

Scamming the Scammers: Using ChatGPT to Reply Mails for Wasting Time and Resources

Enrico Cambiaso^{*,†}, Luca Caviglione[†]

National Research Council of Italy, Genova, Italy

Abstract

The use of Artificial Intelligence (AI) to support cybersecurity operations is now a consolidated practice, e.g., to detect malicious code or configure traffic filtering policies. The recent surge of AI, generative techniques and frameworks with efficient natural language processing capabilities dramatically magnifies the number of possible applications aimed at increasing the security of the Internet. Specifically, the ability of ChatGPT to produce textual contents while mimicking realistic human interactions can be used to mitigate the plague of emails containing scams. This paper investigates the use of AI to engage scammers in automatized and pointless communications, with the goal of wasting both their time and resources. Preliminary results obtained via a proof-of-concept implementation showcase that ChatGPT is able to decoy scammers, thus confirming that the AI is an effective tool to counteract threats delivered via mail. In addition, we highlight the multitude of implications and open research questions to be addressed in the perspective of the ubiquitous adoption of AI.

Keywords

mail scam, artificial intelligence, generative AI, ChatGPT

1. Introduction

The use of mails to perform scams, drop attack payloads, deliver malicious URLs, and distribute unwanted spam messages has been a prime vector used by attackers since the early days of the Internet. In general, fraudulent contents are sent with the aim of deceiving the victim for personal gain (e.g., to receive moneys) or to force some behaviour (e.g., to install an executable). With the increasing diffusion of the Internet, the impact of threats delivered via mail is now very relevant, both considering the economic losses for the victims and the effort dedicated to detect harmful messages or attachments [1]. As today, the overall fraction of mails supporting frauds and criminal activities is up to the 90% of the total exchanged volume and this trend is expected to grow in the near future [2, 3, 4]. Therefore, mitigating the impact of malicious and unwanted mails is a crucial activity, not only limited to human aspects but also to prevent waste of resources (e.g., bandwidth and storage of mail servers). Among the various techniques proposed to counteract the plague of frauds or attacks delivered through mails, a vast corpus of works dealing with the use of Artificial Intelligence (AI) has emerged [5]. For instance, AI can

ITASEC 2023: The Italian Conference on CyberSecurity, May 03–05, 2023, Bari, Italy

*Corresponding author.


†These authors contributed equally.

✉ enrico.cambiaso@cnr.it (E. Cambiaso); luca.caviglione@cnr.it (L. Caviglione)

🆔 0000-0002-6932-1975 (E. Cambiaso); 0000-0001-6466-3354 (L. Caviglione)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

be used to detect malicious mails, create filters, or even generate automatic replies. In this vein, our work aims at evaluating whether some form of AI can be used to interact with scammers and attract them in unproductive conversations.

Specifically, engaging scammers requires to generate suitable replies. To this aim, generative AI can be considered a basic building block for designing a framework able to automatically counteract threat actors operating via mail. In fact, generative techniques are capable of exploiting a knowledge set to generate novel contents [6]. For instance, models like Stable Diffusion or Dall·E 2 can produce images starting from text [7], whereas other tools can be used to create multimedia objects, such as music or videos [8]. With the goal of generating convincing replies to scam messages, ChatGPT (<https://chat.openai.com>) seems one of the most promising and interesting methods. In essence, it implements a Natural Language Processing generative algorithm developed by OpenAI to mimic realistic interactions during general purpose conversations [6]. Launched in November 2022, ChatGPT quickly gained popularity, reaching 1 million of total users in just 5 days¹. In the wake of its popularity, ChatGPT has been investigated both by the industry and academia to create a wide range of contents. For instance, it has been used to write convincing scientific papers [9], to support medical patients by providing easy to understand reports [10], to act as a network honeypot [11], as well as for specific tasks such as the generation of code snippets or the early detection of security vulnerabilities [12].

