RETHINKING PHISHING AWARNES: EVALUATING THE IMPACT OF OPEN-SOURCE AI-SUPPORTED TRAINING TOOLS

By

EHAB BEDIR

B.S., University of Cairo - Egypt
M.S., University of Derby - UK
M.S., Columbus State University - USA

A Research Paper Submitted to the School of Computing Faculty of
Middle Georgia State University in
Partial Fulfillment of the Requirements for the Degree

DOCTOR OF SCIENCE IN INFORMATION TECHNOLOGY

MACON, GEORGIA

2026

Rethinking phishing awareness: evaluating the impact of open-source AI-supported training tools

Ehab Hekal Bedir, *Middle Georgia State University,* ehab.bedir@mga.edu

Abstract

Phishing remains a critical cybersecurity threat that exploits human decision-making through deceptive email-based social engineering. This quantitative study evaluated the effectiveness of an open-source, AI-supported phishing awareness training workflow by measuring participants' behavioral responses to legitimate and simulated phishing emails of increasing difficulty across two phases. An open-source simulation framework was implemented using 'GoPhish' to deliver controlled email campaigns and record behavioral outcomes, including correct decision outcomes, link-click behavior, and phishing reporting behavior. Artificial intelligence was used at the content design level through large language model-assisted generation of phishing email scenarios. A post-simulation 5-point Likert-scale survey was administered to assess participant engagement and perceived training effectiveness. Behavioral and survey outcomes were interpreted in relation to effectiveness patterns reported in the literature on commercial phishing awareness platforms. Although the present study did not implement a fully adaptive AI-supported system, the workflow suggests how an open-source platform can support measurable behavioral improvement while offering architectural flexibility for future API based adaptive extensions. The study aimed to provide evidence-based guidance for organizations considering open-source AI-supported phishing awareness tools as a practical training option, particularly in resource-constrained settings.

Keywords: phishing, social engineering, AI-supported training, cybersecurity awareness, open-source tools, commercial phishing awareness platforms.

Introduction

Phishing attacks continue to pose a substantial threat in the digital landscape, targeting individuals and organizations by exploiting human vulnerabilities through fraudulent emails, malicious links, and other deceptive tactics (Eze & Shamir, 2024). As these attacks become more sophisticated and often enhanced by artificial intelligence (AI), they have become more convincing and harder to detect, necessitating more advanced defense strategies. Traditional awareness training approaches have proven inadequate and struggle to keep pace with the evolving tactics of cybercriminals (Eze & Shamir, 2024).

In response to this challenge, AI-supported technologies, with their ability to provide personalized and adaptive learning experiences, offer a promising alternative for enhancing phishing detection and awareness training (Ansari et al., 2022). Jain (2025) further demonstrated that adaptive AI-based phishing simulations reduced employee click rates on phishing emails by 45% and improved incident-reporting speed by 89% in his organization, highlighting the practical effectiveness of behavior-driven, gamified learning. Similarly, Grimes and Kraemer (2023) analyzed data from the commercial KnowBe4 tool over 60,000 organizations and 32 million users, finding that frequent simulated phishing tests combined with continuous security awareness training improved phishing detection rates by up to 96% and reduced user susceptibility to below 5% within one year, providing large-scale empirical evidence of the effectiveness of commercial AI-supported training programs.

However, while such commercial AI-supported solutions have shown measurable success, their high costs and proprietary nature limit accessibility for many smaller, resource-constrained organizations. Traditional phishing awareness training approaches also remain largely generic and static, making it difficult for users

to keep pace with evolving attacker tactics (Greco et al., 2025). Although AI training shows strong potential through adaptive learning and behavioral modeling (Ansari et al., 2022; Sumner et al., 2022), and applied examples demonstrate meaningful reductions in risky user actions (Jain, 2025), the benefits of these approaches are not equally accessible across organizational contexts (Dewis & Viana, 2022). This creates a practical and research-based gap. Organizations with limited resources need effective and affordable phishing awareness solutions; however, the accessibility and effectiveness of open-source AI-supported training tools remain underexamined in academic research.

Open-source platforms such as GoPhish and King Phisher offer flexible, low-cost alternatives for phishing awareness training, yet there is limited empirical evidence that they improve real user behavior (Luse & Burkman, 2021; Young & Farshadkhah, 2023). Existing research largely emphasizes technical detection systems or commercial training programs, leaving open-source AI-supported approaches insufficiently validated through controlled behavioral measurement. As a result, it remains unclear whether these tools can produce benefits comparable to those reported for commercial AI-supported platforms. In the present study, AI is applied at the content design level through large language model (LLM) generated phishing scenarios rather than as a fully adaptive training engine, while the platform retains the potential for future adaptive extensions through API-based integration and external workflow control.

