue that MIA is difficult on LLMs because of different key reasons. These include the difficulty of handling LLMs pre-trained over billions and trillions of tokens, or the overlap typically exhibited by the underlying token distributions that can be observed in natural language documents, irrespective of their training data membership.

Our research agenda is aimed at extending and leveraging the current membership inference games, by investigating adversarial approaches in order to force the LLM to generate copyrighted text. In this way, we define a framework that can demonstrate copyright violations and overcome MIA's issues related to large dataset and the intrinsic randomness of LLMs.

4. Fighting Fire with Fire: Generative AI to promote Online Safety

LLMs are showcasing remarkable abilities in various Natural Language Processing tasks, making them a highly potent and beneficial tool for everyday life. However, alongside their appealing strengths and widespread adoption, a significant concern is arising regarding their potential role in amplifying the generation and dissemination of misinformation and disinformation. Generative AI technology has significantly empowered malicious actors to produce fake content, which can be disseminated across online social networks and lead to detrimental phenomena, e.g., manipulating public discourse, disseminating hate speech, and sharing fake content.

As a remarkable example, in 2016 Microsoft released the Tay chatbot, which triggered further controversy by posting inflammatory and offensive tweets via its Twitter account, leading Microsoft to shut down the service within just 16 hours [2]. More recently, other works assessed the role of bots and AI agents in conveying and amplifying online discourse about racism and hate speech [14, 15], drawing further attention to this sensitive topic. Thus, as underscored by [16], the scale, velocity and accessibility of generative models present compelling challenges for online platforms, potentially inundating them with a massive amount of fraudulent material and unpredictable social consequences. While policy makers are actively engaged in regulating the use of GenAI tools, the efficacy of these measures remains uncertain. In response, our research group is working towards leveraging Generative AI to enhance online safety. Our objective is to reuse the same technology used to contaminate online discussions for a beneficial purpose in a controlled environment. For instance, [17] demonstrated the potential of a GPT2-like model in crafting tailored responses to combat misinformation regarding the COVID-19 pandemic. Despite this first promising result, there are numerous overlooked opportunities for harnessing GenAI tools to aid online safety. One such opportunity involves the development of automated agents capable of serving as "peace-builders" within online discussions. We aim to train a large language model to generate textual content that, once injected within online social media platforms, can help mitigate polarization and disagreement.

This research line is interesting and open to novel and original developments, but it also faces considerable challenges. A trivial remark is to carefully consider the

[2] https://en.wikipedia.org/wiki/Tay_(chatbot)

ethical implications of using GenAI tools for online safety to ensure responsible use. Second, there are considerable technical challenges regarding the training and/or fine-tuning of these large models due to scalability concerns. Third, evaluating the effectiveness of GenAI interventions in promoting online safety can be demanding and could require a multi-disciplinary approach involving experts from fields such as psychology and sociology.

Another compelling line in our research agenda is to define the aspects to take into account when analyzing the role of LLMs in this context. We are interested in exploring the role of LLMs in contrasting the phenomenon of false information spreading at different levels: detection, mitigation, intervention, and attribution. Our effort is to improve the fake detection models under the constraint of scarcely labeled data, which is a common condition in real scenarios when discovering fakes in new topics and domains. The generative capabilities can be harnessed for exploring innovative augmentation techniques. LLMs can help reduce the learning strategy costs associated with expert interaction (e.g., Active Learning), thereby saving human annotators' time. This can be achieved by effectively integrating LLMs into learning loops at various levels, such as tuple selection and label generation support.

5. Boosting Cybersecurity

The last research line focuses on exploring various scenarios where LLMs can bolster cybersecurity operations. The concept involves utilizing AI-based tools to automate the analysis and processing of vast amounts of semi-structured data. This approach aims to evaluate security risks across systems and infrastructures more efficiently. While Machine and Deep Learning techniques have been widely used to discover deviant behaviors in event logs [18, 19, 20], the adoption of LLMs represents a novel and quite unexplored research line. For instance, in a recent work [21], the authors show how LLMs can be leveraged for analyzing huge volumes of information stored in logs.

A specific research objective is to support the automation of threat assessment. The intervention of the "expert" (i.e., the human operator) is still crucial to evaluate whether the anomalous event can be traced back to an actual attack or threat. Nevertheless, we believe that the adoption of tools based on LLM can support and facilitate this task. Thus, our mid-term research goals are twofold.

- **Improving efficiency.** To enhance response time to potential threats detected through logs, our strategy involves leveraging Active Learning techniques. These techniques enable human operators to actively participate in the model learning process, creating a human-in-the-loop system. Thus, our approach aims to expedite threat response when integrating human expertise into the learning loop of the model, by using post-hoc explanation tools to support the operator in validating the attack and guiding the learning of the model.
- **Data enrichment.** Another critical aspect involves the potential use of LLMs to enhance the security of Internet-wide infrastructures. Numerous protocols and services rely heavily on textual information, such as URLs or configuration data. LLMs can be exploited in generating test cases, particularly for automating periodic assessments aimed at detecting potential deviations in the security posture of a deployment. For example, recent research showcased LLMs' capability to generate attacks against web destinations, particularly in crafting SQL injections [22].