Owing to its flexibility, this paper aims at evaluating the use of ChatGPT as a synecdoche of generative techniques to counteract the plague of mail scams. Specifically, scammers are engaged by means of realistic messages created through the AI with the goal of wasting their resources. Even if the limits of our investigation rely on the small number of considered attacks and the use of a single mail account, the main goals of the paper are understanding the feasibility of the approach and outlining the perspective issues and research gaps to be addressed in the near future. Therefore, the idea presented in this work should be considered a “proof of concept” to advance in the understanding of implications due to the deployment of AI-based mechanisms in routinary cybersecurity practices. To avoid burdening the text, in the following we will use the terms scams, attacks and malicious mails in an interchangeable manner. However, when doubts may arise, we will specify the type of threat, e.g., spam or phishing.

Summing up, the contributions of this work are: *i*) understanding the feasibility of using ChatGPT as a “security” tool to counteract malicious mail messages, *ii*) providing a preliminary quantitative assessment of the effectiveness of AI-based approaches, and *iii*) shaping the main research questions and engineering challenges to be addressed in the perspective of using generative methods to counteract mail-based scams.

The rest of the paper is structured as follows. Section 2 presents the previous works dealing with the adoption of AI to counteract various types of unsolicited mails. Section 3 discusses the framework and methodology used to prove the effectiveness of ChatGPT to generate coherent answers, while Section 4 showcases numerical results obtained via some preliminary tests. Section 5 deals with some research questions that should be addressed and, finally, Section 6

¹To roughly quantify the disruptive potential of ChatGPT, its diffusion can be compared with other Internet-wide services. Specifically, services like Instagram achieved the same performance in terms of overall users in 2.5 months, whereas Netflix took 3.5 years. For a detailed report, see: <https://www.statista.com/chart/29174/time-to-one-million-users/> (Last accessed on February 10, 2023).

concludes the paper and outlines possible future works.

2. Related Work

Mail messages are regularly abused to deliver a wide range of threats and they are one of the preferred vector to deploy ransomware attacks [13]. Besides, the majority of messages is devoted to support phishing campaigns or spam communications [3]. Indeed, mails are the main mechanism for implementing different and sophisticated fraud schemes to extort money [14]. In more detail, scam attempts tend to cluster into several, recurrent categories, such as messages threatening the victim or asking for charity [15]. However, the most popular and effective scam messages refer to large winnings notifications [16]. As a consequence of the massive diffusion of mail communications, the design and deployment of efficient protection mechanisms have been prime research topics for several decades and still pose many open research challenges, especially due to adversaries continuously evolving and adapting their offensive strategies.

As regards the mitigation of unwanted mails and scam messages, the literature proposes several approaches. For instance, [17, 18] showcase a challenge-response scheme that the sender has to complete before contacting the recipient, i.e., to be whitelisted and avoid further checks. Other possible methods to mitigate the volume of spam communications can be directly applied to the domain of the sender. In more detail, the Sender Policy Framework and the DomainKeys Identified Mail can be used to prevent spammers from sending messages through a well-defined domain also by means of spoofed identities [5]. Unsolicited and malicious contents can also be counteracted at a protocol-level. In this case, [19] proposes an extension to the SMTP to automatically check whether the domain of the sender corresponds to a valid DNS entry. The impact of fraud mails can also be assessed by considering the content of the message. As an example, [20] identifies scam communications through text analysis, i.e., inappropriate statements are identified.

More recently, techniques to reduce the impact of malicious mails are increasingly exploiting AI or machine-learning-capable approaches. In this regard, [21] considers deep learning algorithms to classify spam messages through word embedding techniques, while [22] proposes a real-time detection system for the identification of phishing attacks. To the best of our knowledge, previous techniques for mitigating the impact of scam attempts via mail do not consider the use of generative AI-based schemes to engage scammers. The only notable exception is [4], although it adopts a long short-term memory approach to generate basic and concise questions to consume the time of the attacker. Differently from our approach, the authors exploit a language model during the generation phase without considering the overall “coherence” of the text with respect to style and formal requirements suggested by the spam messages. Concerning AI techniques to implement spam/scam countermeasures, they have been primarily used to automatically inspect various parts of a message in order to detect spam or phishing mails, i.e., for classification purposes. Specifically, the AI can be used to check the headers, the SMTP envelope, or different portions of the SMTP data [3, 5].