Accordingly, this quantitative study evaluated an open-source AI-supported phishing awareness workflow by examining the effect of the workflow on users' ability to detect and respond to phishing attempts. Behavioral and perception outcomes were used to assess practical effectiveness and relevance for resource-constrained organizations. This study addressed the following research questions:

RQ1. To what extent do open-source AI-supported phishing awareness tools improve users' detection-related responses to phishing attempts?
RQ2. How do clicking and reporting patterns across the training sequence reflect participant learning and response behavior?
RQ3. To what extent do participants' post-training engagement and perceived effectiveness scores indicate positive perceptions of the open-source AI-supported workflow in relation to outcome patterns reported in the commercial AI platform literature?

Literature Review

Protection Motivation Theory provides a useful interpretive lens for understanding how simulation-based phishing awareness training may influence user behavior. PMT proposes that protective action is shaped by two related appraisal processes: threat appraisal, which reflects perceived severity and vulnerability, and coping appraisal, which reflects beliefs about response effectiveness and self-efficacy (Marikyan & Papagiannidis, 2023). In phishing contexts, repeated simulated exposure may strengthen threat recognition, while guidance, feedback, and practice may increase confidence in avoiding suspicious links and reporting malicious messages. This interpretation is consistent with evidence that simulation-based and structured educational interventions can improve user resilience and protective decision-making over time (Horvat & Radić, 2024). In the present study, Phase 1 provided exposure, feedback, and corrective learning opportunities, whereas Phase 2 assessed whether safer behaviors were retained without immediate prompts. PMT is therefore used as an interpretive lens rather than as a directly tested model.

Beyond behavioral theory, cybersecurity research highlights the complementary role of technical defenses. AI-based phishing detection systems use machine learning to analyze URLs, email content, and user behavior to distinguish phishing attempts from legitimate messages (Sameen et al., 2020; Alsariera et al., 2020). Prior work has reported strong performance using real-time URL detection, hybrid ensemble models, and architectures that combine structural and content-based features (Sameen et al., 2020; Alsariera et al.,

2020; Chen & Chen, 2021). More recently, Lim et al. (2025) introduced EXPLICATE, which combines large language models with explainable AI to generate human-readable explanations for phishing classifications. Although these advances strengthen technical defenses, the human element of phishing risk remains. Human awareness remains central to effective cybersecurity, particularly in higher education and similar environments where social engineering exploits user behavior (Mouwers-Singh & Musikavanhu, 2024; Afolalu & Tsoeu, 2025). Phishing susceptibility is also shaped by emotional and cognitive biases, as users often rely on heuristics rather than systematic evaluation (Goel et al., 2017). Accordingly, effective training should address cues such as legitimacy, authority, and message framing, not only technical indicators of malicious content (Desolda et al., 2021). These findings support the combination of AI detection systems with sustained awareness and training interventions.

This need for integrated, human-centered defense is reflected in the growing literature on AI-supported awareness programs. AI-supported training tools can improve engagement and detection skills through adaptive and personalized learning environments (Ansari et al., 2022). Commercial platforms report that behavior-based personalization and continuous simulation improve engagement and retention (CybeReady, 2024), and large-scale evidence suggests that frequent simulations combined with ongoing awareness training can substantially reduce phishing susceptibility (Grimes & Kraemer, 2023). Ho et al. (2025) likewise found that static awareness formats produced minimal long-term benefit, whereas interactive and context-specific simulations generated modest improvement. Gamified and adaptive simulations have also been linked to stronger engagement and knowledge retention (Jayakrishnan et al., 2022). At the same time, current awareness research rarely distinguishes between simulations using manually written phishing content and those using AI or large language models to generate phishing messages (Greco et al., 2025). Although explainable AI may improve trust in automated systems, these approaches have primarily been developed for analysts rather than end-user training contexts (Lim et al., 2025; Xue et al., 2025). Consequently, there is still limited evidence on whether exposure to LLM-generated phishing emails improves detection accuracy, reporting behavior, or decision-making speed in training environments.

Open-source tools such as GoPhish and King Phisher offer cost-effective and customizable alternatives, making them attractive to non-profits, small businesses, and other resource-constrained organizations (GitHub, 2024; King Phisher, 2024; Khan, 2024; Ansari et al., 2022). However, open source tools often require greater technical expertise and may lack advanced analytics, real-time feedback, and adaptive features commonly associated with commercial platforms (Khan, 2024; Sumner et al., 2022). Commercial platforms such as KnowBe4, Guardz, Phished.io, and Phish Firewall emphasize integrated reporting, automated simulations, post-click learning modules, adaptive learning paths, and progress dashboards, although higher costs may limit adoption among smaller organizations (Guardz, 2025; Phished.io, 2024; Phish Firewall, 2024; Dewis & Viana, 2022). Comparative reviews similarly conclude that open-source tools are often more affordable and adaptable, whereas commercial tools tend to offer stronger support and more advanced features, though such comparisons usually focus on operational characteristics rather than peer-reviewed behavioral outcomes (Khan, 2024).

