

# Supporting the Design of Phishing Education, Training and Awareness interventions: an LLM-based approach

Francesco Greco<sup>1,\*</sup>, Giuseppe Desolda<sup>1</sup> and Luca Viganò<sup>2</sup>

<sup>1</sup> University of Bari "A. Moro", Bari, Italy

<sup>2</sup> King's College London, London, U.K.

## Abstract

Phishing remains one of the most effective cyber threats in our digital world, affecting millions of organizations. Phishing education, training, and awareness programs are used to address employees' lack of knowledge about phishing attacks. However, despite being very expensive, these interventions are not always effective, mainly due to the lack of customization of training materials based on the employees' needs and profiles. In fact, creating customized training content for each employee and each context would require a huge effort from security practitioners and educators thus increasing costs even more. The proposal we present in this paper is to use Large Language Models to automate some steps in the design process of training content, which is tailored to the specific user profile.

## Keywords

phishing education, large language models, warnings, training, simulated campaigns

## 1. Introduction

Phishing is currently one of the most significant cyber-threats, causing substantial losses for companies each year on a global scale [25]. Despite the technological solutions that exist to mitigate phishing attacks [2, 29], criminals are able to succeed due to the exploitation of vulnerabilities that originate from various human factors [16]. Among the primary human factors that can increase a user's susceptibility to phishing, *Lack of Knowledge* and *Lack of Resources* are of particular importance. The former refers to users missing specific knowledge and experience to correctly deal with phishing attacks [18], while the latter refers to the lack of educative resources that can effectively help users recognize phishing attacks [16]. Despite the users being often considered the "weakest link" in the cybersecurity of an organization [41], improving their awareness level can lead to making the employees one of the most valuable defensive assets of an organization [36].

Consequently, companies invest considerable resources [7] in increasing employee awareness and educating them with Phishing Education, Training, and Awareness (PETA)

---

2nd International Workshop on CyberSecurity Education for Industry and Academia (CSE4IA 2024)

\* Corresponding author.

✉ francesco.greco@uniba.it (F. Greco); giuseppe.desolda@uniba.it (G. Desolda); luca.vigano@kcl.ac.uk (L. Viganò)

ORCID 0000-0003-2730-7697 (F. Greco); 0000-0001-9894-2116 (G. Desolda); 0000-0001-9916-271X (L. Viganò)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

programs [40]. However, despite the general agreement among researchers on the usefulness of anti-phishing training [26], its effectiveness can vary considerably [36]. This may be attributed to human factors such as age, gender, technical expertise, and personal traits [26]. To address the ineffectiveness of phishing training and education due to the employees' individual differences, a viable solution would be to provide them with customized training material [17, 31, 50]. Furthermore, training material should be highly engaging and interesting for employees [17, 50], while also being easy and fast to consume. This is because users generally can dedicate little time to security aspects during their work hours [35]. Improving the relevance and quality of training material can indeed result in more effective phishing training programs [9]. This is likely to have the additional benefit of reducing the likelihood of users circumventing the educational process (e.g., by ignoring the training material altogether, or attempting to cheat in assessment quizzes).

However, the design and implementation of PETA material is not a trivial process and requires significant effort and human resources [3]. Typically, simulated phishing campaigns are employed to administer embedded training material [9, 33, 36, 51]. Although training campaigns are one of the most commonly used approaches, they are very expensive to conduct, and may even result in an increase in phishing susceptibility of some employees [36] or in a reduction in their click rate on legitimate links, potentially affecting their productivity [42]. In light of these problems, we deem it necessary to find a solution to address the lack of effective and affordable PETA resources.

This study presents ongoing research aimed at supporting the lightweight creation of effective PETA resources. The effectiveness of the resources will be achieved by tailoring each resource to the specific user, considering their profile created using ad hoc vulnerability assessment techniques (e.g. questionnaire). In this way, each user will be exposed to short, targeted, and relevant resources that they are more likely to accept than the traditional long and generic alternatives used today. The creation of these resources will also be facilitated by the use of LLMs, which, taking into account the user profile and the type of resource to which the user should be exposed (e.g., podcast, document, alert, etc.), will produce PETA resources tailored to the users, covering only their weakness, without exposing the user to aspects in which they are already confident.

This work establishes a first basis for a significant contribution to the broader Italian national project DAMOCLES (Detection And Mitigation Of Cyber attacks that exploit human vulnerabilities), which aims to develop a framework for the Italian Public Administration to assess human factors in cyber incidents and mitigate their impact through security awareness and customized user training. This latter point can indeed be addressed by using technologies like LLMs to support the creation of training material tailored to the individual weaknesses assessed.