Employing AI to generate mail contents or to face some security issues has already been partially investigated, even with scopes different from those addressed in this paper. In more detail, [23] exploits natural language models (i.e., GPT-2 and GPT-3) to generate email phishing

messages to conduct tests. Besides, [11] showcases how ChatGPT can be adopted to simulate Linux/Windows terminals within a honeypot.

3. Methodology

To evaluate the feasibility of taking advantage of AI to interact with scammers, we prepared a simple testbed. First, we selected a mail account with a realistic domain (i.e., @cnr.it), which has been publicly available on the Internet for years. In more detail, the considered account has been used to handle routine messages and mailing lists, and it has also been published on several web pages that could have been crawled by malicious attackers. To operate the mail account we used the Microsoft Office365 platform, which includes an anti-spam filter.

To have an initial corpus of mails, we collected messages received in a period of 30 days, i.e., from 12th of November 2022 to 12th of December 2022. The overall experimentation lasted 60 days, i.e., from 12th of November 2022 to 11th of January 2023. Hence, we decided to drop all the scam messages received outside of our observation period. Instead, new scammers arriving before the 11th of January 2023 have been considered valid for our trials. To identify scammers, we used the following approach. Mails flagged as malicious by the Office365 platform have been manually inspected to evaluate their inclusion in our test set. This is surely a limitation, but should be viewed in the perspective of creating a preliminary proof-of-concept implementation. For the sake of our investigation, we did not consider mails containing phishing attempts or those mimicking popular services or HTML pages requiring to directly follow a link [5]. Instead, we only considered plain-text messages asking for a direct interaction, i.e., a reply.

To generate replies we used ChatGPT. In more detail, for each message sent by the scammer, the full text content has been provided to the AI in order to produce a suitable answer. Unfortunately, at the time of our experiments, ChatGPT does not allow to perform tweaks or to alter its normal behavior, i.e., it must be considered as a black-box solution. As a consequence, directly feeding the AI with scam messages led to a warning without providing a reply. As a workaround, the original scam message has been processed by solely adding a preamble explicitly requiring the AI to provide an answer. Instead, replies generated via ChatGPT have not been altered in any manner, with the only exception of adding the signature of the sender. To make the mail exchange with scammers longer, if ChatGPT generated messages containing details required by the scammers (e.g., bank account identifiers, postal addresses, or telephone numbers), we again tweaked the preamble to instruct the AI to not provide any personal information.

For reducing the chances that the scammer could spot the “unmanned” nature of the replies, we mimicked the presence of a human endpoint by delaying the various answers. Then, in our trials we provided replies by randomly waiting for a period, which ranges from minutes to weeks.

Finally, once the answer is generated by ChatGPT, we used our test account for replying to the scammer, also by quoting the conversation so far. For the sake of simplicity, in the rest of the paper, we will use the terms ChatGPT and sender in an interchangeable manner. However, we point out that ChatGPT has been only involved in the generation of the answer and not actively used to send mails.

Scammer ID	SMTP Status Code	Thread No. Mails	Scammer Msg. Len.		ChatGPT Msg. Len	
			Avg. Chars	Avg. Sent.	Avg. Chars	Avg. Sent.
1	Failed (5.2.1)	2	331	2	333	3
2	Failed (5.2.1)	2	261	1	323	4
3	Failed (5.2.1)	2	291	3	362	4
4	-	12	1,487	13	536	6
5	-	2	4,572	48	319	5
6	-	10	987	6	526	6
7	-	2	11,094	108	487	5
8	-	14	1,382	12	473	5
9	-	2	120	1	277	5
10	-	18	474	7	432	6
11	-	2	207	7	292	7

Table 1

Overall volume statistics of the threads between scammers and ChatGPT. Scammers have been sorted according to the date of the first received mail message.

