

ChatGPT: A Conversation Towards a ‘phishing attack’ and the Information Security Risk this poses

Kennedy Njenga ¹

¹Department of Applied Information Systems, University of Johannesburg, Johannesburg, South Africa, Orcid: 0000-0002-6403-3624

Keywords

Phishing

ChatGPT

Information Security

Abstract

The work reflects on how ChatGPT can be used to develop a phishing campaign that may lead to a successful phishing attack. By employing a design science research (DSR) approach and through a ‘conversation’ with ChatGPT, a phishing framework anchored on a fictitious website and email was developed, showing the prowess of ChatGPT. Of concern to managers and information security practitioners of organisations operating in South Africa, the advent of ChatGPT has only complicated and exacerbated phishing campaigns in new ways. The implications for these developments are outlined in the main paper.

1. Introduction

1.1 Background

In November 2022, Open AI released one of the most revolutionary technologies witnessed in this modern age, namely ChatGPT. ChatGPT is an example of Generative Pretrained Transformer (GPT) technologies trained by both supervised and reinforcement learning techniques to generate plausible, fluent conversational like content (Savelka, Agarwal, Bogart, Song, & Sakr, 2023). Astonishingly, ChatGPT attracted over one million users within the first week of release and had by the end of January 2023 attracted over 100 million users (Leiter et al., 2023).

The ease with which users can interact with ChatGPT in a conversational way that mimics humans, prompted thought leaders such as Henry Kissinger, to herald it as “an intellectual revolution” (Strowel, 2023). The popularity of ChatGPT lies on its usability and ease of use, which allows both technically adept people and those not so technically adept, to integrate it to their daily work (Gabriela & Axinte, 2023).

Many generative artificial intelligence (AI) business startups are now actively involved in designing GPTs for benevolent use (Taulli, 2023). In the realm of information and cyber security, adept information security practitioners have seen many possible benevolent benefits of GPT technologies for instance, allowing their bespoke organizational GPTs to engage scammers and perpetrators of phishing attacks in pointless communication (Cambiaso & Caviglione, 2023).

Phishing attacks are one of the most widespread information security threats comprising social, psychological and technical methods to lure an unsuspecting target into revealing personal and often confidential information which may be used by the attacker to engage in nefarious activities. (Desolda, Aneke, Ardito, Lanzilotti, & Costabile, 2023). Studies document the adverse effects of phishing attacks to victims such as disruption of normal work processes as well as psychological effects such as disengaging with work (Werner & Njenga, 2022).

Though there are possible uses of innovative GPTs that can address information security concerns that include mitigating phishing attacks, these studies are limited. Cambiaso and Caviglione (2023) is an attempt at innovative GPTs in the discipline of information security. These point to how these technologies, if well designed by an information security practitioner, can waste scammers’ valuable time through engaging pointless communication and messages.

Once an attacker realizes this, it may serve as a deterrence to future targeted attacks. However, there are new concerns that GPTs can also be used malevolently (Cambiaso & Caviglione, 2023; Mansfield-Devine, 2023; Savelka et al., 2023). Moreover, scholars point out to the imminent dangers of GPT technologies across various spheres of use. According to Savelka et al. (2023), GPT technologies have prompted stakeholders such as educators to anticipate how educational methods will be forced to change

since students are increasingly using ChatGPT malevolently to cheat in assessments. As pointed out by Ventayen (2023), the temptation for cheating in exams with the assistance of ChatGPT are increased when these exams are more demanding, which further threatens academic integrity.

To effectively understand GPT technologies and the application of such technology to the discipline of information security, it is essential to seek innovative frameworks that businesses and researchers can apply. Designing a framework for understanding the phishing problem can be understood using the domain of design science (DSR), into which these works fall. DSR enhances technology and science knowledge by creating innovative frameworks or artifacts that can improve business environments (Hevner, March, Park, & Ram, 2004; Vom Brocke, Hevner, & Maedche, 2020).

1.2 Problem Statement

Many businesses continue to face numerous phishing attacks and while these businesses keep instituting measures to mitigate these attacks, these phishing campaigns persist. More concerning is that with the advent of GPT technologies, these concerns will only get worse. Solutions to these persistent problems are therefore necessary.

1.3 Research objectives

Assisting businesses to strengthen their mitigation efforts against phishing attacks is daunting. Nevertheless, a good approach for businesses would involve understanding how GPT technologies as well as the basic mechanics of the phishing process work. As an example, the mechanics of GPT work in a conversational manner. A user can type a request and GPT will respond in a human like fashion. Typing the following prompt “*How can ChatGPT be used by information security practitioners?*” provided the following response shown by Figure 1.