## **2. Related work**

### **2.1. Phishing Education, Training, and Awareness (PETA)**

The term "PETA" (Phishing Education, Training, and Awareness) was recently brought up by Sarker et al. [40] to refer to Security Education, Training, and Awareness (SETA) interventions [23] in the domain of phishing. Katsikas et al. [28] define *awareness*, *training*,

and *education* as distinct concepts that together accomplish learning, starting with awareness and culminating in education. However, these terms are often used interchangeably in the literature [23]. Therefore, PETA is a broad concept that encompasses any type of intervention designed to enhance users' awareness and skills, including formal learning (courses, seminars, etc.), simulated phishing campaigns, quizzes, serious games, and anti-phishing warnings.

Simulated phishing campaigns aim to mimic real-world phishing attacks. They are typically part of a dedicated security awareness training program. These campaigns generate simulated phishing emails that closely resemble actual phishing attempts. Employees receive these emails to test their vigilance and response. Generally, *embedded training* is employed in conjunction with simulated phishing campaigns to present employees with landing pages that include educational material immediately after they click on a fake phishing link.

Anti-phishing warnings can constitute a valid intervention for increasing phishing awareness of users. These tools are employed to alert users about potential threats by blocking access to malicious websites [1] or emails [8] and are typically found in browsers (e.g., Google Chrome) or in email clients (e.g., Gmail).

## **2.2. Limitations of current PETA interventions**

A review of the literature conducted by Sarker et al. [40] reveals a number of issues with current PETA material. These include challenges in designing, implementing, and evaluating PETA interventions.

One prominent challenge is the lack of customization of training content [46]. This can lead to employees often being disengaged and uninterested in the training material [5, 50]. Additionally, poorly customized training content can result in being repetitive [9], culturally biased [17], or time-consuming [9, 34] for employees. A noteworthy example is *Anti-Phishing Phil* [43], a serious game for phishing education in which the user is asked to examine URLs to determine if they are associated with malicious or legitimate websites. One of the primary issues with the game was that a significant number of the presented websites were associated with American companies and thus unfamiliar to users outside of the United States. This resulted in some participants in the original study experiencing difficulty in determining if some URLs were legitimate or not. It is similarly important for training material to include different attack scenarios, in order to improve the users' ability to detect a wider spectrum of phishing attacks [32, 50]. Therefore, training material that addresses only, e.g., how to spot phishing URLs will not adequately teach users to defend against more advanced techniques such as spear phishing or persuasion cues [10].

Another critical issue is related to the warnings employed for the protection of users from phishing attacks. These warnings are generally ill-positioned [9, 34, 52], passive (i.e., do not block the user interaction flow) [19], and lead to users becoming habituated [1, 4, 30]. Moreover, the content of warnings usually lacks explanations [3, 6], which can result in users not trusting the system and being less motivated to adopt safe behaviors [6, 8, 48].

Finally, PETA programs tend to be highly expensive in terms of both economic and human resources [7, 46]. Furthermore, the deployment of embedded training requires a significant amount of manual human effort for the production of fake phishing emails, their

evaluation, the management of related tickets, the setting up of firewall rules, and so forth [3, 7]. This can result in the training material, such as anti-phishing recommendations, being outdated or incomplete [38]; it is instead crucial to include recent cyber-attacks and detailed information about how attackers operate and the types of tactics they use [45].

### **2.3. LLMs and education**

Large Language Models (LLM) are gaining traction in the field of education because of their ability to provide tailored feedback and suggestions, saving teachers time and effort in creating personalized materials and tailored feedback [27]. There are some attempts in the literature to use LLMs in fields as diverse as physics education [53] and medical education [39]. Common use cases involve support to educators in assessing and grading written tests, providing feedback to students, and generating educational content [54].

The use of LLMs in education is not without challenges [49]. There are many valid criticisms of these tools; one of the main problems is that they generate responses by predicting the most likely next word without any grasping of the semantical level; their stochastic nature makes the models possibly “hallucinate”, producing seemingly confident responses that are not factual, in part due to incomplete or biased training data.

Nonetheless, LLMs are undeniably performant also on human tasks [24]. For example, medical students use these tools to explain complicated medical concepts in simple terms, generate self-study questions, and create preliminary diagnoses and possible treatment plans [40].

Therefore, although the issue of hallucinations remains challenging to address, the proposal presented in this paper may still prove viable. Moreover, hallucinations can be limited by designing prompts that follow the established guidelines and best practices [14, 44]. For example, approaches like “chain-of-thought” have proved to help LLMs produce more grounded outputs [13].