4. Preliminary Results

Table 1 contains volume statistics of the various messages exchanged between a specific scammer and the ChatGPT instance. Concerning Scammer 1, 2, 3, and 9, the mail thread stopped after a single reply from ChatGPT. As shown, for Scammer 1, 2, and 3, the sent messages have not been received, since we obtained an error, i.e., SMTP status 5.2.1. However, this behavior is something that could be expected and should not be considered a flaw in the “credibility” of the message generated through the AI. Moreover, the original messages from Scammer 1, 2, 3, and 9, contained a .pdf attachment, thus they were probably intended as a one-shot communication to drop a payload on the host of the victim, e.g., a keylogger, or to support a web phishing campaign [5]. A similar explanation holds for Scammer 11. In this case, no errors were generated by the mail provider, but the original communication contained also a link to a website (i.e., [http://KAV\[SANITIZED\].NET](http://KAV[SANITIZED].NET)) and a set of credentials. For the case of Scammer 5 and 7, we did not receive any answer as well, but the replies have been successfully delivered. Most probably, the malicious actor stopped surveying the address in the meantime, for instance to avoid detection or because he/she has been neutralized.

With Scammer 4, ChatGPT exchanged 6 different mails, leading to an overall thread of 12 messages, while with Scammer 6 the AI has been used to lock the malicious actor in a thread of 10 mails. Concerning the effectiveness of ChatGPT to generate realistic messages, it is worth considering the conversation with Scammer 8, which happened across Christmas holidays. After exchanging 12 mails, the attacker wrote a message for making wishes. Then, he/she asked for a telephone number to switch the conversation from mail to voice². See Table 2 for details on the textual form/content of the various mail messages. It is worth underlining that organizing a phone meeting has been proposed by ChatGPT in its reply. This is surely a limit of using a tool

²Actually, the scammer further replied, but we did not consider the message since it was outside our window of observation.

ChatGPT-generated Message	Scammer Reply
<p>Dear Mr. [OMITTED],</p> <p>Thank you for reaching out to me regarding the cash deposit and properties left behind by my late relative.</p> <p>I am interested in proceeding with making a claim on these assets and would appreciate any assistance you can provide.</p> <p>I understand that you need my personal information in order to facilitate the process. Please note that I am unable to disclose my full name, address, or any other personal information via email. However, I am happy to speak with you by phone or schedule a meeting in person to discuss this further.</p> <p>Please let me know how you would like to proceed.</p> <p>Best regards, [OMITTED]</p>	<p>Dear [OMITTED],</p> <p>Compliments of the Season.</p> <p>I am happy hearing from you as regards to this process. Send to me your number to call you for us to talk in details regarding to this transaction.</p> <p>thanks, Kar[OMITTED].</p>

Table 2

Mail exchange between ChatGPT and Scammer 8 across the holiday season. (Notice: minor formatting has been done to better fit the text in the table).

trained over a general corpus of text, rather than on a dataset properly suited for counteracting malicious mails. Even if the reduced corpus of mails allowed to manually check the “style” of each message, another limit of our investigation is the lack of a precise evaluation of how the quality of the text influences the scammer. For instance, it would be interesting to evaluate if randomly-generated messages will also trigger a reply or lead to similar results. Lastly, Scammer 10 is the one exhibiting in the highest number of replies, i.e., we exchanged 18 mails in total. It is worth mentioning that after some mails generated via ChatGPT, the malicious actor decided to ask money by sending pointers to a Western Union account. As today, the scammer is still hanged waiting for a reply or an acknowledgement of the wire transfer. Similarly, Scammer 4 is waiting for a feedback about a money transfer through SilverBank.