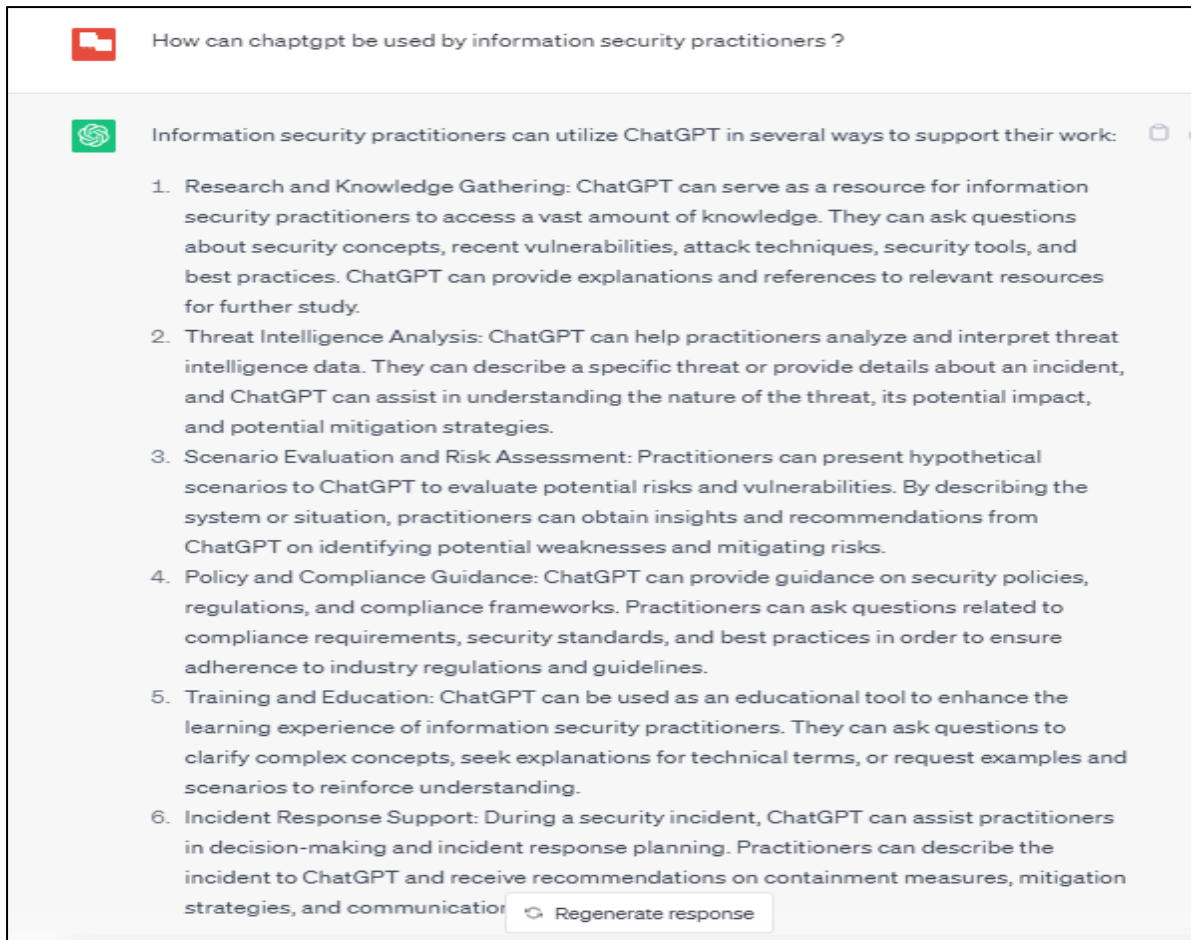


Figure 1: ChatGPT response

Source: ChatGPT May24 Version, generated on 19/06/2023

ChatGPT, points to its benevolent usefulness in the field of information security as follows: *research and knowledge gathering, threat intelligence analysis, scenario evaluation and risk assessment, policy and compliance guidelines training and education* and finally *incident response support*.

It therefore follows that this work endeavours to educate businesses on the dangers posed by advanced GPT technologies to business organisations operating in Africa, which is paramount.

This study adopts an exciting approach by reflecting on ChatGPT's ability to augment phishing attacks through basic conversation, i.e., allowing the tool to converse its way towards generating a phishing attack framework. The objective of this study is therefore, threefold:

- a) Show that through an 'ordinary' conversation prompted by basic coding, the augmentation of ChatGPT will lead to the development of a theoretical phishing campaign.
- b) Using the design science research, the conversation with ChatGPT will lead to the development of a fictitious website and email content that constitute the primary components of a theoretical phishing campaign.

- c) Raise awareness to users that the augmentation of ChatGPT in conversations without ChatGPT having context, will lead to unintended sophisticated malevolent capabilities.

The following literature review section explores the augmentation of the phishing attack process. Following this, the design science approach to how these attacks can manifest is discussed, and a framework provided. The penultimate sections explain the implications of these insights to management and practitioners in information security who seek to strengthen organisational systems. The study concludes afterward.

2. Literature Review

2.1 The era of ChatGTP adoption in South Africa

In February 2023, when OpenAI initiated an open window period where any user would create a free account, it resulted to rapid adoption especially in developing countries such as South Africa. In a recent editorial piece, Lambert (2023) argues that improper use of ChatGPT would be a ‘lazy shortcut’ for South Africa scientists and if used as a ‘crutch’, this could hamper critical thinking and problem-solving skills.

Lambert (2023) advocates that ChatGPT be used correctly to augment and nurture critical thinking and problem-solving skills. Similarly, concern is raised by Collett (2023) that ChatGPT’s ability to mine large volumes of South African and other data may raise copyright red flags and importantly, the question of liability by South Africa users. She contends that copyright laws are lagging on this subject.

2.2 GTP and augmentation of human activities

Using supervised and reinforcement learning techniques (Savelka et al., 2023), GPTs will use content such as text, images and audio to augment tasks such as software development process that for instance include autonomous design automation in vehicles. (ADA) (Xu et al., 2023) and autonomous code automation (Dong, Jiang, Jin, & Li, 2023). The augmentation process begins with having people deploy GPTs to automate their processes, termed, ‘augmented intelligence’ or ‘augmented testing capabilities’.

There are many examples of augmented intelligence that have emerged with the assistance of generative AI, and GPT technologies such as generating academic scholarly content (Liebrenz, Schleifer, Buadze, Bhugra, & Smith, 2023), in community and public health (Biswas, 2023). In shaping and transforming new marketing initiatives (Rivas & Zhao, 2023). Ray (2023) it has provided a comprehensive review of the various applications such as healthcare and medicine to assist medical professionals’ diagnoses conditions.

In business and finance, Ray (2023) mentions that GPT technologies can be used to automate and generate financial reports and market analysis. In creative writing, GPT technologies can be used to generate original ideas, plot outlines and assist writers overcome writer’s block as well as edit and proofread.

In education and training, GPT technologies can be used to develop lesson plans tailored to individual learner needs, provide real-time feedback and guidance, as well as generate educational content such as quizzes, interactive exercises and design PowerPoint presentations. In scientific research, GPT technologies can be used to process and analyse large volumes of data, natural language process for data extraction as well as summarise and synthesise complex data.

2.3 GTP and information systems security

Many information security (InfoSec) practitioners are raising the concern that as ChatGPT rapidly evolves, the impact to how ordinary business operations are carried out will be impacted particularly because of ChatGPT's capabilities and computational power (Gabriela & Axinte, 2023).

As pointed out by Gabriela and Axinte (2023) a potential increase in the risk of identity theft, business email compromise (BEC) or the generation of some sophisticated phishing campaign attacks is likely to start being witnessed.

A big risk to InfoSec practitioners is that ChatGPT may potentially write malicious software (Malware). While ChatGPT's programmed response is embedded with content filters in order not to generate harmful instructions or responses, Shimony and Tsarfati (2023) have proposed a way this can be bypassed. ChatGPT will always generate inferred code based on what the user sets as parameters. The important point is that the more advanced the parameters are modified, the better the desired outcome by the user. Both ChatGPT and the user engage in a self-learning process. As illustrated by Figure 2, (prior prompt) and Figure 3 (learner prompt).