Emerging evidence, nevertheless, suggests that open-source platforms can support meaningful awareness outcomes. Abdykerimov et al. (2025), in a university-based phishing simulation study, reported substantial user interaction with simulated phishing emails and highlighted the importance of non-punitive educational designs. The authors also identified the practicality of tools such as GoPhish and King Phisher for awareness and behavioral measurement. Practitioner evidence likewise suggests value in AI-supported simulations. Jain (2025) reported reduced employee susceptibility and successful breaches following adaptive phishing simulations in a financial organization. Although practitioner findings should be interpreted with caution, the findings align with broader evidence that continuous feedback, adaptive learning, and human-centered training can strengthen organizational resilience (Sumner et al., 2022; Mouwers-Singh & Musikavanhu, 2024). Still, comparative evidence between open-source and commercial platforms remains limited,

particularly in peer-reviewed studies that focus on behavioral outcomes rather than platform features alone. These gaps provide the foundation for the present quantitative study, which evaluates an AI-supported open-source phishing awareness approach while interpreting its outcomes in relation to findings reported for commercial and traditional approaches in prior research.

Methodology

This quantitative study used a single-group repeated-measures design to examine how participants responded to legitimate and simulated phishing emails that varied in difficulty across a two-phase phishing awareness workflow. The goal was to assess behavioral outcomes associated with AI-supported training content delivered through an open-source simulation platform and to interpret those outcomes in light of findings reported in the commercial phishing-awareness literature.

To maintain institutional safety and methodological rigor, all technical infrastructure was kept separate from university systems and managed independently by the researcher. A dedicated research domain and authenticated outbound email configuration were used to support auditable message delivery without relying on institutional mail servers. GoPhish was installed and operated locally on the researcher's laptop and served as the platform for email delivery and behavioral logging. Because GoPhish was developed with a JSON API and official Python client support, GoPhish also offers flexibility for future automation and adaptive integrations beyond its standard interface (GoPhish, n.d.). The platform supported controlled email delivery and automated logging of link clicks and response timing, while email open events were treated as a secondary indicator because of client-side image handling and privacy protections. Artificial intelligence was used at the content design level, via a large language model (LLM), to generate phishing email code elements of varying difficulty by adjusting features such as ambiguity, implied authority, and urgency.
.

Sampling and Recruitment

This study used purposive sampling to recruit participants from local and international educational institutions and small companies. A total of 188 participants were recruited, of whom 116 met the interaction-based inclusion threshold and were retained for the primary inferential analyses. Participants included both students and employees with a minimum age of 18 years, allowing the training workflow to be examined across contexts that face similar phishing risks but may differ in training access and everyday exposure. The primary analysis focused on behavioral outcomes across training phases within a single-group design rather than subgroup comparisons across roles or institutional contexts. This approach aligns with purposive sampling principles, which emphasize selecting participants who are well-suited to the study's objectives and can provide relevant evidence for the research questions (Campbell et al., 2020). Because participants came from varied backgrounds, baseline phishing familiarity was not standardized, and the findings should therefore be interpreted as ecological rather than tightly controlled behavioral observations.

Procedures

The study was conducted in two phases using the same participants, without random assignment or a separate control group. In Phase 1, participants received one legitimate control email and three simulated phishing emails of increasing difficulty. The Phase 1 landing pages included informational guidance and a non-sensitive decision prompt to support reflection after initial engagement. In Phase 2, simulated phishing emails of increasing difficulty were delivered without corrective prompts to assess learning outcomes following exposure in Phase 1. Across both phases, behavior was recorded through GoPhish logs of clicks and timing, while reporting behavior was captured through the embedded reporting mechanism and

designated reporting inbox. No credentials were requested, no malware or attachments were used, and all interactions were designed to remain low-risk and ethically compliant.

Email exposure occurred in a controlled session with sequential delivery and immediate survey completion to reduce recall bias and limit peer discussion. Automated OSINT-based personalization was excluded to minimize privacy risk and maintain IRB-aligned safeguards. Commercial phishing awareness platforms were not tested, and comparisons were limited to contextual interpretation through the literature.

Data Collection Tools

A- Phishing Simulations: Behavioral data were collected through GoPhish campaign logs and the study reporting workflow. GoPhish provided link-click outcomes and click-timing data for each message exposure. Reporting behavior was captured via the designated reporting mechanism embedded in the email interface and recorded in the reporting inbox, with timestamps and reference identifiers.

B- Perceptions and Experiences Survey (Likert Scale): A 5-point Likert-scale questionnaire was administered to assess user engagement, perceived effectiveness, and overall experience with the training.