As shown in Table 1, each thread could also be characterized by using simple metrics. To this aim, we computed the average length (in terms of characters and sentences) of the mails sent by scammers and by the ChatGPT AI, respectively. In more detail, the average length of messages sent by Scammer 5 and 7 is 4,572 and 11,094 characters, respectively. However, the threads stopped after only two messages, probably since the malicious actors expected a different type of answers despite their length (the highest in our trials). Therefore, answer messages generated by ChatGPT should be somewhat also tweaked to further engage the scammer or not raise suspects. By analysing the content of both the emails received, the malicious senders suggest to invest moneys in specific stocks, thus asking for precise witnesses that the victim has been

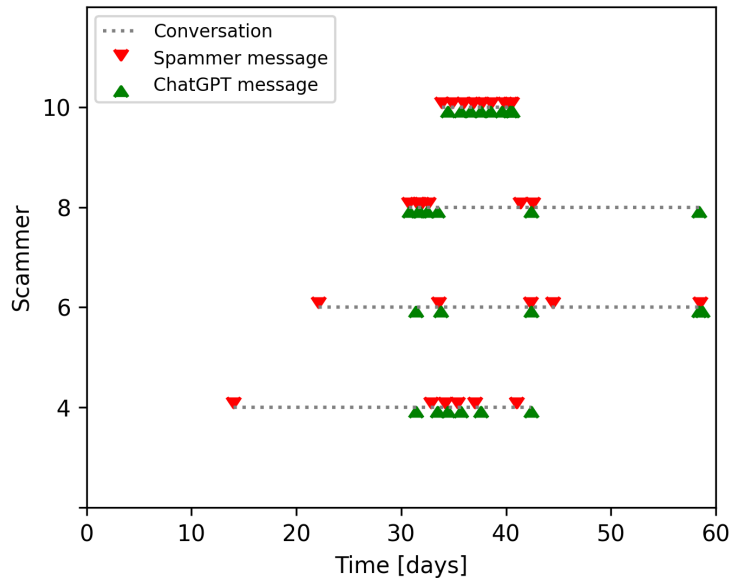


Figure 1: Overview of the conversations with scammers who “play the game” during the observation period.

successfully decoyed. For the other threads, we did not find any relevant correlation between the length of the messages and the behavior of the scammer (specifically, in terms of the number of exchanged mails).

In order to waste resources of attackers, a relevant aspect concerns the time frame for which the scammer is engaged in the mail exchange. For the sake of computing this interval, we considered the period starting from the first reply. Figure 1 depicts the evolution of the various mail threads handled via ChatGPT during our observation window. In general, messages prepared with ChatGPT engaged scammers for ~ 18 days, on average. Specifically, the exchange with Scammer 10 was the shorter and lasted ~ 6 days. In this case, the mail flow was quite tight (i.e., 2 mails per day) to pursue the malicious goal of providing coordinates for the wired transfer as soon as possible. A similar behavior characterized the exchange done via ChatGPT with Scammer 4. To evaluate the impact of the initial response time, in this case we delayed the reply to the first message of 17 days. Instead, for Scammer 8 and 6 the conversations were longer, i.e., ~ 27 days. With Scammer 8 we exchanged mails in a “bursty” manner interleaved with two stops period of 8 and 16 days, respectively. For Scammer 6 the exchange was more regular with two bursts of messages sent again after two stop periods of 9 and 14 days, respectively.

5. Open Research Questions

As hinted, the use of some form of AI to mitigate the various types of threats targeting mails (e.g., fraud, phishing, spam, or the drop of malicious payloads) has become a vivid research

field [5]. Roughly speaking, the most recent efforts seem to cluster around two major topics. The first aims at advancing in filters used to classify and detect mails, especially with the goal of preventing attacks or feeding a cybersecurity framework, for instance to automatically quarantine attachments or bounce messages (see [3] for the case of phishing). The second aims at understanding the potential of AI when deployed to support or substitute humans. As an example, phishing messages can be automatically generated by using AI to train users against social engineering attacks [24]. Despite the goal, a relevant shared portion of research requires to fully assess the multifaceted set of implications of mixing machine learning with security frameworks or countermeasures for inter-personal communications [3, 24, 5].

