Simulating AI-Driven Social Engineering Attacks in Ethical Hacking Using Microsoft 365 Defender

Yazan M. Abu-Aisheh, Mousa M. Farajallah,Ibrahim R Al-Sharif

***Abstract*— Phishing remains one of the most prevalent and effective social engineering attacks, primarily targeting human psychology rather than technical loopholes [1]. The recent development of artificial intelligence (AI), especially with advanced language models like OpenAi's GPT-4, has provided cybercriminals as well as cybersecurity experts with powerful tools to create exceptionally realistic phishing messages [2]. The present research presents a comparative evaluation of AI-generated phishing emails compared to conventional phishing email simulations in the context of an ethical hacking exercise, using Microsoft 365 Defender for Office 365 [3]. The customized and dynamic phishing emails were prepared using GPT-4, while the traditional static templates used were representative of common attack tactics. The experimental setup entailed a pilot group of 300 participants and measured several metrics, such as email opening rate, click-through rate (CTR), credential submission, realism perception, and improvement in awareness. The results show that AI-generated phishing emails outperformed conventional attacks across all the criteria measured, with a 48% CTR and a 34% increase in post-training awareness. The results highlight the increased realism and potency of AI-based phishing simulations [4], while also emphasizing improved, experiential security training in organizations. Future research projects will look to increase the number of participants, include statistical confirmation, and explore other modalities like voice and SMS phishing (vishing and smishing).**

I. INTRODUCTION

Phishing and social engineering attacks continue to evolve, posing significant threats to cybersecurity by taking advantage of the behavior of humans instead of purely Internet-based technology vulnerabilities. CISA has continually highlighted the central role that social engineering plays as a prime method of cyber infiltration, particularly by email phishing [5]. These attacks often utilize emotional triggers such as urgency or fear to induce victims to provide sensitive information or open malicious attachments [6].

The advent of generative artificial intelligence, particularly in the form of large language models (LLMs), has also given malicious actors the capability to generate advanced and context-sensitive phishing content in bulk [7]. Along with this, ethical hackers and information security experts leverage the same AI technologies for the development of simulated phishing campaigns, thus helping organizations to detect weaknesses in human behavior and improving awareness training programs.

This paper addresses these shortcomings by introducing and evaluating an end-to-end simulation solution that integrates AI-based phishing emails into Microsoft 365 Defender's Attack Simulation Training. The solution monitors critical behavioral metrics such as email open rates, click-through rates, and credential submission attempts and automatically offers targeted awareness training to vulnerable users. It seeks to build organizational social engineering resilience by leveraging AI not only as a future threat, but as a powerful tool for ethical, instructional simulations.

II. LITERATURE REVIEW

Social engineering attacks, and phishing more broadly, have long been one of the usual topics of concern in the field of cybersecurity. Traditional approaches to simulating such forms of attacks in ethical hacking exercises used to heavily depend on hand-authored phishing emails or employing pre-built phishing templates in applications like GoPhish or the Microsoft 365 Defender Attack Simulator. Artificial Intelligence (AI), and specifically the use of large language models (LLMs), has, however, introduced different dynamics to simulate and investigate social engineering attacks.

Early social engineering simulation research focused on human-designed phishing campaigns. For example, research on user vulnerability to phishing in the form of emails and their content was performed by Alsharnouby et al. [8]. Tools like GoPhish offered automation of campaign distribution but not adaptability in content generation [9].

With the arrival of LLMs such as GPT-3 and GPT-4, researchers began to experiment on how to employ these models to generate false messages automatically. In [10], Bhatia et al. utilized GPT-3's capacity for crafting phishing emails that are linguistically credible. The study discovered that AI-generated emails were more persuasive and grammatically sound than traditional phishing kits. Dai et al. [11] also proposed an automated phishing model based on Open AI models and demonstrated that they can bypass spam filters.

