An Adaptive Phishing Awareness Training Environment based on LLMs and Interactive Learning

Kapischan Sriganthan and Felix Härer

University of Applied Sciences and Arts Northwestern Switzerland
Peter Merian-Strasse 86, 4002 Basel, Switzerland
felix.haerer@fhnw.ch

Abstract. Phishing remains a major cybersecurity threat, exploiting human vulnerabilities through deceptive tactics. Traditional awareness training often involves simulated phishing campaigns via e-mail and similar channels, in addition to training materials for employees to complete. Prior work indicates limitations in effectiveness, highlighting the lack of interactivity and contextual factors. This paper explores an interactive, user-adaptive, LLM-based approach to phishing awareness training, combining simulated phishing campaigns with adaptive feedback mechanisms from interactive learning. A training environment was developed as a research prototype, supporting experiments on adaptive training methods, including modes of interaction, target customization, feedback, and adaptivity to user environments. The prototype demonstrates the potential of realizing these aspects with LLMs according to interactive learning principles, such as targeted user engagement and instantaneous feedback. It is intended to serve as a basis for studies on effectiveness, efficiency, and adaptive methods in corporate settings, education, or phishing prevention initiatives.

Keywords: Phishing · Awareness Training · LLM · Cybersecurity Simulation · Interactive Learning

1 Introduction

Although advancements in encryption, firewalls, and multi-factor authentication have strengthened technical defenses, human error steadily remains the weakest link in the security chain, causing high-severity breaches through an ever-increasing number of phishing attack techniques [8]. Reports from the Anti-Phishing Working Group (APWG), a global initiative against cybercrime, indicate that nearly 900,000 phishing websites were detected in the second quarter of 2024 and highlight the recent severity of "business e-mail compromise" (BEC) attacks, causing substantial financial and reputational damages [1]. These attacks not only compromise the security of users and enterprises, but increasingly

involve identity theft and impersonation that convinces employees to reveal privileged information or to transfer assets, reportedly causing damage amounting to 2.9 billion USD in 2023 within the US [1].

Addressing these attacks requires effective phishing awareness training that engages users beyond providing information or course material. Training methods often involve simulated phishing campaigns where users are subjected to phishing e-mails or websites [4]. Those users who fall victim to these simulated attacks are often asked to complete training materials or receive information personally or in workshops. In particular, prior research has shown that established training methods mostly fail to engage users effectively, lacking interactivity and real-world applicability [2]. Furthermore, simulation and training methods are limited in terms of adaptivity [2,4], especially for training against targeted high-severity attacks such as BEC or targeted "whale phishing" attacks, aimed at high-level employees. Adaptivity allows simulations and training to target individual users and enterprise environments as opposed to simulating generic campaigns and providing training not specific to the campaign, the user, and the business setting.

Based on the premise that interactivity and adaptivity need to be addressed in future phishing awareness training, this paper applies principles of interactive learning [7,3] and introduces adaptivity in simulations and training by LLMs. An adaptive Phishing Awareness Training Environment has been developed as a research prototype. Towards strengthening phishing awareness training efficiency and effectiveness against modern, targeted attacks in different enterprise settings, the aim of the prototype is to show the feasibility of adaptivity and interactive learning concepts and to serve as a basis for further experimentation and studies.

This paper follows this structure: Section 2 outlines background and the core concept of the adaptive approach. The research prototype with architecture and implementation is discussed in Section 3. Section 4 concludes the paper by summarizing findings and outlining directions for future research.

2 Background and Concept

This section introduces the background and the approach to phishing awareness training, centered around interactive learning and adaptivity by utilizing Large Language Models (LLMs) in an interactive environment.

2.1 Background

Interactive learning calls for learning by interacting with the subject matter, allowing learners to engage or experiment in direct contact with a subject, receive instant feedback on their actions or questions, and explore the subject based on their own experiences [3]. In terms of phishing, interactive learning can support the ability of users to recognize phishing tactics over time [8].