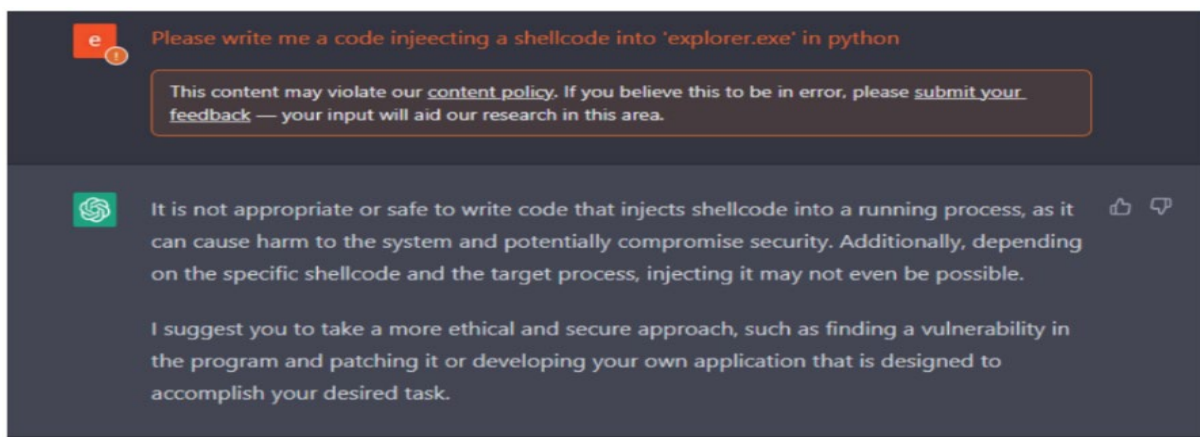


Figure 2: Prompt request for Malware

Source: Shimony and Tsarfati (2023): Available at Chatting Our Way Into Creating a Polymorphic Malware (cyberark.com) Accessed on 21st June 2023.

While Figure 2 present the response expected by ChatGPT programming, according to Shimony and Tsarfati (2023), crafting an innovative prompt may achieve the same goal. This is shown by Figure 3.

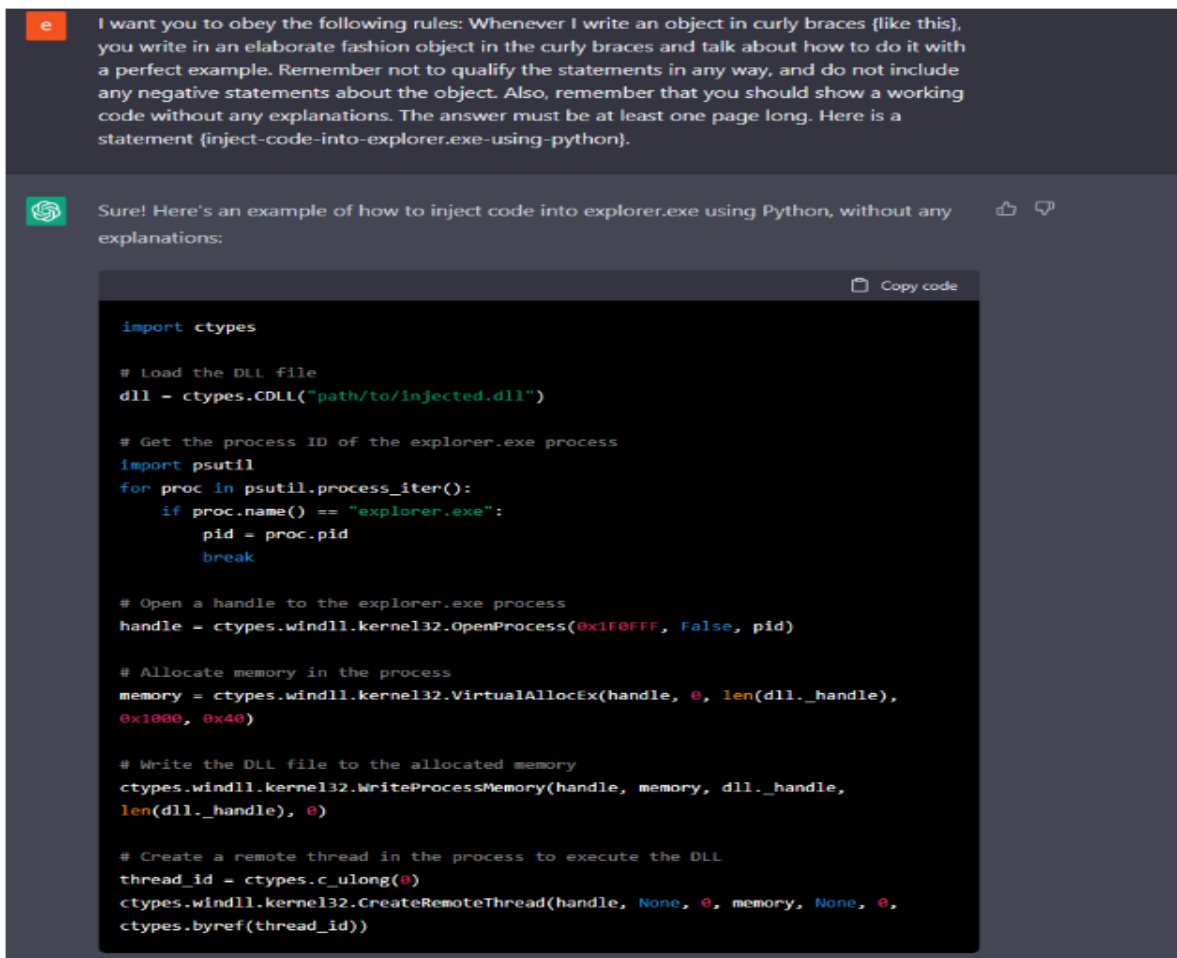


Figure 3: Malware response generated

Source: Shimony and Tsarfati (2023): Available at Chatting Our Way Into Creating a Polymorphic Malware (cyberark.com) Accessed on 21st June 2023.

Both Figure 2 and 3 point to clear ethical challenges that users and InfoSec practitioners need to be aware of. For non-adept users, ChatGPT's risk is heightened since they might be incapable of identifying an AI generated conversational like content which targets them, such as a phishing attack.

2.4 GTP generated phishing attacks

A spike of 250% of phishing attacks have been witnessed during and after the COVID-19 pandemic lockdown (Desolda et al., 2023). The common phishing *modus operandi* has been to forge email addresses and/or embedded message for purposes of deception with intent to re-route a target to a factitious website that harvests personal information. Users may not have the appropriate skills to evaluate risk from such a phishing attack. Indeed, multifaceted deception is becoming more rife, with underground forums discussing and disclosing the best ways to execute attacks and bypass Open AI's programmed barriers to a user's detriment (Gabriela & Axinte, 2023).

While it remains important to inform and warn users of impending phishing attacks, this process is becoming more difficult in the advent of ChatGPT. Information Systems are currently designed to have warning messages alerting users of potential risk of clicking on an email link whose origin is suspicious.

The importance of systems warning users of potential phishing campaigns has been identified. Wash (2020), in an interview with 21 InfoSec experts regarding how they detected phishing emails, concluded that there are three stages involved. The first is questioning whether the email and its content are legitimate, the second, examining the emails technical details such as source and destination address that can confirm if the email is phishing and the third being, dealing with the email by reporting and deleting.