Nguyen and Lin [12] took this a step further by investigating the realism of AI-generated phishing emails through user studies. Through their results, they highlighted how over 65% of users were unable to distinguish AI-written phishing emails from genuine ones. This not only indicates the potential for harm from these tools when used for nefarious purposes but also their potential in ethical hacking training.
For simulation software, Microsoft 365 Defender comes equipped with a pre-installed phishing attack simulator. Its templates are static, though. Recent research proposes

978-9950-8599-0-6

enhancing such software with AI to create adaptive, context-dependent email content, thus making it more realistic in user awareness testing [13].

A number of studies attempted to compare AI-based simulations with traditional methods as well. Kumar et al. [14], for instance, compared rates of clicks on GPT-constructed phishing emails with those written manually and observed a significant increase in user interaction with AI-constructed emails, particularly spear-phishing.

In contrast, this research addresses these difficulties through the implementation of AI-phishing simulations with GPT-4 in Microsoft 365 Defender. Not only is it feasible to achieve realism in phishing emails, but real-time logging of user interactions, automated assignment of training, and feedback through behavior are also achieved—features not common to earlier research. Through the comparison of real-world performance on the same indicators in an enterprise environment, this paper offers actionable data on the employment of defensive AI in ethical hacking.

III. Methodology and Framework

This chapter outlines the conduct of phishing attack simulations with the aid of artificial intelligence in the context of ethical hacking using Microsoft 365 Defender for Office 365. The simulation had two main purposes: first, to test users' susceptibility to advanced AI-created phishing content, and second, to measure the effectiveness of the built-in security awareness training. The approach was divided into four phases: phishing content generation, campaign setup, simulation run, and post-simulation evaluation.

A. AI-Generated Phishing Content

The aim of this simulation was to create highly realistic phishing emails based on the OpenAI GPT-4 language model. The emails were designed to bear strong resemblance to actual phishing use, e.g., password reset schemes, urgent business announcements, or cautionary notices to check shared documents. The templates were tailored to precise branding, tone, and content so that they were as realistic as possible.

To achieve this, researchers used targeted prompts to instruct GPT-4 to produce authentic-sounding messages. For example, prompts had instructions like: “Compose an email from the HR department, asking the recipient to update their password ASAP in a professional tone,” or “Compose an urgent notice requesting the user to check a shared invoice.” The prompts were crafted with the help of cybersecurity specialists to ensure emails appeared authentic and relevant to actual circumstances.

Each phishing email included the following elements:

- Personalization components (recipient's name, job title, business logo)
- phishing-style call to action (e.g., “Check your invoice”, “Re-set your password”
- An innocuous, but trackable, URL in a Microsoft simulated environment

The phishing content, made possible by AI, varied and was more contextually specific compared to traditional static templates.

B. Simulation Environment: Microsoft 365 Defender for Office 365

Microsoft Office 365 Defender was leveraged in order to deliver the simulations through its Attack Simulation Training feature. It is built into Microsoft Outlook and Exchange to enable simulated phishing email to be sent to real user mailboxes in a controlled, safe environment. User behavior—email opens, link clicks, and credential input—were automatically logged.

Microsoft Defender was selected due to:

- Native analytics for measuring campaigns
- Compliant, risk-free simulation that doesn't involve revealing real credentials
- Automatically assigned post-simulation training modules
- It offers repeat campaigns and user targeting

This setup enabled easy simulation without the need for supporting infrastructure or privacy concerns with open-source solutions.

C. Evaluation Metrics

Some of the main measurements for both learning achievement and behavioral change taken from Microsoft Defender’s dashboard and subsequent user surveys included:

- Email Open Rate: Percentage of recipients who opened the phishing email
- Click-Through Rate (CTR): Percentage that clicked on the phishing embedded link
- Credential submission rate: percentage who tried to log in on phishing landing page
- Time-to-Click: Average time gap between getting email and interacting
- Realism rating: How realistic the participants evaluated the email as being (post-simulation questionnaire)
- Awareness Enhancement: Change in pre- and post-training quiz scores as a measure of learning

These tests also gave quantitative and qualitative information on how the participants responded to the AI-powered phishing simulations.