The interactive learning principles are combined with and supported by LLMs that have shown potential in their ability to generate text according to instructions and prompting, interactions for assistance, or question answering [6].

Their ability to analyze textual cues and infer intent makes them particularly suited for phishing awareness training, where identifying subtle manipulative tactics is crucial. Their application allows for training according to interactive learning, adapted to phishing messages, users, and business settings. On this basis, training can involve teaching phishing indicators, help users recognize deceptive messages or suspicious links, and identify social engineering techniques.

2.2 Concept for Interaction Design

This approach builds upon interactive phishing awareness methods and draws inspiration from the serious games and chatbot training systems [10,9]. The concept of an interactive phishing awareness training system is applied, where simulated phishing scenarios are followed by coaching, here in the form of an LLM-based coach, acting as both a trainer and a guide. Users are exposed to the scenarios within the platform while the coach is present to evaluate user reactions and provide guidance based on their behavior [10]. Integrated into the environment, the coach engages with users when identifying phishing attempts, using interaction design principles towards improving their detection skills. Language models are utilized in terms of their increasingly capable language understanding in the interaction design that consists of three distinct phases:

Message generation For campaigns simulating both benign and phishing messages, content is generated with contextualization. User and environment data form the basis for creating targeted messages. Few-shot prompting and constraint techniques [5] are employed to augment LLM prompts with environment descriptions. For example, the business areas of a company, division, or team; the organizational structure, such as names and roles of co-workers and superiors; and user-specific details, such as the employee's role, tasks, and profile information.

Engaging with users interactively To evaluate user responses to phishing content, users are exposed to generated messages and asked how they perceive them. The ensuing dialogue incorporates contextual information about the environment and the user to personalize and target the interaction [6]. The system replicates message-based interactions such as e-mail exchanges and prompts users to assess individual messages, e.g., by reviewing links. It then asks for a decision on whether to trust a message or report it as phishing. The scenario example includes security alert messages prompting immediate action, phishing surveys attempting to collect sensitive information, and spoofed messages impersonating trusted services [2].

Coaching based on response behavior Users are coached based on their response behavior through explanations and discussions. The moment when a user reacts needs to be immediately utilized for feedback [3], in order to create

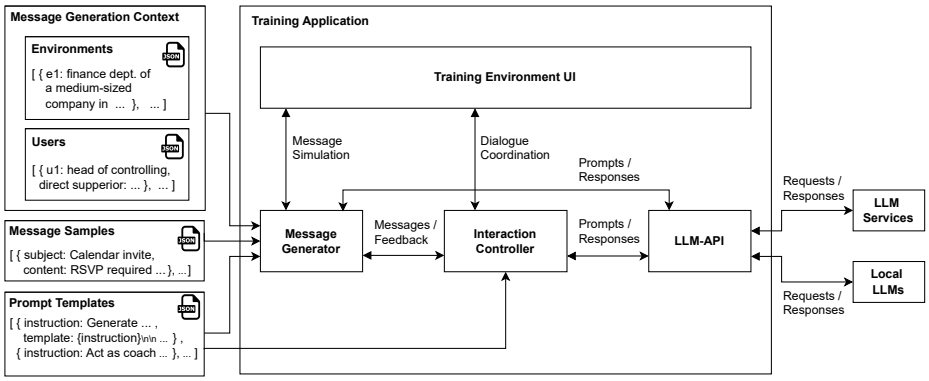

Fig. 1. Architecture of the adaptive phishing awareness training system, illustrating the main components and data flows for message generation, user interaction, and LLM integration.

a "teachable moment" to enhance learning outcomes. Here, the user's understanding of the situation is crucial. Instantaneous feedback reflects their decision within the context of guidance, whether phishing occurred and how the consequences could affect, e.g., the company finances or personal data. Users can actively discuss and ask questions in this situation. The coach will explain why certain elements may indicate a phishing attempt and how to verify authenticity. The coach highlights missed warning signs, such as manipulative urgency cues, deceptive URLs, or inconsistencies in sender details. Positive reinforcement is used in case of correct identifications.