We also foresee the adoption of LLMs as tools for analysing textual descriptions of system configurations, in order to detect potential risks and vulnerabilities relative to such configurations.

A further relevant application of LLMs is the creation of a new-wave of tools to perform fuzz testing, especially for handling network protocols [23]. This is particularly relevant for a twofold reason. First, ubiquitous containerized/virtualized frameworks are progressively migrating to the intrinsically networked microservice paradigm. Second, the emerging plague of malwares exploiting information hiding is hard to mitigate, especially since it requires knowing in advance where the attacker will cloak the data [24].

In this perspective, LLMs could be used to discover in advance protocol fields, metadata, header information, or text segments in software that could be abused to conceal arbitrary/malicious content. For the case of networked (micro)services, fuzzers can be used to learn the grammar ruling a protocol starting from RFC documents [25]. These testing tools can hence be guided to explore interactions among containers or to fuzz specific operations, e.g., the setup/teardown of a connection.

For the case of information-hiding-capable malware, detection and sanitization are tightly coupled with the abused resource (e.g., digital media vs network traffic), and the number of features and ambiguities that can be exploited is almost unbounded. Therefore, fuzzers can be built by starting from datasets of pre-existent information-hiding-capable-attacks or trained over well-known clocking patterns [26]. Thus, LLMs can lead to guided fuzzers, which demonstrated their ability to reveal corner cases or uncommon anomalous templates [23].

A midterm goal is then to tweak an LLM to evaluate the limits of protocols when containing arbitrary information for implementing a covert communication. The

use of LLMs will be particularly efficient for protocols like HTML and MQTT, which are based on large portions of textual information, especially in the header [27]. Moreover, we also plan to investigate if LLMs can be used to improve the performance of our pre-existent AI/ML mechanisms for the detection of covert communications [28, 29].

6. Conclusions

LLMs present a spectrum of opportunities and challenges within the cybersecurity domain. We've delved into four primary research avenues, each addressing distinct problems and proposing corresponding solutions. These areas include:

- Watermarking and Detection of Generative Content: Developing methods to embed unique identifiers into data for tracking and authentication purposes, alongside techniques for detecting generative content to combat potential trustworthiness and security risks.
- Membership Inference and Data Provenance: Addressing concerns related to establishing the origin of training data, crucial for ensuring data integrity, privacy.
- Misinformation Mitigation/Intervention: Implementing strategies to combat misinformation and ensure online safety, particularly in the context of rapidly evolving online information landscapes.
- Log Analysis and Stress Testing in Infrastructure Protection: Analyzing system logs and subjecting infrastructures to stress tests to assess their resilience against cyber threats, essential for maintaining robust security measures.

We have devised specific solutions within the context of three research projects funded by the Italian Ministry of Research. These solutions aim to address various cybersecurity challenges and enhance overall digital security measures,

Acknowledgments

This work was partially supported by the following projects: 1) WHAM! - Watermarking Hazards and novel perspectives in Adversarial Machine learning (B53D23013340006); 2) SERICS - SEcurity and RIghts in the CyberSpace (PE00000014); 3) MIRFAK - Limiting MIsinformation spRead in online environments through multi-modal and cross-domain FAKe news detection (P2022C23K9), funded under the NRRP MUR program funded by the EU - NGEU. A part of the work was also supported by: Project RAISE (ECS00000035); MUR on D.M. 351/2022, PNRR Ricerca, CUP H23C22000550005; MUR on D.M. 352/2022, PNRR Ricerca, CUP H23C22000440007.

References

[1] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, Y. Zhang, A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly, High-Confidence Computing (2024) 100211.

[2] E. Cambiaso, L. Caviglione, Scamming the Scammers: Using ChatGPT to Reply Mails for Wasting Time and Resources, arXiv preprint arXiv:2303.13521 (2023).

[3] X. Zhao, Y.-X. Wang, L. Li, Protecting Language Generation Models via Invisible Watermarking, in: International Conference on Machine Learning, 2023, pp. 42187–42199.

[4] L. Caviglione, C. Comito, M. Guarascio, G. Manco, Emerging Challenges and Perspectives in Deep Learning Model Security: A Brief Survey, Systems and Soft Computing 5 (2023) 200050.

[5] N. Agarwal, A. K. Singh, P. K. Singh, Survey of Robust and Imperceptible Watermarking, Multimedia Tools and Applications 78 (2019) 8603–8633.

[6] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models, in: ICML, volume 202 of *Proceedings of Machine Learning Research*, 2023, pp. 17061–17084.