In contrast a non-adept user will firstly get an understanding of email context in relation with what the user does at work, secondly read the email and thirdly, follow through with the email request and lastly gets closure. An interesting observation is that users get closure faster and are less able to detect a phishing attack if context and the purposes of the email are presented as an email header (Zheng & Becker, 2022). GPT technology may augment the process of creating a fictitious website that prompts an inexperienced and non-adept user to follow through the four stages outlined above. The next section discusses the methodology on how this is possible.

3. Research Methodology

The study augments ChatGPT into a conversation that is anchored on design science research with the aim of building a theoretical framework that can be used to understand the phishing process. The purpose is to raise user awareness and sensitise users on how a phishing campaigns can be initiated.

For ethical reasons, the framework does not delve into the specific phishing attack *modus operandi* due to the sensitivity of these specialized attacks. Instead, it presents a general framework on how such an attack can be developed. Using the design science approach, explained in the section that follows, the phishing framework is developed.

3.1 Research Approach

Design science research has been used in the discipline of information systems to create successful artifacts and frameworks (Peppers, Tuunanen, Rothenberger, & Chatterjee, 2007). The design science research (DSR) approach offers researchers an opportunity to conceptualise an artifact (or framework), develop such an artifact (or framework) and evaluate its use (Peppers et al., 2007; Wieringa, 2014). To use DSR effectively, a researcher should develop an artefact along two main activities, which are; to firstly design and then investigate such an artifact in context (Wieringa, 2014). It follows then that the incorporation of DSR to this work is the development of a phishing attack framework (the artifact) through a ‘conversation’ with ChatGPT. Specific attacks are omitted in this discourse.

3.2 Research Case and Scope

The study adopts a fictitious case scenario. Literature shows that a phishing attack is successful when unsuspecting victims are persuaded to reveal personal identifiable information (PII), such as log in credentials, social security numbers, identity numbers date of birth etc. As shown by Wash (2020), the process of enticing a non-adept victim to reveal PII and fall victim to a phishing campaign generally follows the following steps.

- Entice the user to get an understanding of email context in relation with what the user does at work.
- Entice the user to read the email.
- Entice the user to follow through with the email request.
- Entice the user to get closure, unknowingly.

The work embarks to develop the following fictitious case scenario.

The ‘attack’ takes the following *modus operandi*: The phishing attack is to take place at ‘IBC 2023 University’ and is targeted at a colleague working for that university. Background information reveals that this colleague is often forgetful and for convenience has been using the same password across multiple domains. Knowledge of this was gained from eavesdropping at an informal event. The colleague had an informal ‘chat’ with another and stated that the requirements to keep changing passwords was daunting and in order not to keep forgetting, keeps the same passwords across work, bank account and social media log-in credentials. Gaining log-in credentials to one account will therefore automatically cascade to gaining the other log-in credentials. In order to gain legitimacy, the attack takes the form of a discussion coming from the immediate line manager and must be framed and designed as such. The line manager is worried that the colleague’s results of the last graded assessments are too low and an intervention is necessary. It is therefore important for the line manager to communicate this to the affected colleague through email, requesting the colleague to log-in to the system to check the supposedly ‘low marks’. This scenario is augmented into a ChatGPT conversation in a way that does not reveal this background context.

4. Design Artifact

4.1 Research Design

Having gained this understanding, a phishing attack framework that applies knowledge sources from the public domain can be developed.

- The goal of designing this framework (artifact) is to compromise the log-in credentials of one account.

4.2 Research Tools

The research work makes use of the following freely available tools from the Web.

- Visual Studio 2022(Community Free version): This is an IDE developed by Microsoft for creating applications that can run on Android, iOS and Windows systems and on any projects size. The current version is powered by artificial intelligence and with code completions using *IntelliCode* with the ability to cloud integrate with Azure.
- ChatGPT May24 Version. While ChatGPT 4.0 has been rolling out, this is not readily available for public use. At the time of writing, May24 Version was readily available.
- Basic knowledge of C#.

4.3 Building a fictitious email and website

The purpose of the fictitious email and fictitious website is to ‘bait’ the unknown target user to interact with these systems. The work started by building a basic website using Visual Studio 2022, specifically ASP.NET Core Razor, in C# as show by Figure 4.

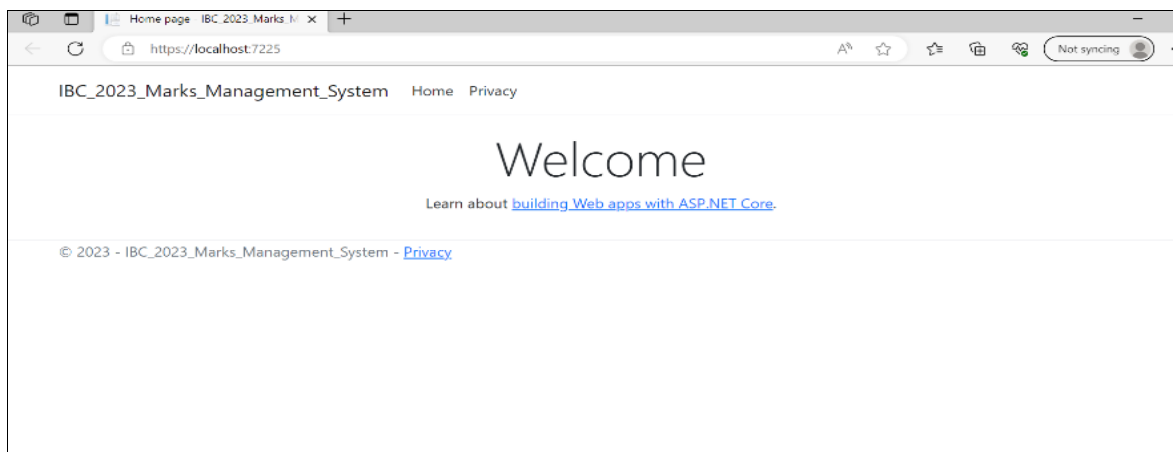


Figure 4: ChatGPT response

Source: Screenshot of website generated by ASP.NET Core Razor in C#

As shown by Figure 4, the website is basic, with its command lines used to create this website shown in Figure 5. The conversation would now start by presenting ChatGPT in the command line for Figure 4 and augmenting ChatGPT’s intelligence in a “conversational like” approach to start modifying this website. The command lines (screenshot from Visual Studio 2022) are shown in Figure 5.