Overall, this iterative process is aimed at strengthening users' abilities to recognize phishing tactics over time [8].

3 Research Prototype

For evaluating the adaptive approach by further experiments and future studies, a research prototype architecture and implementation have been developed. This section introduces the architecture and a concrete implementation in Python.

3.1 Architecture

Figure 1 presents the architecture of the LLM-based adaptive awareness training system. The main application encompasses four components:

The `Message Generator`, handling the processing and generation of phishing messages from three inputs. First, Message Generation Context consisting of environment and user information is loaded for providing context when constructing a prompt. Secondly, from a collection of Message Samples, a message is loaded to adapt it for the given context. Thirdly, the prompt is constructed by loading a template from Prompt Templates and instantiating its variables,

in particular instruction, context, and message. For example, a prompt will be generated from a sample message on a calendar invite as a custom message with details of the environment and user such as the department, position, and superiors or co-workers.

The `Training Environment UI` hosts both the simulated phishing messages and the dialogue with the LLM-based coach in an embedded chat. Here, a generated message is displayed together with real, unchanged messages. Following message generation, the system displays messages and engages with the user according to the interaction design. When selecting and viewing a message, the user can determine in discussion with the coach whether or not it is a phishing message and label the message.

The dialogue is coordinated by the `Interaction Controller`. The coach engages in context-dependent discussions after the initial context is set using an initiation prompt template. Per instruction to the LLM, it acts as a coach, providing contextual information and discussing the message contents. Once a message is labeled, the behavior of the user is discussed by reflecting their actions, stating whether it is a phishing message, and providing further information and guidance such as missed warning signs.

The `LLM-API` component invokes LLM services or local LLMs. To evaluate LLMs and parameter configurations, multiple LLMs are supported. Local LLMs are especially relevant since they are openly accessible, can be fine-tuned, and further customized. In particular, instruction tuning can be modified, which would otherwise restrict a model, e.g., not allowing phishing message generation. LLM services commonly apply instruction tuning and other restrictions, which limits their application for message generation. In addition, local LLMs allow for privacy and data protection, not exposing user and company information or message contents to services.

3.2 Prototype Implementation

A Python implementation of the architecture[1] has been developed that aims to determine the general feasibility of the approach and serves as a basis for further experimentation and studies.

Training Environment UI The user interface component is realized by a web-based UI with a Flask web-server, communicating with two modules that contain the message simulation and interaction controller components. Figure 2 shows the training environment. Users are presented with a dashboard displaying incoming messages in the fashion of an e-mail inbox, including real business or personal communications in addition to potential phishing threats. Phishing messages as well as legitimate ones were derived from sample messages by the LLM with the aforementioned context, including user and company information. According to the approach, users interact with messages and analyze each e-mail to decide whether to report it as phishing or deem it legitimate. Figure 2

[1] https://github.com/fhaer/adaptive-phishing-awareness-training

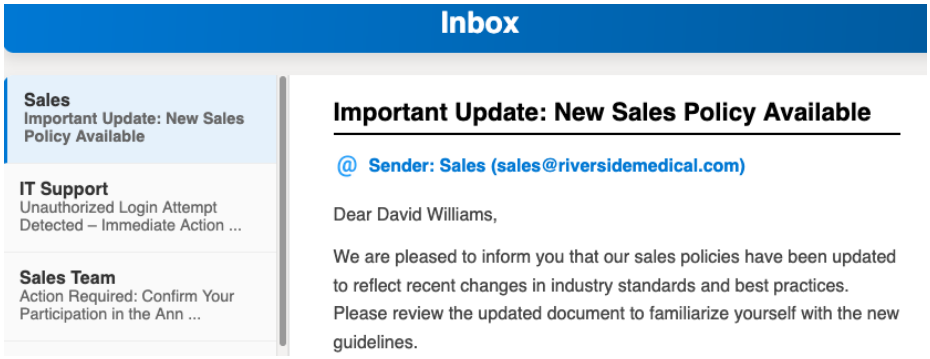

Fig. 2. Prototype user interface simulating an e-mail inbox with phishing and non-phishing messages. A legitimate message is visible, generated by the LLM Qwen 2.5 7B based on a sample message.