The survey items were adapted from previously validated instruments in human-computer interaction and information systems research. Questions 1 through 4 were adapted from the User Engagement Scale developed by O'Brien and Toms (2010), which measures focused attention, perceived reward, and involvement in digital environments. The wording was contextualized for phishing awareness training while preserving the original intent of the scale to capture absorption, time distortion, and perceived value. Questions 5 through 7 were adapted from the personalization and perceived usefulness measures used by Tam and Ho (2006). These items assess whether users perceive the content as tailored to user needs and responsive to user behavior, and they were rephrased to reflect adaptive training scenarios in phishing awareness. Questions 8 and 9 were developed for this study to assess the perceived usefulness of the training, confidence in detecting phishing emails, and perceived relevance for organizational adoption. These items were conceptually aligned with perceived efficacy and capability constructs commonly used in technology acceptance and training research, while being tailored to AI-supported phishing awareness training. Since the items were adapted and combined from existing validated instruments, internal consistency, including Cronbach's alpha for each subscale, was examined using the research data to provide evidence of reliability and construct alignment in the context of phishing awareness.

Data Preparation and Analysis Plan

Behavioral data from GoPhish campaign exports were imported into SPSS and organized at the email exposure level, with each row representing one participant's exposure to one email condition across the eight email sequences. Survey responses were retained in a separate anonymous dataset because no personal identifiers were collected, thereby preserving confidentiality. As a result, behavioral and survey outcomes were analyzed as complementary datasets rather than linked at the individual level.

Behavioral variables were coded numerically for analysis. Binary indicators were created for key outcomes, including clicks, reports, and email opens, as well as for design variables such as phase, email type, email difficulty, and exposure order. Additional derived variables were created to support the interpretation of engagement and learning, including interaction occurred, corrective behavior evidence, and correct decision. Correct decision was defined as whether the participant's observed response aligned with the email condition, allowing appropriate responses to both simulated phishing emails and legitimate control emails to be evaluated consistently.

A participant-level summary file was then created using the SPSS Aggregate procedure. This dataset produced one record per participant and calculated totals across exposures, including total exposures, measurable interactions, clicks, reports, and correct decisions. It was used to assess participation completeness and apply the inclusion threshold for the primary inferential analyses.

Initial screening included frequency distributions and descriptive statistics to verify coding accuracy, examine missing data, and assess overall data quality. Delivery success and delivery error variables were reviewed as quality control checks to distinguish behavioral nonresponse from technical delivery failure. The timing variable was also examined to determine whether parametric assumptions were reasonable and whether nonparametric procedures were more appropriate for timing-related analyses.

Inferential analyses were selected based on variable measurement level and the research questions, consistent with a quantitative approach in which measurable variables are analyzed statistically (Creswell & Creswell, 2023). Because the study was designed as a single-group exploratory evaluation rather than a comparative trial, the findings reflect training-associated change rather than superiority over traditional or commercial training models. Categorical behavioral outcomes, including clicks, reports, and correct decisions, were examined using cross-tabulations and chi-square tests across phase and email difficulty. For the principal chi-square analyses, Cramér's V was used to estimate effect size and support the interpretation of the strength of association. Survey items were coded on a 5-point Likert scale, and composite mean scores were calculated to represent key perception constructs. Internal consistency was assessed using Cronbach's alpha before reporting the composite results.

Inclusion and Exclusion Criteria

For the primary inferential behavioral analysis, an interaction-based inclusion threshold was applied to ensure meaningful exposure to the training intervention. Participants were eligible for the primary analysis if they demonstrated at least four measurable interactions across the eight email exposures. This threshold reflected sufficient engagement with the simulation sequence and repeated opportunities to interact with the training workflow and feedback elements intended to support phishing recognition and decision making. Participants who did not meet this threshold were retained in the full dataset for transparency and descriptive reporting but were excluded from the primary effectiveness analyses due to limited engagement.

Variables

The behavioral dataset included variables designed to capture participant responses across the eight email exposures. Each record was associated with an anonymous participant identifier to allow all exposures for a given participant to be grouped during analysis. Study condition variables included phase and email difficulty, with phase distinguishing the first and second training periods and difficulty ranging from the legitimate control condition to high difficulty phishing conditions. The primary behavioral outcome variables were click behavior, report behavior, and correct decision, each coded as binary indicators. Correct decision was defined as whether the participant's observed response was appropriate for the email condition, allowing consistent evaluation of both simulated phishing emails and the legitimate control email. Additional indicators were included to support the interpretation of participant response patterns. Interaction occurred was used to identify exposures with measurable participant activity and to restrict the main inferential analyses to exposures in which an observable response was recorded. Evidence of corrective behavior was included as a process indicator to capture learning-related responses following an unsafe action. Data quality variables, including send success and delivery error flag, were also reviewed to verify that nonresponse was not primarily attributable to delivery failure.