D. Awareness Training and Feedback

Microsoft Defender automatically enrolled those individuals who had interacted with the phishing messages (i.e., opened a link or entered credentials) in short training sessions on phishing signs, email authentication practices, and reporting incidents.

In addition, a specially crafted feedback questionnaire—distributed through Microsoft Forms—was also used in order to obtain subjective data regarding realism of the phishing content and worth of the training.

To measure knowledge retention, pre-and post-simulation phishing awareness tests were given. The tests measured knowledge and the ability to identify typical phishing indicators and proper user actions, with scores compared to identify awareness increase rates.

IV. Results

The phishing simulation evaluation was conducted with a pilot sample of 300 users in a corporate environment. Two campaigns were run using the Microsoft 365 Defender platform: one using phishing emails created by GPT-4, and the other using conventional template-based phishing messages. The following table provides an overview and comparative analysis of the key metrics.

The findings that followed were:

Metric	**AI-Generated Phishing**	**Traditional Phishing**
Email Open Rate	92%	78%
Click-Through Rate (CTR)	48%	31%
Credential Submission Rate	26%	14%
Average Time-to-Click	3 minutes	4.5 minutes
Realism Rating	4.3 / 5	3.2 / 5
Awareness Improvement Score	+34%	+19%

key findings:

- The 92% email open rate confirms that the subject line and sender details in AI-composed emails were especially compelling.
- The click-through rate (48%) indicates a high rate of user engagement with AI-created emails due to contextual and emotional cues.
- Credential Submission (26%): Highlights that artificially generated content was more deceptive, which led to users revealing their credentials.
- Time-to-Click (3 mins): Shows impulsivity, with faster reaction times for AI emails.
- Realism Rating (4.3/5): Corroborates the accepted authenticity of artificially generated emails.
- Awareness Improvement (+34%): Demonstrates the effectiveness of simulation and training in improving phishing recognition.

V. Discussion

The results establish that phishing emails composed using GPT-4 are more effective at simulating real-world attacks than traditional template-based emails. The high rates of openings and clicks highlight the effective features of AI-composed content, which mimics human styles of writing and leverages psychological triggers, such as urgency and authority.

The 26% observed rate of credential submission represents a high risk, implying that over one-quarter of users were prone to content artificially generated. These findings reiterate the need for ongoing, interactive training—particularly in environments where staff routinely deal with emails.

Users determined that emails produced by artificial intelligence were substantially more genuine than traditional phishing attacks. Their perceived authenticity, combined with the contextual and emotional importance of the messages, made them more difficult to distinguish from authentic emails.

The observed improvement of +34% in awareness scores highlights the efficacy of the combined training model employed by Microsoft 365 Defender. These results align with modern research in experiential learning, which suggests that active interaction with threats supports stronger long-term retention than passive methods of instruction.

Finally, the shorter average time-to-click for AI phishing emails suggests impulsivity plays a critical role in vulnerability. Future training should emphasize skepticism toward time-sensitive or emotionally charged messages.

VI. Conclusion and Future Work

This research demonstrates the effectiveness of using AI-powered phishing simulation in the framework of ethical

hacking with Microsoft 365 Defender for Office as a trustworthy and enterprise-level application. Using advanced language models for generating convincing phishing attacks enabled the simulation of authentic cyber-attacks, which mirrored the tactics adopted by real attackers. The results reflected high user interaction with a measurable increase in security awareness through targeted training, thus legitimizing the method as being both efficient and practical.

The combined capabilities of Microsoft 365 Defender, including user activity monitoring, provision of timely feedback, and protection for data, played a critical role in the successful conduct of the simulation. The findings identify the possibility for enhancing the cybersecurity posture for organizations by integrating artificial intelligence with established forms of security in a governed and ethical framework.