The current limitations and future prospects of LLMs in education will constitute a valuable source of discussion during the workshop.

## **3. An LLM-based approach to mitigate challenges in PETA**

To address some of the key challenges in producing quality PETA material, i.e., *high costs*, *lack of customization*, and *ineffective warnings*, we propose an approach that leverages LLMs to help security practitioners and educators automate some tasks in the design process.

Automating the creation of training materials has already been indicated as being potentially beneficial to IT security teams across several dimensions, including reducing deployment and maintenance efforts, and the amount of human hours required, ultimately reducing costs for organizations [40]. Automation can also facilitate the delivery of tailored, recurrent, and relevant training interventions, reducing the costs associated with manually customizing training content [32].

### 3.1. Addressing the Lack of customization in PETA content

#### 3.1.1. Customized Simulated Phishing Campaigns

In order to create customized training content, it is first necessary to take into account the different psychological and demographic factors of the employees. Recently, in the context of the DAMOCLES Italian project, we proposed an approach to systematically conduct and assess the individual vulnerabilities of employees in an organization [20]. This approach has the goal of investigating, in the context of a simulated phishing campaign, the interaction between attack techniques (persuasion principles [10] and emotional triggers) and user personality traits, to determine which characteristics of phishing emails maximize the effectiveness of the attacks for specific employees. Therefore, once the employee's profile has been gathered in terms of the Big 5 model [37] (e.g., collected by administering the NEO Five-Factor-Inventory-3 [11]), tailored phishing emails can be crafted and delivered to test their susceptibility under challenging conditions. Obviously, the difficulty of the email is an important factor to consider in a phishing campaign: for example, the level of challenge could start low, by presenting users with phishing emails that are easy to detect and testing the user with very difficult emails towards the end of the campaign [12].

The role of the employee within the organization also plays a critical factor in the design of tailored phishing emails. For example, it would be an obvious red flag for a CEO to receive a phishing email sent by another "CEO" of their own company; therefore, such an attack would certainly be recognizable, even if the most effective social engineering techniques are used, based on his or her profile. Another factor that could be considered is the set of websites that employees visit most commonly to generate phishing URLs that resemble domains that are relevant to them. These could be collected either automatically by analyzing which websites the employee most frequently visits during their work hours, or by asking him or her to provide them spontaneously through a questionnaire. Moreover, the URLs on the organization's internal Domain Name System (DNS) server can be used to include domains that resemble the legitimate ones used by the company [26]. Finally, the employee's demographic information, such as name and gender, is a valuable source of information for creating spear-phishing emails that are more relevant, e.g., that do not contain generic greetings and address them by name.

LLMs can be used to automate the process of writing convincing emails which can include different topics and/or different persuasion principles (e.g., see [21, 22]). It is worth noting that commercially available LLMs like ChatGPT cannot be directly used to generate phishing emails, as this is not considered an ethical activity even for white-hat purposes. Therefore, less ethical tools such as Worm-GPT might be considered for generating phishing emails.

Once we have all the information about a specific employee, we can generate a phishing email by filling out the following prompt and feeding it to the LLM:

*"Pretend to be a security practitioner at [organization name] who is planning a simulated phishing campaign. You must create a fake phishing email in HTML format that is tailored to an employee of the organization. The email must be addressed to [employee name] (use [employee pronouns]), who is [employee role] in the organization. The email must be about [topic]; use Cialdini's persuasion principle of [persuasion principle] and include sentences that*

*leverage [emotional trigger]. You must use fake [organization nationality] real names for the sender. Create a phishing link URL to include in the email that mimics one of the following legitimate links: [URLs list]; for example, a fake URL for the website 'paypal.com' could be 'https://www.paypal-refund-claim.com'".*

The persuasion principle refers to Cialdini's theory of persuasion [10] and can be one of the following: authority, scarcity, reciprocation, social proof, liking, or consistency. The emotional trigger can refer to one among the most leveraged emotions used in phishing attacks: curiosity, fear, greed, anger, joy, confusion, or empathy. The choice of which persuasion principle and emotional trigger to use is dictated by the employee's personality traits. The topic can vary to generate different emails covering different plausible scenarios such as "request of password reset", "account blocked", "free giveaway", "payment request", etc. It is worth noting that in this example the prompt is considering exactly one persuasion principle and one emotional trigger at once, for simplicity; however, more complex attacks may also include a combination of two or more persuasion principles (e.g., authority and scarcity) and/or emotional triggers to create more effective phishing emails. An example of the usage of this prompt is presented in Appendix A.