The perception dataset included post-simulation survey variables used to assess participants' views of the training experience. Primary survey items measured perceived usefulness, confidence in detecting phishing,

and perceived value for organizational adoption. These items were coded on a 5-point Likert scale and were also combined into a composite mean score representing perceived usefulness and effectiveness. Together, these variables supported the interpretation of participant perceptions following completion of the phishing awareness workflow.

Results

The results are organized around the study's three research questions. RQ1 examined whether the open-source, AI-supported phishing awareness workflow was associated with improvements in detection-related behavioral responses across the training sequence. RQ2 examined how clicking and reporting patterns reflect participants' learning and response behavior across phases and difficulty conditions. RQ3 examined whether post-training engagement and perceived effectiveness scores reflected positive participant perceptions of the workflow. The analyses reported below address these questions through both behavioral and perception-based findings.

Before hypothesis testing, the behavioral dataset was prepared and screened in SPSS to verify coding accuracy, data quality, and sample adequacy. The dataset was structured at the email-exposure level, with each row representing one participant's exposure to a single email condition. Variables were coded to match the study design, including binary behavioral outcomes coded as 0 and 1, phase and difficulty condition variables, and a timing variable retained as a scale measure. Variable labels and value labels were also applied to support consistent interpretation during analysis.

Initial screening of the exposure level dataset used frequency distributions and descriptive statistics. The dataset contained 1,378 exposure records and showed no missing values in the primary behavioral or quality control variables examined. Delivery quality checks indicated that email transmission was highly reliable, with 99.9 percent of sent events recorded as successful and only 0.1 percent flagged as delivery errors. This supported the interpretation that low or absent interaction was primarily behavioral rather than the result of technical delivery failure. Descriptive analysis of TimeToClickSeconds identified 434 valid timing records, matching the total number of click events, and showed substantial variability consistent with a non-normal distribution. As a result, timing-related analyses were reserved for later assumption testing and nonparametric procedures where appropriate.

Because the behavioral dataset was organized at the exposure level, a participant-level summary dataset was created in SPSS using the Aggregate procedure to assess sample adequacy and participation completeness. This screening identified 188 unique participants, exceeding the minimum target of 100. Most participants completed the full sequence of eight email exposures, with 151 participants, or 80.3 percent, completing all eight exposures and 21 participants, or 11.2 percent, completing seven. Interaction-based participation was also strong, with 168 participants, or 89.4 percent, showing at least one measurable interaction and 20 participants, or 10.6 percent, showing none. After the inclusion criterion was applied to the participant-level summary dataset, 116 participants, or 61.7 percent, met the threshold for the primary analysis, while 72 participants, or 38.3 percent, were excluded because of lower engagement. The primary inferential tests reported below were conducted on the interaction filtered exposure dataset, as noted.
Because the key outcome variables were categorical, chi-square tests of independence were used to evaluate associations among study phase, email difficulty, and correct decision (McHugh, 2013). Although CorrectDecision was coded as 0 and 1, it represents a binary categorical outcome rather than a continuous measure, making chi-square more appropriate than ANOVA for the primary inferential analyses.

Correct decision improvement across phases

To examine whether decision quality improved across the training sequence, a chi-square test of independence was conducted between study phase and correct decision, using only exposures with

measurable interaction (InteractionOccurred = 1). The association between phase and correct decision was statistically significant, $\chi^2(1, N = 897) = 8.674$, $p = .003$. Cramér's V = .098, indicating a small effect size. The proportion of correct decisions increased from 51.3 percent in Phase 1 (244 of 476 interacted exposures) to 61.0 percent in Phase 2 (257 of 421 interacted exposures). A bar chart was generated to visualize this increase (see Figure 1).

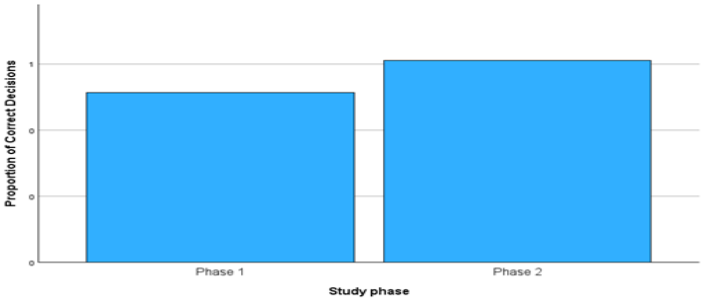

Figure 1. Correct decision rates increased from Phase 1 to Phase 2 among interacted exposures

A supplementary participant-level analysis was also conducted to examine whether correct-decision counts improved from Phase 1 to Phase 2 among participants who met the interaction-based inclusion threshold. Because the paired correct count variables were not assumed to be normally distributed, a Wilcoxon signed-rank test was used. The analysis showed a statistically significant increase in 'correct decision' counts from Phase 1 to Phase 2 ($Z = -2.700$, $p = .007$), based on 116 paired cases. Rank results indicated that 58 participants showed improvement, 34 showed lower performance in Phase 2, and 24 showed no change, see Table 1. This supplementary result strengthens the main exposure level findings by indicating that improvement was also observed at the participant level among those who engaged sufficiently with the training workflow.