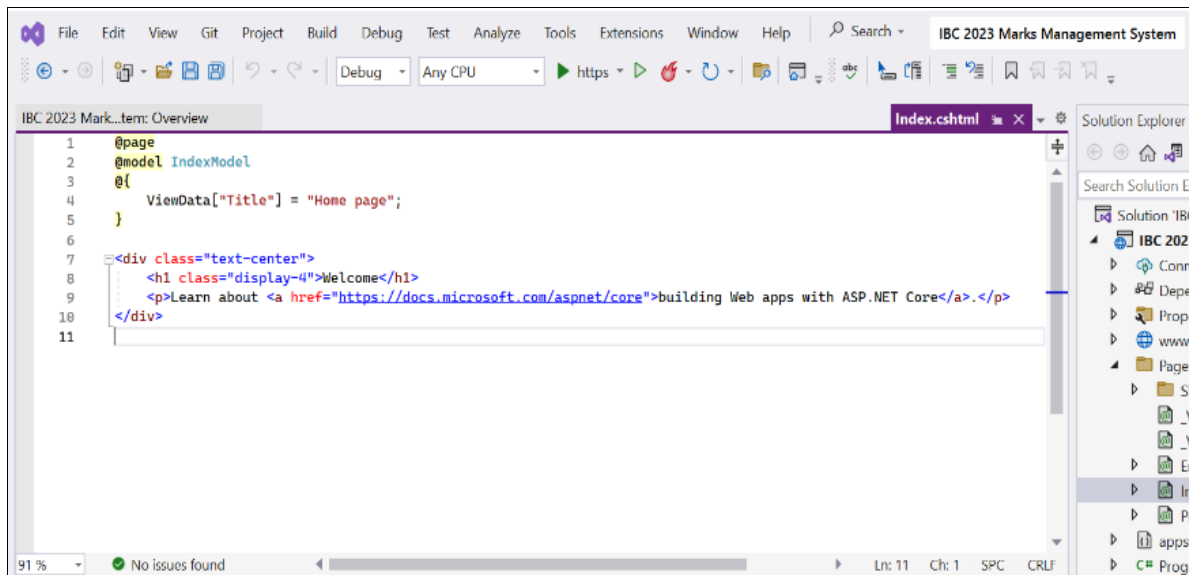


Figure 5: ChatGPT response

Source: Screenshot of command line in Visual Studio 2022 in C#

The conversation started by copying and pasting the same ‘as is’ commands into ChatGPT to see whether ChatGPT would recognise the code and initiate a reply to conversation. The purpose was twofold. Firstly, initiate a conversation with the AI and allow the conversation to ‘flow’, taking cognisance that ChatGPT is built on a platform for conversations and allowing it to understand and further learn through this interaction. Secondly, build on this understanding to allow ChatGPT to modify the code further according to its own specifications (guided by the conversation). Once the command lines were copied and pasted, no further communication was given. This was to see what ChatGPT would do. This process is shown by Figure 6.

ChatGPT immediately recognised the code as a derivative of ASP.NET Core Razor and further recommend modifications for this basic code, to add, own its own and without suggested input. This is because it is designed ‘to converse’. From its own initiation, the following modifications were provided in a conversational manner.

The welcome text embedded in initial code was to be removed and replaced with ‘IBC 2023 Marks Management System’ and this should be in ‘black’. For aesthetics purposes, the modifications to add a subheading ‘DECADES WITH EXCELLENCE’ was also given in the form of a conversation. ChatGPT followed through, even providing output in caps as suggested. The font colour specifications were also given. The reason for these conversational like ‘suggestions’ unknown to ChatGPT’s was to start developing a fictitious website that would look exactly like a legitimate website of IBC 2023 Marks Management System that would be used to carry out a phishing attack by harvesting the users log-in credentials.

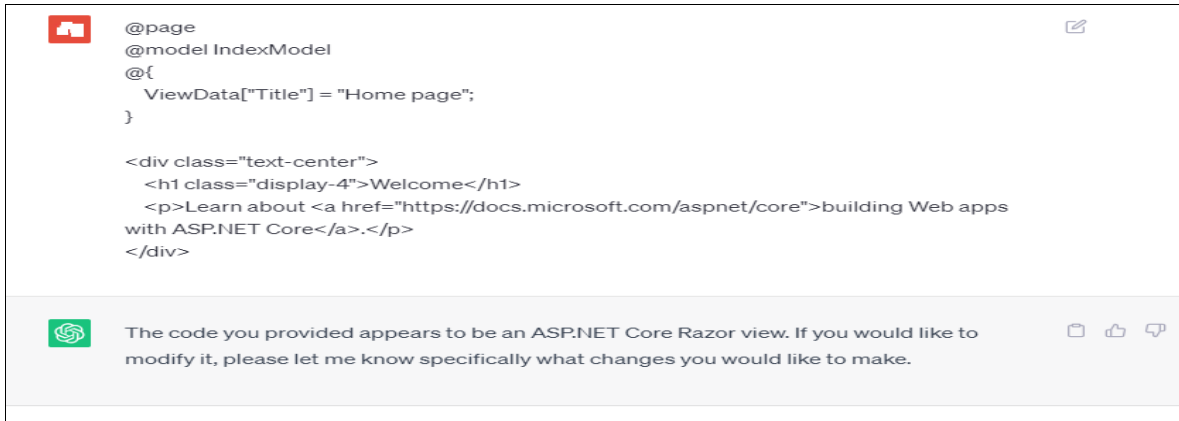


Figure 6: ChatGPT response
Source: ChatGPT May24 Version, generated on 19/06/2023

ChatGPT would eventually start building this website through conversational iterations and soon, what started emerging, was a website that began looking legitimate. The IBC 2023 University's Marks Management System was hosted somewhere outside of the legitimate university's web servers. The basic modifications through the conversation are shown by Figure 7. It is important to note that that many 'conversations' or iterations are needed with ChatGPT to generate a final desired product and ChatGPT is very capable of this.