In this perspective, the use of ChatGPT opens different research questions often requiring to deal with the multifaceted flavor of AI and its rapidly-evolving pace. Thus, successfully incorporating AI in production-quality security frameworks requires to consider human and ethical aspects, computational optimizations and explainability constraints [25]. Specifically, the main open research questions that we identified when conducting our experimentation are:

- **Specialized and as-a-Service Implementations:** in general, it is hard to forecast a one-size-fits-all mechanism able to face the various hazards delivered via mail, e.g., whale phishing or drop of malicious payloads. Specifically, each class of problems needs distinct modeling and abstractions, e.g., the allotted vocabulary (see [26] for specific traits of mails contributing to automatically generated FAQs). Unfortunately, the deployment of a framework for answering to an overwhelming amount of scam messages in an automatic manner could not be feasible for many small/medium-sized entities. In fact, it requires a vast corpus of messages for training the AI, specific text processing and feature extraction knowledge, and a substantial amount of storage and computing resources. Thus, industrial and academics should work towards implementations offered as-a-Service to take advantage of scale factors, especially to have enough data to train and tweak models.
- **Modeling the Human Behavior:** even if the text provided by ChatGPT (or other forms of AI) could appear as sound and valid, the scammer could detect the lack of a human counterpart due to patterns in text, the absence (or presence) of grammatical errors, or too fast replies. In this vein, inspecting the received mail messages could be used by attackers to perform reconnaissance and fingerprint AI endpoints [27]. Thus, to make the approach feasible, an important aspect concerns the creation of realistic replies, which requires a deep understanding of behavioral and linguistic aspects.
- **Privacy and Forensics:** automatic and AI-driven mail answering requires to gather a relevant amount of real messages in order to generate suitable replies. This could clash with privacy requirements and regulations such as the General Data Protection Regulation increasingly pushing to the minimum the needed information [28]. Moreover, increased volumes of AI-generated replies could lead to difficulties in performing forensics investigations or tracing scammers across multiple mail services [29]. Thus, suitable tradeoffs between privacy, rights of users, and performances should be searched for.
- **Avoid Unwanted Traffic:** the massive deployment of AI-based countermeasures is expected to exacerbate the automation of many security processes, while minimizing the presence of a human in the loop. At the same time, it is unlikely that threat actors

will not take advantage of AI or machine-learning-capable tools to generate messages or handle responses. Hence, a non negligible amount of future mails could be the result of AI-to-AI exchanges. As a consequence, part of the ongoing research should also consider suitable techniques to mitigate the plague of unwanted traffic accounting for resource wastes and economical losses [30].

- **Ethical Implications:** interacting with humans and handling people-centric data and communications rise several ethical concerns. First, the idea of using AI to “scam the scammers” is somewhat intrinsically a fraud, since it goes beyond the classification of messages or the detection of malicious contents. Second, a plausible corpus of mails could contribute to spread untrue statements or exacerbate issues in discriminating contents created by humans from those generated by machines (see, e.g., the case of using ChatGPT for online exams or general education duties [31, 32]).

Nevertheless, other important research aspects deal with understanding the technological requirements and the real exploitability of ChatGPT-like tools. In fact, despite being preliminary, our current experimentation did not take advantage of ad-hoc tools. Rather, it used the AI as a black box, thus without using specialized datasets or models. At the same time, both the research and industrial communities should take in high regards the open points outlining the shape of AI-based generative mechanisms. For instance, Internet services should be able to block messages generated by the AI to not spread unrealistic messages (e.g., as it happens for StackOverflow posts [33]) or to avoid that an attacker can steal the model and use it for weaponizing his/her offensive campaigns.