Table 1. Participant Level Change in Correct Decisions from Phase 1 to Phase 2.

Outcome category	n	%
Improved in Phase 2	58	50.0
No change	24	20.7
Declined in Phase 2	34	29.3
Total	116	100.0

Note: n = number of participants. The improvement in Phase 2 indicates a higher number of correct decisions than in Phase 1.

To further clarify whether improvement reflected fewer repeated errors among participants who initially struggled, an additional supplementary participant-level learning analysis was conducted to examine whether eligible participants who made at least one incorrect decision in Phase 1 demonstrated fewer incorrect decisions in Phase 2. The analysis was restricted to participants who met the interaction-based inclusion threshold (Total Interactions at least 4) and had Incorrect_P1 greater than 0, yielding 100 paired cases. Because the paired incorrect count variables were not assumed to be normally distributed, a Wilcoxon signed-rank test was used. Results indicated a statistically significant reduction in the number of incorrect decisions from Phase 1 to Phase 2, $Z = -3.552$, $p < .001$. Rank results indicated that 49 participants showed improvement, 23 showed higher incorrect counts in Phase 2, and 28 showed no change. This supplementary finding supports the phase-based improvement pattern by indicating fewer repeated errors among participants who initially made mistakes during Phase 1 (Table 2).

Table 2. Participant-Level Change in Incorrect Decision Counts from Phase 1 to Phase 2

Outcome category	n	%
Improved in Phase 2 (Incorrect_P2 < Incorrect_P1)	49	49.0
No change (Incorrect_P2 = Incorrect_P1)	28	28.0
Declined in Phase 2 (Incorrect_P2 > Incorrect_P1)	23	23.0
Total	100	100.0

Note: n = number of participants. The improvement in Phase 2 indicates fewer incorrect decisions in Phase 2 than in Phase 1.

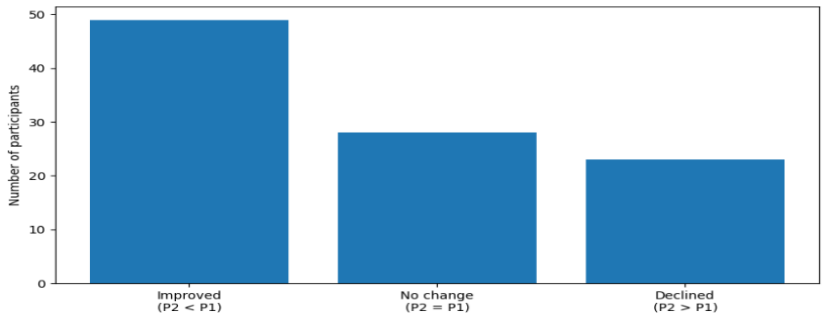

Figure 2. Change in Incorrect Decision Counts from Phase 1 to Phase 2. (Wilcoxon signed-rank sample, N=100)

Together, these findings provide the primary behavioral evidence addressing RQ1, indicating that the training workflow was associated with improvements in detection-related responses across phases.

Correct decision patterns by difficulty and phase

A second chi-square test examined the association between email difficulty and correct decision in the same interaction filtered dataset and was statistically significant, $\chi^2(3, N = 897) = 8.474$, $p = .037$, Cramér's $V = .097$. Correct decision rates were 49.1 percent for legitimate control emails, 53.1 percent for low-difficulty phishing emails, 60.9 percent for medium-difficulty phishing emails, and 61.1 percent for high-difficulty phishing emails (see Table 3).

Table 3. Overall correct decision by difficulty levels across interacted exposures.

Email Difficulty	Incorrect n (%)	Correct n (%)	Total n
Legitimate	88 (50.9)	85 (49.1)	173
Low	152 (46.9)	172 (53.1)	324
Medium	88 (39.1)	137 (60.9)	225
High	68 (38.9)	107 (61.1)	175
Total	**396 (44.1)**	**501 (55.9)**	**897**

Note: Values are frequencies with percentages in parentheses. The legitimate condition refers to the control email rather than a phishing difficulty level.

To evaluate whether difficulty effects differ across phases, phase-specific cross-tabulations were conducted. In Phase 1, correct decision rates differed significantly across difficulty categories, $\chi^2(3, N = 476) = 23.231$, $p < .001$, with Cramér's $V = .221$, indicating a modest effect size, ranging from 32.1 percent for legitimate control emails to 66.1 percent for medium difficulty emails. In Phase 2, the association

between email difficulty and correct decision was not statistically significant, $\chi^2(3, N = 421) = 2.126$, $p = .547$, with Cramér's $V = .071$, and correct decision rates were more consistent across categories, ranging from 55.1 percent to 63.5 percent (Table 4).