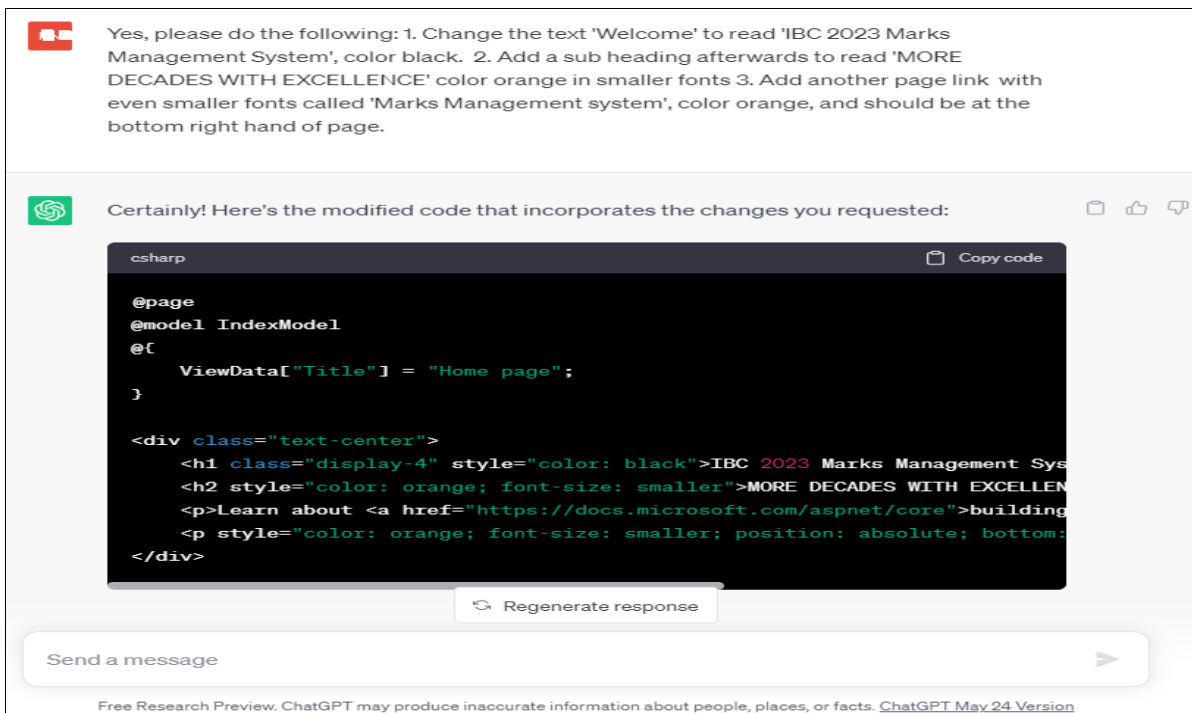


Figure 7: ChatGPT response
Source: ChatGPT May24 Version, generated on 19/06/2023

The code provided in Figure 7 was copied and pasted back into Visual Studio 2022 and the output is presented by Figure 8. For purposes of adhering to academic ethics, no further 'conversations' and iterations were carried out beyond Figure 7.

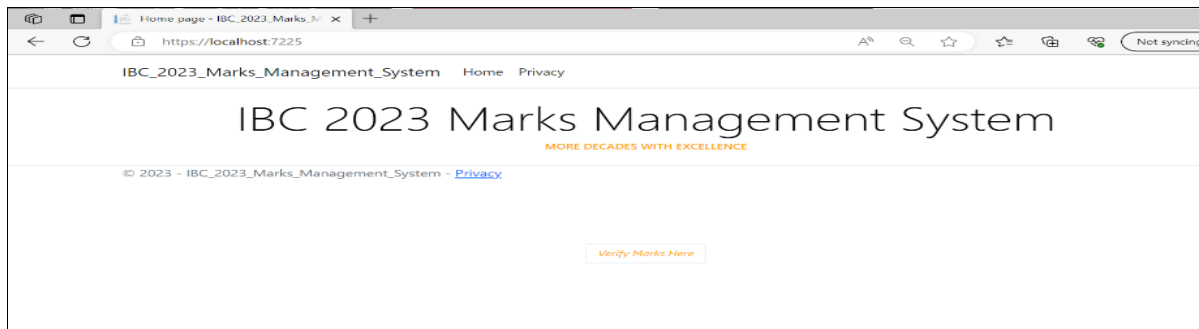


Figure 7: ChatGPT response

Source: Screenshot of website generated by ASP.NET Core Razor in C#

Based on the fictitious case outlined in the previous section, a conversation was initiated with ChatGPT instructing it to write an email purporting to come from the Head of Department of IBC 2023 University requesting a colleague to verify the marks on the Marks Management System. (*ChatGPT does not have any context to the prompts in any conversation*). The number of words which were specified and the context of the email that there is concern about low marks given, ChatGPT innovatively generated this request. ChatGPT's output is shown by Figure 8.

5. Investigate Artifact in Context

5.1 Bringing this all together: Schematic of the Phishing attack

The previous sections showed that it is possible to create a phishing campaign through having a conversation with ChatGPT directing it to create a fictitious phishing email and fictitious website. ChatGPT was able to also come up with what looked like a legitimate email heading. The phrase 'urgent' in the heading is interesting because this did not constitute part of the conversation request, but for some interesting reasons, ChatGPT 'decided' to fit in those words. It was also noted that ChatGPT provided guidelines on how the user should log-in to the website. The conversation to ChatGPT was that the 'colleague should log in', however, the fictitious email generated went further to outline the steps, thus further augmenting and collaborating innovatively with the request.



Write approximately 150 word email from the IBC 2023 university Head of Department to another colleague asking them to verify the marks on the Marks Management System (MAMS) There is concern that the marks for the class this colleague is teaching are too low. The colleague is to log onto this system called 'IBC 2023 Marks Management System'.



Subject: Urgent Request for Verification of Marks on IBC 2023 Marks Management System (MAMS)



Dear [Colleague's Name],

I hope this email finds you well. I am writing to address a matter of concern regarding the marks recorded in the IBC 2023 Marks Management System (MAMS) for the class you are currently teaching.

It has come to our attention that the marks appear to be significantly lower than expected, raising concerns about the fairness and accuracy of the grading process. In order to ensure the integrity of our assessment system, I kindly request your immediate assistance in logging onto the IBC 2023 Marks Management System (MAMS) and verifying the marks you have assigned to your students.

Please access the system by following these steps:

1. Visit the IBC 2023 University website at [insert website URL].
2. Log in using your faculty credentials.
3. Navigate to the "Marks Management System (MAMS)" section within your faculty portal.
4. Locate your class and click on the corresponding link to view the student marks.
5. Carefully review each student's marks, considering the grading criteria and rubrics provided for each assessment.

If you come across any discrepancies or anomalies, kindly make a note of them for further investigation. Your feedback and insights into the assessment process will be crucial in ensuring fairness and accuracy.

I would greatly appreciate it if you could complete the verification process as soon as possible and provide me with the results of your review. If you have any questions or need assistance, please do not hesitate to contact me.

Thank you for your prompt attention to this matter. Your cooperation in maintaining academic integrity is highly valued.

Best regards,

[Your Name]

Head of Department
IBC 2023 University

Figure 12: ChatGPT response
Source: ChatGPT May24 Version, generated on 19/06/2023

Through this conversation, it follows that ChatGPT was able to create unknowingly, a phishing campaign framework. This framework is shown in Figure 13.