To investigate the output of the model, it is possible to ask the LLM to explain how the persuasion principles and emotional triggers are addressed by individual sentences or words in the generated text; we can also ask the LLM to report the legitimate URL that was mimicked with the phishing URL. To do this, we can extend the previous prompt with the following text:

*"For each persuasion principle and emotional trigger used in the email, produce an explanation that points out the pieces of text used to employ them. Finally, report the legitimate URL that is being mimicked in the email.*

*Output format:*

*[EMAIL]*

*-----*

*[EXPLANATION]"*.

### **3.1.2. Customized Embedded Training**

In addition to customizing the emails in the simulated phishing campaign, the educational content for conducting the embedding training must be generated and personalized taking into account both the type of techniques used in the phishing attacks and the information about the employee [26]. For example, the employee who falls victim to a simulated phishing email must be presented with a customized landing page that:

- Debriefs them about the simulated attack, addressing them by their name.
- Explains the social engineering techniques used in the email (e.g., authority principle and urgency) and some tips on how to avoid them.
- Reports the fake phishing URL they clicked on and presents it next to the legitimate one, highlighting the phishing cues that the victim should have noticed, such as the top-level domain being misplaced, or URL spoofing.

The educational level for generating tailored training material must also be considered in the generation of the educational content: for example, users who are more familiar with technical aspects of IT security might benefit from explanations that include technical jargon and/or more details; therefore, the explanation could, e.g., include terms such as “domain”, “homograph attack”, etc., which might be more relevant to them. This allows us to generate explanations that are more appropriate to who receives the explanation [47].

A possible prompt to generate such embedded training content would be:

*“Pretend to be a cybersecurity educator who teaches employees how to recognize an email of phishing. [organization] organized a simulated phishing campaign to test its employees' susceptibility to phishing. Specifically, [employee name] clicked on a phishing link in one of the fake emails and was redirected to a landing page containing training information. The email used [social engineering principles] principles; moreover, the phishing link was [phishing URL], which mimicked the legit link [mimicked URL].*

*Create a short explanation webpage that:*

*1) debriefs them about the simulated attack*

*2) explains the techniques that were used and some tips on how to avoid them*

*3) reports both URLs (phishing and mimicked), highlighting the URL spoofing techniques that were used.*

*Consider that the employee is a [employee role] and is [expertise level] of cybersecurity, and tailor the explanation accordingly.”*

### **3.2. LLMs for improving anti-phishing warnings**

In order to constitute a valuable phishing awareness resource for users, warning dialogs must include content that is relevant and educational to the user. Determining a priori whether the warning content is helpful to the user is not an easy task, as it is heavily affected by the user’s knowledge of phishing, security, and IT. Moreover, employees usually have limited time to dedicate to reading warnings, as cybersecurity is usually a secondary task for them [35].

Therefore, if we want explanations in warning messages to be considered, they must be designed to be readable, understandable, and alerting [15]. In addition, warning messages should vary so that users do not easily become habituated to seeing the same warning under different risk circumstances [4]. Since generating high-quality warning messages is an onerous task, it is simply not feasible to manually create diverse content that also adapts to each user’s knowledge level.

A possible solution to this problem is to use LLMs to automatically generate warning dialogs that include explanations that (i) are dynamically generated (thus helping to avoid habituation), (ii) address the specific phishing threat (thus potentially improving decision-making), and (iii) adapt to the user’s knowledge (thus being relevant and understandable).

The explanation in a warning dialog should follow the best practices established in the literature of warning design and have a standard structure such as those proposed in [15]. Specifically, it should be formed by three parts: 1) a description of the phishing feature that is being explained, 2) an explanation of the hazard of phishing, and 3) potential consequences of a successful attack. Hereafter, we propose a possible prompt to generate such comprehensive warning dialogs:

*“Construct a brief explanation message (max 50 words) directed to [employee expertise level] that will follow this structure:*

- 1. description of the most relevant phishing feature*
- 2. explanation of the hazard*
- 3. consequences of a successful phishing attack*

*For example, a message that explains that a URL in the email (PHISHING\_URL) is imitating another legitimate one (SAFE\_URL), would be:*

*“The target URL (PHISHING\_URL) is an imitation of the original one, (SAFE\_URL). This site might be intended to take you to a different place. You might be disclosing private information.”.*

#### **4. Conclusions and future work**

While phishing remains a critical problem, user education, training, and awareness can mitigate the success of these attacks and improve an organization’s susceptibility. In fact, addressing human vulnerabilities can make employees an essential line of defense against phishing attacks [41]. Since PETA interventions are often very expensive for organizations, automation can help reduce the burden and produce high-quality training materials more easily. In this paper, we have proposed LLMs as a tool to reduce manual effort and produce highly customizable training material that can be tailored to user needs.