Table 4. Correct decision rates by email difficulty across Phase 1 and Phase 2 among interacted exposures.

Email Difficulty	Phase 1 Incorrect n (%)	Phase 1 Correct n (%)	Phase 1 Total n	Phase 2 Incorrect n (%)	Phase 2 Correct n (%)	Phase 2 Total n
Legitimate	53 (67.9)	25 (32.1)	78	35 (36.8)	60 (63.2)	95
Low	113 (51.4)	107 (48.6)	220	39 (37.5)	65 (62.5)	104
Medium	40 (33.9)	78 (66.1)	118	48 (44.9)	59 (55.1)	107
High	26 (43.3)	34 (56.7)	60	42 (36.5)	73 (63.5)	115
Total	**232 (48.7)**	**244 (51.3)**	**476**	**164 (39.0)**	**257 (61.0)**	**421**

Note: Values are frequencies with percentages in parentheses. Phase 1 represents the guided learning phase. Phase 2 represents the independent assessment phase.

Clicking and reporting patterns by difficulty and phase

To examine behavioral outcomes beyond correct decision, phase-stratified analyses evaluated clicking and reporting by difficulty within the interaction-filtered exposure dataset. For clicking, the association between email difficulty and clicking was statistically significant in Phase 1, $\chi^2(3, N = 476) = 13.153$, $p = .004$, and Phase 2, $\chi^2(3, N = 421) = 10.893$, $p = .012$. To directly assess phase change across difficulty levels, Phase 1 versus Phase 2 comparisons were conducted for clicking. Clicking increased significantly for the legitimate control condition, $\chi^2(1, N = 173) = 15.240$, $p < .001$, from 33.3 percent to 63.2 percent, while clicking decreased significantly for the low difficulty condition, $\chi^2(1, N = 324) = 4.231$, $p = .040$, from 54.5 percent to 42.3 percent. Phase differences in clicking were not statistically significant for medium difficulty, $\chi^2(1, N = 225) = 1.452$, $p = .228$, or high difficulty, $\chi^2(1, N = 175) = 1.538$, $p = .215$ (see Table 5).

Reporting behavior also differed significantly by difficulty within each phase, with Phase 1 showing $\chi^2(3, N = 476) = 55.816$, $p < .001$, and Phase 2 showing $\chi^2(3, N = 421) = 14.263$, $p = .003$. Phase 1 versus Phase 2 comparisons by difficulty indicated significant reporting increases for the legitimate control condition, $\chi^2(1, N = 173) = 5.748$, $p = .017$, from 5.1 percent to 16.8 percent, and for the low difficulty condition, $\chi^2(1, N = 324) = 4.255$, $p = .039$, from 24.5 percent to 35.6 percent. Reporting decreased significantly for the medium difficulty condition, $\chi^2(1, N = 225) = 4.346$, $p = .037$, from 48.3 percent to 34.6 percent, while the high difficulty condition showed no statistically significant phase difference, $\chi^2(1, N = 175) = 1.605$, $p = .205$. Given the workflow in which participants could report after viewing landing page guidance, reporting was treated as a protective and potentially corrective behavior indicator, and shifts across phases were interpreted as changes in response patterns rather than a uniform increase across all difficulty levels.

Table 5. Clicking and reporting rates by email difficulty across Phase 1 and Phase 2.

Email difficulty	Phase 1 clicked n (%)	Phase 2 clicked n (%)	Chi-square for clicking	p	Phase 1 reported n (%)	Phase 2 reported n (%)	Chi-square for reporting	p
Legitimate	26 (33.3)	60 (63.2)	15.240	< .001	4 (5.1)	16 (16.8)	5.748	.017
Low	120 (54.5)	44 (42.3)	4.231	.040	54 (24.5)	37 (35.6)	4.255	.039
Medium	49 (41.5)	53 (49.5)	1.452	.228	57 (48.3)	37 (34.6)	4.346	.037
High	32 (53.3)	50 (43.5)	1.538	.215	30 (50.0)	46 (40.0)	1.605	.205

Note. Values are frequencies with percentages in parentheses. Clicking on the legitimate control email may reflect appropriate engagement, whereas clicking on phishing emails reflects risky behavior. Reporting was interpreted as a protective and potentially corrective response.

These clicking and reporting patterns address RQ2 by showing how participant response behavior varied across the training sequence and difficulty conditions.