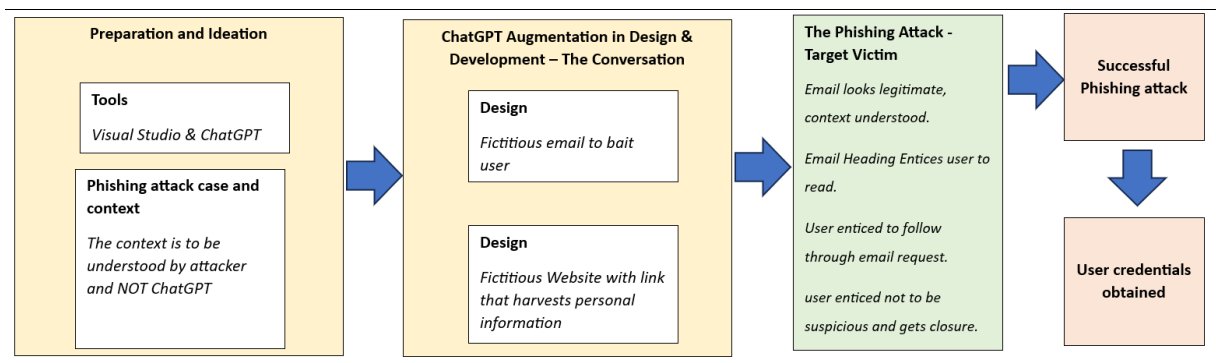


Figure 13: ChatGPT response
Source: ChatGPT May24 Version, generated on 19/06/2023

The framework developed and augmented by ChatGPT through conversation outline three important phases. In the first phase, namely the **preparation and ideation phase**, the attacker generates a fictitious context that a user would ‘buy-into’ and then scouts for the tools necessary to accomplish the campaign. In the second phase, namely the **augmentation in design and development**, the attacker, uses ChatGPT’s technical and intellectual prowess to design, build and execute what it is directed to. The third phase, namely the **phishing attack** phase, the campaign is executed. The success of the phishing attack is dependent on how well the three phases are conceptualised.

5.2 Ethics and the Framework

As a theoretical discourse, this study did not use any human data and present a framework which includes information that is readily and freely obtainable from the public domain. The discussions are open and derived from secondary sources that are appropriately cited.

6. Results and Findings

6.1 Literature on phishing frameworks and comparison with developed framework

It remains imperative to note that the framework developed may vary depending on unique contexts, however, the basic steps and approach can be compared with others who have done similar studies of phishing attacks. What is unique is that none of these studies augment ChatGPT, but essentially point to the process illustrated by Figure 14.

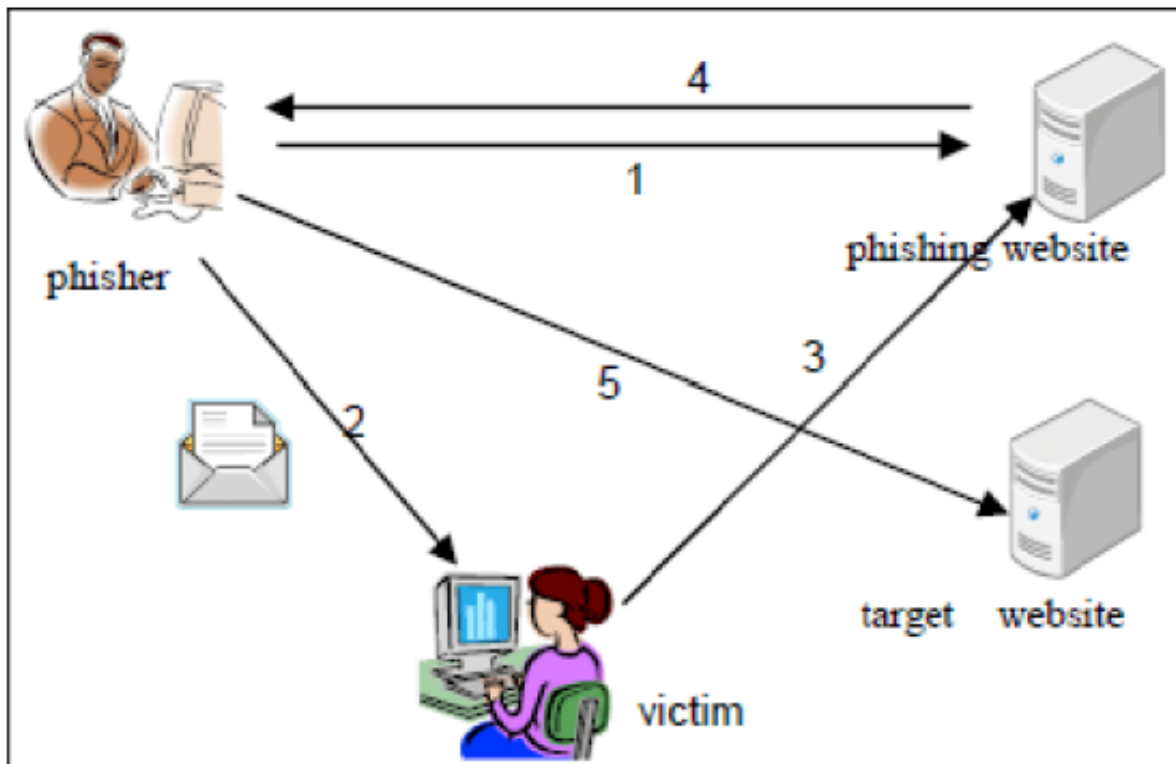


Figure 14: Phishing attack process
Source: Shankar, Shetty, and Nath (2019)

In the conversational phishing framework developed using the DSR approach, the process shown in Figure 14 and presented by Shankar et al. (2019), ChatGPT can be guided to make this process even more complex for the average user (victim) in a much faster way, reducing the cycle time between campaign generation and phishing success, therefore increasing the number of victims within a short period of time. The conversational phishing framework can be compared with previous studies on phishing campaigns as shown in Table 1.

Table 1: Comparing Conversational framework with phishing frameworks in literature

Framework	Steps in phishing campaign				
<i>This research work</i> <i>Conversational ChatGPT framework</i>	Preparation and Ideation	Augmentation in design and development	Phishing attack		
<i>Shankar et al. (2019)</i>	Plan	Compose email	Attack	Gather data	Fraud
<i>Chhikara, Dahiya, Garg, and Rani (2013)</i>	Attacker obtains E-mail addresses	The attacker generates an E-mail that appears legitimate	The attacker sends the E-mail to the intended victims in a way that appears legitimate	Depending on the content of the E-mail, the recipient opens a malicious attachment	The attacker harvests the victim's sensitive information and may exploit it in the future
<i>Alkhalil, Hewage, Nawaf, and Khan (2021)</i>	Planning phase: Target information gathering	Preparation phase: Exploit vulnerabilities	Attack phase: Attacker sends threats		

The conversational framework presented from this research has 3 steps namely: Step i) Preparation and Ideation; Step ii) Augmentation in design and development and Step iii); Phishing attack. In comparison to other frameworks for phishing identified through literature review, Shankar et al. (2019) outlines 5 steps, Chhikara, Dahiya, Garg, and Rani (2013) outlines 5 steps while Alkhalil, Hewage, Nawaf, and Khan (2021) outlines 3 steps.

Ideation which forms part of DSR is a crucial step in the preparational and ideation phase which seems not to have been considered by other phishing frameworks (Alkhalil et al., 2021; Chhikara et al., 2013; Shankar et al., 2019). This is perhaps because this phase is crucial in the advent of ChatGPT and other generative AI where it is necessary to brainstorm on its most effective use. All phases ultimately target the victim by identifying the best ways that the victim can be compromised. The next section details the implications of these attacks to management of organisations.