LLMs are a relatively new technology that, despite their impressive performance on various human tasks [24], still have clear limitations, such as suffering from *hallucinations*, and thus need to be carefully supervised by human experts. However, we envision a future in which human-AI collaboration is fundamental to support complex tasks that traditionally belong to humans. Therefore, this approach may still be valuable for security practitioners and educators to produce effective PETA interventions much more efficiently.

In future work, we plan to implement the approach presented in this work and to include it as a possible mitigation strategy against phishing attacks for public administration within the DAMOCLES Italian project. The effectiveness of the proposed approach will be evaluated by iteratively assessing an organization’s susceptibility to phishing over time. Specifically, the effectiveness of an LLM-powered simulated phishing campaign will be evaluated by measuring the employees’ click rate at time zero (e.g., if it is too low, the phishing emails are probably too easy to detect); on the other hand, the effectiveness of the educational content will be measured by monitoring the click-rate over time for the exposed users. Finally, it will be of paramount importance to collect the feedback of employees who are exposed to LLM-generated training content both during the design phase and once the system is deployed.

#### **Acknowledgements**

This work has been supported by the Italian Ministry of University and Research (MUR) and by the European Union - NextGenerationEU, under grant PRIN 2022 PNRR "DAMOCLES: Detection And Mitigation Of Cyber attacks that exploit human vulnerabilities" (Grant P2022FXP5B).



This work is partially supported by the co-funding of the European Union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 - Partnerships extended to universities, research centres, companies and research D.D. MUR n. 341 del 5.03.2022 - Next Generation EU (PE0000014 - "Security and Rights In the CyberSpace - SERICS" - CUP: H93C22000620001).

The research of Francesco Greco is funded by a PhD fellowship within the framework of the Italian "D.M. n. 352, April 9, 2022"- under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 3.3 - PhD Project "Investigating XAI techniques to help user defend from phishing attacks", co-supported by "Auriga S.p.A." (CUP H91122000410007).

## References

- [1] Akhawe, D. and Felt, A.P., Alice in warningland: a large-scale field study of browser security warning effectiveness. *USENIX conference on Security*. Washington, D.C., 2013, pp. 257-272.
- [2] Almomani, A., Gupta, B.B., Atawneh, S., Meulenberg, A. and Almomani, E., A Survey of Phishing Email Filtering Techniques, *IEEE Commun. Surv. Tutor.* 15 (2013) pp. 2070-2090. doi:10.1109/SURV.2013.030713.00020.
- [3] Althobaiti, K., Meng, N. and Vaniea, K., I Don't Need an Expert! Making URL Phishing Features Human Comprehensible. *Conference on Human Factors in Computing Systems*. Yokohama, Japan, 2021, pp. 1-17. 10.1145/3411764.3445574
- [4] Anderson, B.B., Kirwan, C.B., Jenkins, J.L., Eargle, D., Howard, S. and Vance, A., How Polymorphic Warnings Reduce Habituation in the Brain: Insights from an fMRI Study. *ACM Conference on Human Factors in Computing Systems*. Seoul, Republic of Korea, 2015, pp. 2883-2892. 10.1145/2702123.2702322
- [5] Arachchilage, N.A.G., Love, S. and Beznosov, K., Phishing threat avoidance behaviour: An empirical investigation, *Comput. Hum. Behav.* 60 (2016) pp. 185-197. doi:10.1016/j.chb.2016.02.065.
- [6] Bravo-Lillo, C., Cranor, L.F., Downs, J., Komanduri, S. and Sleeper, M., Improving Computer Security Dialogs. *International Conference on Human-Computer Interaction*. Berlin, Heidelberg, 2011, pp. 18-35.
- [7] Brunken, L., Buckmann, A., Hielscher, J. and Sasse, M.A., "To Do This Properly, You Need More Resources": The Hidden Costs of Introducing Simulated Phishing Campaigns. *32nd USENIX Security Symposium 2023*, pp. 4105-4122.
- [8] Buono, P., Desolda, G., Greco, F. and Piccinno, A., Let warnings interrupt the interaction and explain: designing and evaluating phishing email warnings. *CHI Conference on Human Factors in Computing Systems*. Hamburg Germany, 2023, pp. 1-6. 10.1145/3469886
- [9] Caputo, D.D., Pfleeger, S.L., Freeman, D.J. and Johnson, M.E., Going Spear Phishing: Exploring Embedded Training and Awareness, *S&P* 12 (2014) pp. 28-38. doi:10.1109/MSP.2013.106.
- [10] Cialdini, R.B., *Influence: The Psychology of Persuasion*. revised ed, Harper Collins.2009.
- [11] Costa, P.T. and McCrae, R.R., Four ways five factors are basic, *Pers. Individ. Dif.* 13 (1992) pp. 653-665. doi:10.1016/0191-8869(92)90236-I.
- [12] CybSafe. *The ultimate people-centric guide to simulated phishing*, 2023. Url: <https://www.cybsafe.com/value/simulated-phishing/> Last Access 21 Apr 2024.