[7] T. Lee, S. Hong, J. Ahn, I. Hong, H. Lee, S. Yun, J. Shin, G. Kim, Who Wrote this Code? Watermarking for Code Generation, arXiv abs/2305.15060 (2023).

[8] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, X. Zhang, Membership inference attacks on machine learning: A survey, ACM Comput. Surv. 54 (2022). doi:10.1145/3523273.

[9] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, F. Tramer, Membership Inference Attacks From First Principles, 2022. arXiv:2112.03570.

[10] D. Chen, N. Yu, Y. Zhang, M. Fritz, GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models, in: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, CCS '20, ACM, 2020.

[11] J. Dubiński, A. Kowalczuk, S. Pawlak, P. Rokita, T. Trzciński, P. Morawiecki, Towards More Realistic Membership Inference Attacks on Large Diffusion Models, 2023. arXiv:2306.12983.

[12] S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha, Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, in: 2018 IEEE 31st Computer Security Foundations Symposium, 2018, pp. 268–282.

[13] M. Duan, A. Suri, N. Mireshghallah, S. Min,

W. Shi, L. Zettlemoyer, Y. Tsvetkov, Y. Choi, D. Evans, H. Hajishirzi, Do Membership Inference Attacks Work on Large Language Models?, 2024. arXiv:2402.07841.

[14] J. Uyheng, D. Bellutta, K. Carley, Bots amplify and redirect hate speech in online discourse about racism during the covid-19 pandemic, Social Media + Society 8 (2022) 205630512211047. doi:10.1177/20563051221104749.

[15] J. Uyheng, K. M. Carley, Bots and online hate during the covid-19 pandemic: case studies in the united states and the philippines, Journal of Computational Social Science 3 (2020) 445 – 468. URL: https://api.semanticscholar.org/CorpusID:224818205.

[16] S. Feuerriegel, R. DiResta, J. A. Goldstein, S. Kumar, P. Lorenz-Spreen, M. Tomz, N. Pröllochs, Research can help to tackle ai-generated disinformation, Nature Human Behaviour 7 (2023) 1818–1821.

[17] B. He, M. Ahamad, S. Kumar, Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 2698–2709.

[18] A. Cuzzocrea, F. Folino, M. Guarascio, L. Pontieri, A Multi-view Learning Approach to the Discovery of Deviant Process Instances, in: On the Move to Meaningful Internet Systems: OTM 2015 Conferences - Confederated International Conferences: CoopIS, ODBASE, and C&TC 2015, volume 9415 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 146–165.

[19] F. Folino, G. Folino, M. Guarascio, L. Pontieri, Semi-Supervised Discovery of DNN-Based Outcome Predictors from Scarcely-Labeled Process Logs, Business & Information Systems Engineering 64 (2022) 729–749.

[20] F. Folino, G. Folino, M. Guarascio, L. Pontieri, Data- & Compute-efficient Deviance Mining via Active Learning and Fast Ensembles, Journal of Intelligent Information Systems (2024).

[21] Z. Ma, A. R. Chen, D. J. Kim, T.-H. Chen, S. Wang, LLMParser: An Exploratory Study on Using Large Language Models for Log Parsing, in: 2024 IEEE/ACM 46th International Conference on Software Engineering, IEEE Computer Society, 2024, pp. 883–883.

[22] R. Fang, R. Bindu, A. Gupta, Q. Zhan, D. Kang, LLM Agents can Autonomously Hack Websites, arXiv preprint arXiv:2402.06664 (2024).

[23] S. Mallissery, Y.-S. Wu, Demystify the Fuzzing Methods: A Comprehensive Survey, ACM Computing Surveys 56 (2023) 1–38.

[24] L. Caviglione, W. Mazurczyk, Never Mind the Malware, Here's The Stegomalware, IEEE Security & Privacy 20 (2022) 101–106.

[25] C. S. Xia, M. Paltenghi, J. Le Tian, M. Pradel, L. Zhang, Fuzz4all: Universal Fuzzing with Large Language Models, Proc. IEEE/ACM ICSE (2024).

[26] S. Wendzel, S. Zander, B. Fechner, C. Herdin, Pattern-based Survey and Categorization of Network Covert Channel Techniques, ACM Computing Surveys 47 (2015) 1–26.

[27] T. Schmidbauer, S. Wendzel, SoK: A Survey of Indirect Network-level Covert Channels, in: Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security, 2022, pp. 546–560.

[28] N. Cassavia, L. Caviglione, M. Guarascio, A. Liguori, M. Zuppelli, Ensembling Sparse Autoencoders for Network Covert Channel Detection in IoT Ecosystems, in: International Symposium on Methodologies for Intelligent Systems, 2022, pp. 209–218.

[29] N. Cassavia, L. Caviglione, M. Guarascio, A. Liguori, M. Zuppelli, Learning Autoencoder Ensembles for Detecting Malware Hidden Communications in IoT Ecosystems, Journal of Intelligent Information Systems (2023) 1–25.