In future studies, expansion of the parameters to explore a diverse variety of unique user populations from different industries can provide a greater understanding of trends in vulnerabilities. Additionally, future studies could include a comparative evaluation of traditional methods of phishing with AI-initiated communications in order to determine their effectiveness. Of further value is examining the simulation of two alternate social engineering methods, i.e., voice phishing (vishing) and SMS phishing (smishing), together with exploration into the long-term effects of training interventions. As AI continues to evolve, its dual role as both risk and protector will be a vital aspect in the discipline of cybersecurity.

REFERENCES

[1] M. JAKOBSSON AND P. FINN, SOCIAL ENGINEERING: THE SCIENCE OF HUMAN HACKING, 2ND ED. HOBOKEN, NJ, USA: WILEY, 2018.

[2] N. BHATIA, S. JAIN, AND P. SHARMA, "USING AI TO ENHANCE PHISHING AWARENESS TRAINING: A COMPARATIVE STUDY OF EMAIL REALISM AND USER RESPONSE," J. CYBERSECURITY DIGIT. TRUST, VOL. 8, NO. 3, PP. 111–123, 2022.

[3] MICROSOFT, "ATTACK SIMULATION TRAINING IN MICROSOFT DEFENDER FOR OFFICE 365," MICROSOFT LEARN, 2024. [ONLINE]. AVAILABLE: HTTPS://LEARN.MICROSOFT.COM/EN-US/MICROSOFT-365/SECURITY/OFFICE-365-SECURITY/ATTACK-SIMULATION-TRAINING

[4] A. HADNAGY, *SOCIAL ENGINEERING: THE ART OF HUMAN HACKING*, INDIANAPOLIS, IN, USA: WILEY, 2010.

[5] CISA, "PHISHING GUIDANCE," CYBERSECURITY AND INFRASTRUCTURE SECURITY AGENCY, 2023. [ONLINE]. AVAILABLE: HTTPS://WWW.CISA.GOV/NEWS-EVENTS/PHISHING-GUIDANCE

[6]A. JAIN AND B. B. GUPTA, "PHISHING DETECTION: ANALYSIS OF VISUAL SIMILARITY BASED APPROACHES," *SECURITY AND PRIVACY*, VOL. 1, NO. 1, PP. E9, JAN. 2018. DOI: 10.1002/SPY2.9

[7] M. HARWELL AND D. MALIK, "AUTOMATING CYBER ATTACKS WITH AI: IMPLICATIONS FOR SOCIAL ENGINEERING," *CYBER DEFENSE REVIEW*, VOL. 7, NO. 1, PP. 50–65, 2023.

[8] N. Alsharnouby, F. Alaca, and S. Chiasson, "Why phishing still works: User strategies for combating phishing attacks," *International Journal of Human-Computer Studies*, vol. 82, pp. 69–82, Apr. 2015.

[9] GoPhish: Open-Source Phishing Toolkit. [Online]. Available: https://getgophish.com/

[10] R. Bhatia, A. Jain, and A. Sharma, "Harnessing GPT-3 for Simulating Phishing Attacks: An Empirical Evaluation," *IEEE Access*, vol. 10, pp. 106324–106335, 2022.

[11] M. Dai, H. Li, and Y. Chen, "Phishing with AI: A Threat Assessment of GPT-based Attacks," in *Proc. of the 2023 IEEE Symposium on Security and Privacy Workshops*, pp. 201–208.

[12] T. Nguyen and D. Lin, "Can AI Fool You? Evaluating the Realism of GPT-generated Phishing Emails," in *Proc. of the ACM CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–12.

[13] Microsoft 365 Defender Attack Simulator Documentation. [Online]. Available: https://learn.microsoft.com/en-us/microsoft-365/security/office-365-security/attack-simulation-training-overview

[14] A. Kumar, S. Alavi, and T. Soh, "Comparative Study of Traditional vs AI-Based Phishing Simulation in Enterprise Environments," *Journal of Cybersecurity Technology*, vol. 7, no. 2, pp. 105–123, 2023.