- [13] DAIR.AI. *Chain-of-Thought Prompting - Prompt Engineering Guide*, Url: <https://www.promptingguide.ai/techniques/cot> Last Access 7 May 2024.
- [14] DAIR.AI. *Prompt Engineering Guide*, Url: <https://www.promptingguide.ai/> Last Access 24 Jan. 2024.
- [15] Desolda, G., Aneke, J., Ardito, C., Lanzilotti, R. and Costabile, M.F. 2023. Explanations in warning dialogs to help users defend against phishing attacks. In *International Journal of Human-Computer Studies*, 20.
- [16] Desolda, G., Ferro, L.S., Marrella, A., Catarci, T. and Costabile, M.F. 2021. Human Factors in Phishing Attacks: A Systematic Literature Review. In *ACM Computing Survey ACM*, 35.
- [17] Dixon, M., Arachchilage, N.A.G. and Nicholson, J., Engaging Users with Educational Games: The Case of Phishing. *Conference on Human Factors in Computing Systems*. Glasgow, Scotland Uk, 2019, pp. 1-6. 10.1145/3290607.3313026
- [18] Dupont, G., *The Dirty Dozen Errors in Maintenance*. Proceedings of the 11th Meeting on Human Factors In Aviation Maintenance and Inspection.1997.
- [19] Egelman, S., Cranor, L.F. and Hong, J., You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. *SIGCHI Conference on Human Factors in Computing Systems*. 2008, pp. 1065–1074 10.1145/1357054.1357219
- [20] Greco, F., Buono, P., Desiato, D., Desolda, G., Lanzilotti, R. and Ragone, G., Unlocking the Potential of Simulated Phishing Campaigns: Measuring the Impact of Interaction among Different Human Factors. *DAMOCLES: Detection And Mitigation of Cyber attacks that exploit human vulnerabilities workshop, co-located with AVI '24*. Arenzano (Genoa), Italy, 2024, pp. 10.
- [21] Greco, F., Desolda, G., Esposito, A. and Carelli, A., David versus Goliath: Can Machine Learning Detect LLM-Generated Text? A Case Study in the Detection of Phishing Emails. *The Italian Conference on CyberSecurity*. Salerno, Italy, 2024, pp.
- [22] Hazell, J. *Spear Phishing With Large Language Models*, 2023. url: <https://arxiv.org/abs/2305.06972> Last Access.
- [23] Hu, S., Hsu, C. and Zhou, Z., Security Education, Training, and Awareness Programs: Literature Review, *JCIS* 62 (2022) pp. 752-764. doi:10.1080/08874417.2021.1913671.
- [24] HuggingFace. *Open LLM Leaderboard*, 2024. Url: [https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard) Last Access 13 Feb. 2024.
- [25] IBM 2023. Security X-Force Threat Intelligence Index.
- [26] Jampen, D., Gür, G., Sutter, T. and Tellenbach, B., Don't click: towards an effective anti-phishing training. A comparative literature review, *Hum. Cent. Comput. Inf. Sci.* 10 (2020) pp. 33. doi:10.1186/s13673-020-00237-7.
- [27] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J. and Kasneci, G., ChatGPT for good? On opportunities and challenges of large language models for education, *ChatGPT for good? On opportunities and challenges of large language models for education* 103 (2023) pp. 102274. doi:10.1016/j.lindif.2023.102274.
- [28] Katsikas, S.K., Health care management and information systems security: awareness, training or education?, *Int. J. Med. Inform.* 60 (2000) pp. 129-135. doi:10.1016/S1386-5056(00)00112-X.