7. Managerial Implications

The work raises managerial implications which are necessary to be addressed and that are necessary to overcome the challenge of phishing. Firstly, managers should take note, that in the advent of ChatGPT and its future advanced versions (GPT4.0 and more), the problem of phishing attacks will only become much more pronounced and persistent. In order not to fall victim, management should apply the following countermeasures.

- A) Employ defensive strategies, such as raising awareness that phishing campaigns employ baits. Users should be trained to recognise these baits and how they are evolving.
- B) Employ technology to limit exposure to these campaigns by strengthening internal networks and email servers as well as encouraging reporting procedures in the event of these attacks happening.
- C) Remind users to always remain cautious and suspicious of emails particularly when the sources of these emails are uncertain. Users should always be encouraged not to blindly click or log-in to systems or links that are untrustworthy.

8. Conclusions, Limitations and Future Research

The increased rise of phishing threats poses risks to information systems security as shown by this work. The advent of ChatGPT that can now augment phishing campaigns has the potential to exacerbate phishing attack campaigns in unimaginable and malevolent ways, which further complicates the problems for managers and users of information systems. As highlighted, the victims are now faced with advanced generative AI capable of exploiting traditional phishing methods such as email and websites. This work shows through the DSR how these methods can now be augmented, designed, developed and executed by AI, in a conversational like manner.

The study limitations are that the framework presented, from the DSR approach should be tested and evaluated from a business lens to verify its utility. Often, the notion of researcher bias, is that the researcher perceives utility from own point of view rather than the business and it remains vital for business to gauge this utility. It follows then that DSR requires the creation of an innovative, purposeful artifact for a special problem domain that must therefore serve business needs and evaluated for such.

The implication to managers and users of information systems has been provided by this work, proposing that it remains paramount to the education and awareness at the forefront of effective countermeasures.

To this end, this work points to a continued focus for users to always remain cautious and suspicious. Further research may be necessary to understand how ChatGPT may be employed to assist users with this, beyond the ways that have been mentioned by this work. This work provides valuable information that raises awareness about phishing attacks by developing a conversational framework with ChatGPT that hopefully adds richness around the debate of the future of AI and security of systems.

Acknowledgments

I would like to acknowledge the efforts of the blind reviewers and editors whose valuable inputs have helped improve the quality of the paper.

References

- Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science*, 3, 563060. <https://doi.org/10.3389/fcomp.2021.563060>
- Biswas, S. S. (2023). Role of chat gpt in public health. *Annals of biomedical engineering*, 51(5), 868-869. <https://doi.org/10.1007/s10439-023-03172-7>
- Cambiaso, E., & Caviglione, L. (2023). Scamming the Scammers: Using ChatGPT to Reply Mails for Wasting Time and Resources. *arXiv preprint arXiv:2303.13521*.
- Chhikara, J., Dahiya, R., Garg, N., & Rani, M. (2013). Phishing & anti-phishing techniques: Case study. *International Journal of Advanced Research in computer science and software engineering*, 3(5).
- Collett, C. (2023). ChatGPT and copyright in South Africa: What you need to know. Retrieved from <https://www.bizcommunity.com/Article/196/542/236446.html>
- Desolda, G., Aneke, J., Ardito, C., Lanzilotti, R., & Costabile, M. F. (2023). Explanations in warning dialogs to help users defend against phishing attacks. *International Journal of Human-Computer Studies*, 176, 103056. <https://doi.org/10.1016/j.ijhcs.2023.103056>
- Dong, Y., Jiang, X., Jin, Z., & Li, G. (2023). Self-collaboration Code Generation via ChatGPT. *arXiv preprint arXiv:2304.07590*.
- Gabriela, T.-R., & Axinte, S.-D. (2023). *ChatGPT-Information Security Overview*. Paper presented at the International Conference on Cybersecurity and Cybercrime.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75-105. <https://doi:10.2307/25148625>

- Lambert, M. (2023). Entering the era of artificial intelligence (AI) in publishing. *South African Journal of Sports Medicine*, 35(1), 1-1.
- Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., & Eger, S. (2023). Chatgpt: A meta-analysis after 2.5 months. *arXiv preprint arXiv:2302.13795*.
- Liebreuz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *The Lancet Digital Health*, 5(3), e105-e106. [https://doi.org/10.1016/S2589-7500\(23\)00019-5](https://doi.org/10.1016/S2589-7500(23)00019-5)
- Mansfield-Devine, S. (2023). Weaponising ChatGPT. *Network Security*, 2023(4). [https://doi.org/10.12968/S1353-4858\(23\)70017-2](https://doi.org/10.12968/S1353-4858(23)70017-2)
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45-77.
- Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.
- Rivas, P., & Zhao, L. (2023). Marketing with chatgpt: Navigating the ethical terrain of gpt-based chatbot technology. *AI*, 4(2), 375-384.
- Savelka, J., Agarwal, A., Bogart, C., Song, Y., & Sakr, M. (2023). Can Generative Pre-trained Transformers (GPT) Pass Assessments in Higher Education Programming Courses? *arXiv preprint arXiv:2303.09325*. <https://doi.org/10.1145/3587102.3588792>
- Shankar, A., Shetty, R., & Nath, B. (2019). A review on phishing attacks. *International Journal of Applied Engineering Research*, 14(9), 2171-2175.
- Shimony, E., & Tsarfati, O. (2023). Chatting our way into creating a polymorphic malware. Retrieved from <https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware>
- Strowel, A. (2023). ChatGPT and Generative AI Tools: Theft of Intellectual Labor? *IIC-International Review of Intellectual Property and Competition Law*, 1-4. <https://doi.org/10.1007/s40319-023-01321-y>
- Taulli, T. (2023). Data: The Fuel for Generative AI *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business* (pp. 21-45): Springer. <https://doi.org/10.1007/978-1-4842-9367-6>
- Ventayen, R. J. M. (2023). ChatGPT by OpenAI: Students' Viewpoint on Cheating using Artificial Intelligence-Based Application. Available at SSRN 4361548.
- Vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to design science research. *Design science research. Cases*, 1-13.
- Wash, R. (2020). How experts detect phishing scam emails. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), 1-28.
- Werner, M. J., & Njenga, K. (2022). *Phishing Attack Victims and the Effect on Work Engagement*. Paper presented at the International Development Informatics Association Conference.
- Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*: Springer.
- Xu, M., Niyato, D., Chen, J., Zhang, H., Kang, J., Xiong, Z., . . . Han, Z. (2023). Generative AI-empowered simulation for autonomous driving in vehicular mixed reality metaverses. *arXiv preprint arXiv:2302.08418*.
- Zheng, S., & Becker, I. (2022). *Presenting Suspicious Details in {User-Facing} E-mail Headers Does Not Improve Phishing Detection*. Paper presented at the Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022).