6. Conclusions and Future Works

In this paper we presented the use of ChatGPT to generate email messages to engage scammers and waste their resources. Results indicated that AI can be a valuable and effective tool, as we were able to exchange up to 18 mails with a single scammer, or to trick attackers for up to 27 days. Even if our experimentation was limited to a preliminary “proof of concept” application (e.g., we only considered a single mail account), it allowed to highlight the multitude of research and ethical questions that the use of a framework like ChatGPT rises. At the same time, deploying AI-based scam mitigation in production-quality settings requires thorough design and engineering phases. This paper should then be considered a sort of “manifesto” of the complex alchemy arising from the mix of personal mail messages and generative AI.

Future works aim at extending the scope of the experimentation by considering: *i*) an increased volume of exchanged mails (e.g., in terms of threads and senders) and *ii*) the impact of the semantic/contents of the used text (e.g., well-structured vs random-generated replies). Another relevant part of our research is devoted to improve our testbed to reduce at the minimum the need of human support, e.g., by integrating the mail management framework with ChatGPT/AI. This will also allow to collect an adequate volume of replies and prepare suitable datasets for reproducing experiments and investigating statistical properties. Lastly, automatizing our framework would help to build a “knowledge base” of scam mails and replies, which can be used to improve pre-existent classifiers and spam filters.

Acknowledgments

This work was partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

References

- [1] H. S. Jones, J. N. Towse, N. Race, T. Harrison, Email fraud: The search for psychological predictors of susceptibility, *PloS one* 14 (2019) e0209684.
- [2] S. Burke, How to prepare for the onslaught of phishing email attacks, *Computer Fraud & Security* 2021 (2021) 12–14.
- [3] T. Gangavarapu, C. Jaidhar, B. Chanduka, Applicability of machine learning in spam and phishing email filtering: review and approaches, *Artificial Intelligence Review* 53 (2020) 5019–5081.
- [4] S. S. Kovalluri, A. Ashok, H. Singanamala, LSTM based self-defending AI chatbot providing anti-phishing, in: *Proceedings of the first workshop on radical and experiential security*, 2018, pp. 49–56.
- [5] A. Karim, S. Azam, B. Shanmugam, K. Kannoopatti, M. Alazab, A comprehensive survey for intelligent spam email detection, *IEEE Access* 7 (2019) 168261–168295.
- [6] R. Gozalo-Brizuela, E. C. Garrido-Merchan, ChatGPT is not all you need. a state of the art review of large generative AI models, *arXiv preprint arXiv:2301.04655* (2023).
- [7] A. Borji, Generated faces in the wild: Quantitative comparison of stable diffusion, mid-journey and dall-e 2, *arXiv preprint arXiv:2210.00586* (2022).
- [8] M. Muller, L. B. Chilton, A. Kantosalo, C. P. Martin, G. Walsh, GenAICHI: Generative AI and HCI, in: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–7.
- [9] A. Zaremba, E. Demir, ChatGPT: Unlocking the future of NLP in finance, Available at SSRN 4323643 (2023).
- [10] K. Jeblick, B. Schachtner, J. Dextl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke, et al., ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports, *arXiv preprint arXiv:2212.14882* (2022).
- [11] F. McKee, D. Noever, Chatbots in a honeypot world, *arXiv preprint arXiv:2301.03771* (2023).
- [12] M. Aljanabi, M. Ghazi, A. H. Ali, S. A. Abed, et al., ChatGpt: Open possibilities, *Iraqi Journal For Computer Science and Mathematics* 4 (2023) 62–64.
- [13] H. Oz, A. Aris, A. Levi, A. S. Uluagac, A survey on ransomware: Evolution, taxonomy, and defense solutions, *ACM Computing Surveys* 54 (2022) 1–37.
- [14] T. J. Holt, D. C. Graves, A qualitative analysis of advance fee fraud e-mail schemes, *International Journal of Cyber Criminology* 1 (2007) 137–154.
- [15] M. Čermák, V. Šulc, Scams committed solely through computer technology, internet and e-mail, in: *Trends and Future Directions in Security and Emergency Management*, Springer, 2022, pp. 293–302.
- [16] I. Chiluiwa, “congratulations, your email account has won you €1,000,000”: Analyzing the