- [29] Khonji, M., Iraqi, Y. and Jones, A., Phishing Detection: A Literature Survey, *IEEE Commun. Surv. Tutor.* 15 (2013) pp. 2091-2121. doi:10.1109/SURV.2013.032213.00009.
- [30] Kim, S. and Wogalter, M.S., Habituation, Dishabituation, and Recovery Effects in Visual Warnings, *Habituation, Dishabituation, and Recovery Effects in Visual Warnings* 53 (2009) pp. 1612-1616. doi:10.1177/154193120905302015.
- [31] Kirlappos, I. and Sasse, M.A., Security Education against Phishing: A Modest Proposal for a Major Rethink, *S&P 10* (2012) pp. 24-32. doi:10.1109/MSP.2011.179.
- [32] KnowB4. *Whitepaper: Building an Effective and Comprehensive Security Awareness Program* Url: <https://info.knowbe4.com/wp-building-effective-comprehensive-sat> Last Access 21 Apr 2024.
- [33] Kumaraguru, P., Cranshaw, J., Acquisti, A., Cranor, L., Hong, J., Blair, M.A. and Pham, T., School of phish: a real-world evaluation of anti-phishing training. *Symposium on Usable Privacy and Security*. Mountain View, California, USA, 2009, pp. 1-12. 10.1145/1572532.1572536
- [34] Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E., Protecting people from phishing: the design and evaluation of an embedded training email system. *SIGCHI Conference on Human Factors in Computing Systems*. San Jose, California, USA, 2007, pp. 905-914. 10.1145/1240624.1240760
- [35] Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F. and Hong, J., Teaching Johnny not to fall for phish, *Trans. Internet Technol.* 10 (2010) pp. 1-31. doi:10.1145/1754393.1754396.
- [36] Lain, D., Kostianen, K. and Čapkun, S., Phishing in Organizations: Findings from a Large-Scale and Long-Term Study, in: *Proceedings of the 2022 IEEE Symposium on Security and Privacy, SP*, Year, pp. 842-859. doi:10.1109/SP46214.2022.9833766.
- [37] McCrae, R.R. and Costa, P.T.J., *The five-factor theory of personality*. Handbook of personality: Theory and research, The Guildford Press, New York, NY, US.2008.
- [38] Mossano, M., Vaniea, K., Aldag, L., Düzgün, R., Mayer, P. and Volkamer, M., Analysis of publicly available anti-phishing webpages: contradicting information, lack of concrete advice and very narrow attack vector, in: *Proceedings of the IEEE European Symposium on Security and Privacy Workshops, EuroS&PW '20*, IEEE, Year, pp. 130-139. doi:10.1109/EuroSPW51379.2020.00026.
- [39] Pandey, M. and Mishra, V., Large Language Models in Medical Education and Quality Concerns, *JQHE* 6 (2023) pp. 3. doi:10.23880/jqhe-16000319.
- [40] Sarker, O., Jayatilaka, A., Haggag, S., Liu, C. and Babar, M.A., A Multi-vocal Literature Review on challenges and critical success factors of phishing education, training and awareness, *JSS* 208 (2024) pp. 111899. doi:<https://doi.org/10.1016/j.jss.2023.111899>.
- [41] Sasse, M.A., Brostoff, S. and Weirich, D., Transforming the 'Weakest Link' - a Human/Computer Interaction Approach to Usable and Effective Security, *Transforming the 'Weakest Link' - a Human/Computer Interaction Approach to Usable and Effective Security* 19 (2001) pp. 122-131. doi:10.1023/A:1011902718709.
- [42] Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L.F. and Downs, J., Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. *SIGCHI Conference on Human Factors in Computing Systems*. Atlanta, Georgia, USA, 2010, pp. 373-382. 10.1145/1753326.1753383

- [43] Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E., Anti-Phishing Phil: the design and evaluation of a game that teaches people not to fall for phish. *Symposium on Usable privacy and security*. Pittsburgh, Pennsylvania, USA, 2007, pp. 88–99. 10.1145/1280680.1280692
- [44] Shieh, J. Best practices for prompt engineering with OpenAI API OpenAI.
- [45] TerranovaSecurity. *Phishing Benchmark Global Report*, 2021. Url: <https://www.terrnovasecurity.com/resources/guides/gone-phishing-report-2021> Last Access 21 Apr 2024.
- [46] Tessian 2022. Phishing Awareness Training: How Effective is Security Training? In *Advanced Email Threats*.
- [47] Viganò, L. and Magazzeni, D., Explainable Security. *IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence*. 2018, pp. 7. arXiv:1807.04178
- [48] Vilone, G. and Longo, L., Notions of explainability and evaluation approaches for explainable artificial intelligence, *Inf Fusion* 76 (2021) pp. 89-106. doi:10.1016/j.inffus.2021.05.009.
- [49] Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P.S. and Wen, Q. *Large Language Models for Education: A Survey and Outlook*, 2024. url: <https://arxiv.org/abs/2403.18105> Last Access.
- [50] Wen, Z.A., Lin, Z., Chen, R. and Andersen, E., What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. *Conference on Human Factors in Computing Systems*. Glasgow, Scotland Uk, 2019, pp. 1-12. 10.1145/3290605.3300338
- [51] Wright, R.T., Jensen, M.L., Thatcher, J.B., Dinger, M. and Marett, K., Research Note: Influence Techniques in Phishing Attacks: An Examination of Vulnerability and Resistance, *Inf. Syst.* 25 (2014) pp. 385-400.
- [52] Wu, M., Miller, R.C. and Garfinkel, S.L., Do Security Toolbars Actually Prevent Phishing Attacks? *SIGCHI Conference on Human Factors in Computing Systems*. 2006, pp. 601–610. 10.1145/1124772.1124863
- [53] Wulff, P., Physics language and language use in physics - What do we know and how AI might enhance language-related research and instruction, *Eur. J. Phys.* 45 (2024) pp. 023001. doi:10.1088/1361-6404/ad0f9c.
- [54] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y. and Gašević, D., Practical and ethical challenges of large language models in education: A systematic scoping review, *BJET* 55 (2024) pp. 90-112. doi:10.1111/bjet.13370.