illustrates an example where the coach provides feedback after a user did not label a phishing e-mail correctly in part (1). Parts (2) and (3) show the LLM's behavior when answering questions during coaching.

LLM Implementation and Setup Regarding LLM implementation, the LLM-API component utilizes libraries for the OpenAI GPT-4o API and, as an interface to LLMs running locally, the well-known Ollama and "llm"[2] libraries. Common APIs following the OpenAI standards enable the support for further local LLM runtime environments such as LocalAI. Local LLMs are particularly well-suited for this task, as they can run on-device without causing information leakage and do not impose restrictions on phishing message generation.

When executing the platform on a laptop or workstation, local LLMs need to be selected according to memory size and compute performance. In the initial experiments, a MacBook with M2 processor and 24 GiB RAM was running the following state-of-the-art on-device LLMs with 7 to 14 billion parameters: Llama 3.1 8B, Qwen 2.5 7B, Qwen 2.5 14B. These relatively small models were capable of the generation tasks, regarding message generation and dialogue interactions. All models produced convincing messages and helpful coaching interactions. Occasionally, odd phrasing was observed when generating messages adapted from the sample messages by Qwen 2.5 7B; this was not apparent in the other models. Generation time ranged between 10 and 20 seconds for message generation and answers in coaching. In a production environment, LLMs could be run on a company server or workstation with GPU acceleration for near-instant responses, or in a cloud environment that meets data protection requirements. Overall, the initial results indicate that smaller privacy-friendly on-device models can support the adaptive approach in message generation and coaching as demonstrated in Figure 2 and 3.

[2] https://llm.datasette.io/en/stable/

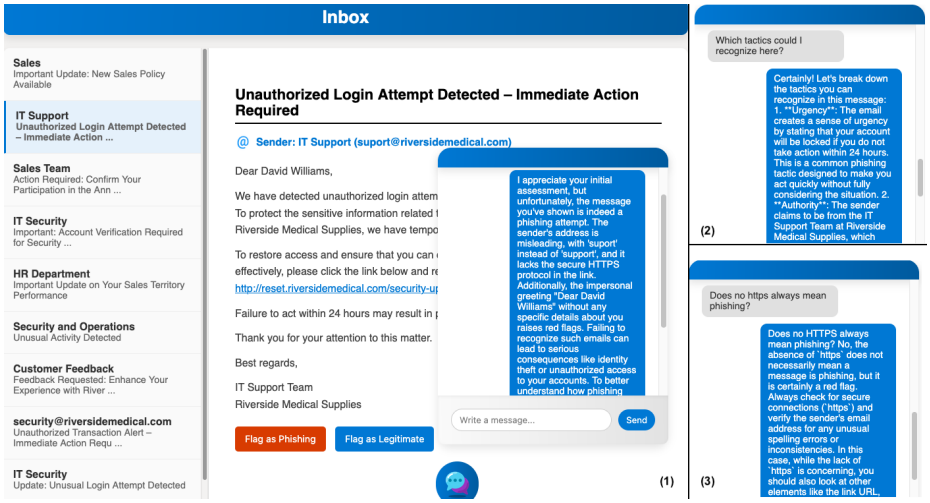

Fig. 3. Prototype interface during a coaching session involving a phishing message generated by the Qwen 2.5 7B LLM. In part (1), the LLM-based coach responds after the user incorrectly labels the message as legitimate. Parts (2) and (3) show the coach answering follow-up questions.