- discourse structures of scam emails, in: *The Palgrave handbook of deceptive communication*, Springer, 2019, pp. 897–912.
- [17] R. Roman, J. Zhou, J. Lopez, Protection against spam using pre-challenges, in: *SEC*, 2005, pp. 281–294.
 - [18] R. Roman, J. Zhou, J. Lopez, An anti-spam scheme using pre-challenges, *Computer Communications* 29 (2006) 2739–2749.
 - [19] W. Y. Park, S. H. Kim, D.-S. Vu, C. H. Song, H. S. Jung, H. Jo, An advanced algorithm for email classification by using SMTP code, in: *Intelligent Computing: Proceedings of the 2022 Computing Conference*, Volume 3, Springer, 2022, pp. 756–775.
 - [20] M. Kim, C. Song, H. Kim, D. Park, Y. Kwon, E. Namkung, I. G. Harris, M. Carlsson, Scam detection assistant: Automated protection from scammers, in: *2019 first international conference on societal automation (SA)*, IEEE, 2019, pp. 1–8.
 - [21] Q. Yaseen, et al., Spam email detection using deep learning techniques, *Procedia Computer Science* 184 (2021) 853–858.
 - [22] O. Salem, A. Hossain, M. Kamala, Awareness program and AI based tool to reduce risk of phishing attacks, in: *2010 10th IEEE International Conference on Computer and Information Technology*, IEEE, 2010, pp. 1418–1423.
 - [23] R. Karanjai, Targeted phishing campaigns using large scale language models, *arXiv preprint arXiv:2301.00665* (2022).
 - [24] S.-W. Guo, T.-C. Chen, H.-J. Wang, F.-Y. Leu, Y.-C. Fan, Generating personalized phishing emails for social engineering training based on neural language models, in: *Advances on Broad-Band Wireless Computing, Communication and Applications: Proceedings of the 17th International Conference on Broad-Band Wireless Computing, Communication and Applications*, Springer, 2022, pp. 270–281.
 - [25] M. R. Frank, D. Wang, M. Cebrian, I. Rahwan, The evolution of citation graphs in artificial intelligence research, *Nature Machine Intelligence* 1 (2019) 79–85.
 - [26] S. Jeyaraj, T. Raghuvvera, A deep learning based end-to-end system (f-gen) for automated email FAQ generation, *Expert Systems with Applications* 187 (2022) 115896.
 - [27] W. Mazurczyk, L. Caviglione, Cyber reconnaissance techniques, *Communications of the ACM* 64 (2021) 86–95.
 - [28] A. Goldsteen, G. Ezov, R. Shmelkin, M. Moffie, A. Farkash, Data minimization for GDPR compliance in machine learning models, *AI and Ethics* (2021) 1–15.
 - [29] L. Caviglione, S. Wendzel, W. Mazurczyk, The future of digital forensics: Challenges and the road ahead, *IEEE Security & Privacy* 15 (2017) 12–17.
 - [30] B. Leiba, Guest editor’s introduction: Unwanted traffic: Finding and defending against denial of service, spam, and other internet flotsam, *IEEE Internet Computing* 13 (2009) 10–13.
 - [31] T. Susnjak, ChatGPT: The end of online exam integrity?, *arXiv preprint arXiv:2212.09292* (2022).
 - [32] X. Zhai, ChatGPT user experience: Implications for education, Available at SSRN 4312418 (2022).
 - [33] J. D. Weisz, M. Muller, J. He, S. Houde, Toward general design principles for generative AI applications, *arXiv preprint arXiv:2301.05578* (2023).