## A. Appendix: An example of an email created using an LLM

**F You**

Pretend to be a security practitioner at University of Bari who is planning a simulated phishing campaign. You must create a fake phishing email in HTML format that is tailored to an employee of the organization. The email must be addressed to Dr. Francesco Greco (use he/him pronouns), who is a researcher in the organization. The email must be about account password reset; use Cialdini's persuasion principle of Scarcity and include sentences that leverage Urgency. You must use fake Italian real names for the sender. Create a phishing link URL to include in the email that mimics one of the following legitimate links: [www.uniba.idp.cineca.it, www.uniba.it, www.u-gov.uniba.it]; for example, a fake URL for the website 'paypal.com' could be 'https://www.paypal-refund-claim.com'

**ChatGPT**

Subject: Urgente: Reset della Password del Tuo Account

Caro Dr. Francesco Greco,

Spero che questa email ti trovi bene. Sono Martina Rossi dal dipartimento di tecnologia informatica presso l'Università di Bari.

Stiamo effettuando un'importante operazione di manutenzione dei sistemi informatici dell'università, e abbiamo bisogno del tuo tempestivo aiuto. Come parte di questo processo, abbiamo individuato un potenziale rischio di sicurezza relativo ai tuoi dati di accesso.

Per garantire la massima sicurezza dei tuoi dati e dei sistemi dell'università, ti chiediamo gentilmente di procedere con il reset della tua password dell'account.

Per favore, clicca sul seguente link per completare il processo di reset della password:  
[<https://www.uniba-idp-cineca-passwordreset.com>]

Ricorda che questa operazione è fondamentale per garantire la sicurezza dei tuoi dati personali e dei dati dell'università. Ti preghiamo di agire tempestivamente.

Ti ringraziamo per la tua collaborazione e comprensione in questa materia.

Cordiali saluti,

Martina Rossi  
Dipartimento di Tecnologia Informatica  
Università di Bari

Nota: Questa email è stata inviata automaticamente. Si prega di non rispondere direttamente.

**Figure 1.** An example of usage of the prompt presented in Section 3.1.1 to produce an email for a simulated phishing campaign, using ChatGPT-3.5. We filled in the prompt with mock parameters, indicating the University of Bari as the organization and one of the authors as the victim of the attack. In the prompt, we indicated three different legitimate domains commonly used in the University of Bari for web activities.

The email was generated in Italian since ChatGPT recognized that the University of Bari is an Italian institution. We asked subsequently to translate it to English to present it to our English-speaking readers. In the following there is the translated email:

*Subject: Urgent: Reset Your Account Password*

*Dear Dr. Francesco Greco,*

*I hope this email finds you well. I am Martina Rossi from the Department of Computer Technology at the University of Bari.*

*We are conducting an important maintenance operation on the university's computer systems, and we need your prompt assistance. As part of this process, we have identified a potential security risk related to your login credentials.*

*To ensure the utmost security of your data and the university's systems, we kindly ask you to proceed with resetting your account password.*

*Please click on the following link to complete the password reset process:*

*[<https://www.uniba-idp-cineca-passwordreset.com>]*

*Please note that this operation is essential to ensure the security of your personal data and the university's data. We kindly request you to act promptly.*

*Thank you for your cooperation and understanding in this matter.*

*Best regards,*

*Martina Rossi  
Department of Computer Technology  
University of Bari*

*Note: This email has been sent automatically. Please do not reply directly.*