4 Conclusion and Future Work

This study explored the approach, architecture, and prototype implementation of an adaptive phishing awareness training system. The initial prototype demonstrates the feasibility of applying Large Language Models (LLMs) for adaptivity and interactive learning. In particular, the results show that LLMs can be applied in phishing awareness training for (1) generating messages targeted at specific users and (2) conducting coaching sessions that adapt to users and help them recognize phishing by discussing individual messages and tactics. The tested open source LLMs, running locally, were found to be generally suitable and beneficial in terms of privacy and data protection compliance. Overall, the LLM-based approach enables adaptivity, immediate feedback, and interaction, indicating potential for cybersecurity awareness training beyond traditional online training. In future research, it is planned to develop the approach and prototype further, conduct a user study, and begin experiments on the effectiveness and efficiency of adaptive awareness training methods.

References

1. Anti-Phishing Working Group (APWG): Phishing Attack Trends Report – 2Q 2024. Tech. rep. (2024), `https://apwg.org/trendsreports/`
2. Blancaflor, E., Calpo, A.H., Cebrian, S.J., Siquioco, F.: A Comprehensive Review of Neural Network-Based Approaches for Predicting Phishing Websites and

URLs. In: 2024 5th International Conference on Industrial Engineering and Artificial Intelligence (IEAI). pp. 96–101. IEEE, Bangkok, Thailand (2024). https://doi.org/10.1109/IEAI62569.2024.00025

3. Goldin, I., Narciss, S., Foltz, P., Bauer, M.: New Directions in Formative Feedback in Interactive Learning Environments. International Journal of Artificial Intelligence in Education **27**(3), 385–392 (Sep 2017). https://doi.org/10.1007/s40593-016-0135-7
4. Jampen, D., Gür, G., Sutter, T., Tellenbach, B.: Don't click: towards an effective anti-phishing training. A comparative literature review. Human-centric Computing and Information Sciences **10**(1), 33 (Aug 2020). https://doi.org/10.1186/s13673-020-00237-7
5. Lin, Z., Chan, C., Song, Y., Liu, X.: Constrained Reasoning Chains for Enhancing Theory-of-Mind in Large Language Models. In: Hadfi, R., Anthony, P., Sharma, A., Ito, T., Bai, Q. (eds.) PRICAI 2024: Trends in Artificial Intelligence. pp. 354–360. Springer Nature, Singapore (2024). https://doi.org/10.1007/978-981-96-0119-6_34
6. Sai, S., Yashvardhan, U., Chamola, V., Sikdar, B.: Generative AI for Cyber Security: Analyzing the Potential of ChatGPT, DALL-E, and Other Models for Enhancing the Security Space. IEEE Access **12**, 53497–53516 (2024). https://doi.org/10.1109/ACCESS.2024.3385107
7. Stranieri, A., Yearwood, J.: Enhancing learning outcomes with an interactive knowledge-based learning environment providing narrative feedback. Interactive Learning Environments **16**(3), 265–281 (Dec 2008). https://doi.org/10.1080/10494820802114176
8. Tandale, K.D., Pawar, S.N.: Different Types of Phishing Attacks and Detection Techniques: A Review. In: 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC). pp. 295–299. IEEE, Aurangabad, India (2020). https://doi.org/10.1109/ICSIDEMPC49020.2020.9299624
9. Waghmare, C.: Enhancing Business Communication with ChatGPT. In: Unleashing The Power of ChatGPT, pp. 79–92. Apress, Berkeley, CA (2023). https://doi.org/10.1007/979-8-8688-0032-0_4
10. Wen, Z.A., Lin, Z., Chen, R., Andersen, E.: What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12. ACM, Glasgow Scotland Uk (2019). https://doi.org/10.1145/3290605.3300338, https://dl.acm.org/doi/10.1145/3290605.3300338