Phase 1 learning process indicator

Phase 1 process analyses were used to describe learning phase-behavior during the training workflow. In Phase 1, clicking varied by difficulty, with click rates of 33.3 percent for legitimate control (26 of 78), 54.5 percent for low difficulty (120 of 220), 41.5 percent for medium difficulty (49 of 118), and 53.3 percent for 'high difficulty' (32 of 60). Reporting also varied by difficulty, with rates of 5.1 percent for legitimate control (4 of 78), 24.5 percent for low difficulty (54 of 220), 48.3 percent for medium difficulty (57 of 118), and 50.0 percent for 'high difficulty' (30 of 60). Corrective behavior evidence was treated as a Phase 1 learning indicator because Phase 1 included training guidance and a second opinion opportunity designed to prompt reflection after initial engagement. Within Phase 1, evidence of corrective behavior occurred in 34.0 percent of interacted exposures (162 of 476).

Likert scale survey

Although 114 respondents submitted the survey, the perceived usefulness and effectiveness composite was computed only for respondents with complete, usable responses to the three component items: Usefulness, Confidence, and OrgAdoption. As a result, 97 cases were retained as valid for the composite and reliability analyses, and the remaining cases were excluded because they had missing responses on one or more of the required items, ensuring consistent item information for the reported scale results.

During data preparation, composite perception scores were computed from the 5-point Likert items to represent the main constructs. For each respondent, an arithmetic mean was computed for each construct using $\bar{X} = (\Sigma X_i) / k$, where X_i is the item score and k is the number of items. Engagement was calculated as the mean of 'FocusedAttention', 'TimeLoss', Involvement, and Reward; personalization or adaptivity was calculated as the 'mean' of Personalized and Adaptive; and perceived usefulness and effectiveness were calculated as the 'mean' of Usefulness, Confidence, and OrgAdoption. Because the perception items were adapted and combined from validated instruments, internal consistency was examined using Cronbach's alpha to provide evidence of reliability and construct alignment.

Perceived usefulness and effectiveness were the primary perception outcomes and were operationalized as a 3-item composite derived from Usefulness, Confidence, and 'OrgAdoption'. The usefulness and effectiveness scale demonstrated strong internal consistency in the final sample, with an alpha = .814 (3 items, valid cases = 97). Descriptive results indicated generally positive perceived usefulness and effectiveness, with a mean of 3.68 (SD = 0.90, N = 97) on a 1-5 scale. Consistent with methodological guidance cautioning against rigid reliance on fixed alpha cutoffs, the alpha value was interpreted in context rather than treated as a pass-or-fail threshold (Hussey et al., 2025).

Discussion

This study examined whether an open-source, AI-supported phishing awareness workflow implemented through GoPhish was associated with behavioral improvement and favorable participant perceptions. Behavioral outcomes were assessed using exposure-level campaign logs and reporting records, with emphasis on correct decision, clicking, and reporting across two phases and multiple difficulty levels. A post simulation survey assessed perceived usefulness and effectiveness. Overall, the findings suggest that an open-source simulation-based approach can support meaningful phishing awareness outcomes in resource-constrained settings.

The clearest evidence of effectiveness was the improvement in correct decision performance from Phase 1 to Phase 2 among exposures with measurable interaction. Because Phase 2 removed the corrective prompts available during the learning phase, this pattern suggests carryover learning rather than immediate correction alone. Supplementary participant-level Wilcoxon analyses supported this interpretation by

showing both an increase in correct decisions and a reduction in incorrect decisions from Phase 1 to Phase 2 among eligible participants. Although the effects were small, they indicate measurable behavioral improvement within the study period. The association between difficulty and correct decision was significant in Phase 1 but not in Phase 2, suggesting that repeated exposure and guided practice may have reduced performance variability across message types and produced more stable decision behavior in the later phase.

Clicking and reporting patterns offered additional insight into participant responses. Clicking declined across phases for the low difficulty phishing condition, which is consistent with reduced risky engagement, while clicking increased for the legitimate control condition, which may reflect more appropriate engagement with safe content. Reporting showed a mixed pattern, increasing for legitimate and low difficulty conditions, decreasing for the medium difficulty condition, and showing no significant change for the high difficulty condition. This suggests that reporting may be more sensitive to message difficulty, confidence, and the structure of the learning sequence than a simple linear improvement pattern would indicate.

The Phase 1 findings also support the instructional role of the learning phase. Corrective behavior evidence suggests that many participants adjusted their responses after viewing guidance content and using the second opinion opportunity. Because Phase 2 functioned as an assessment phase without corrective prompts, independent performance was interpreted primarily through correct decision and the related patterns of clicking and reporting. Survey findings complemented the behavioral results by indicating generally positive perceived usefulness and effectiveness, supported by strong internal consistency of the primary perception scale. Although favorable perceptions do not guarantee behavioral change, their convergence with the behavioral findings strengthens the practical significance of the workflow examined in this study.

A broader aim of the study was to examine whether an open-source, AI-supported workflow could produce outcomes relevant to limitations often associated with traditional static awareness methods and comparable in form to behavior-based indicators emphasized in the commercial training literature. Commercial platforms commonly highlight features such as AI-assisted content generation, automated delivery, one-click reporting, post-click learning, adaptive pathways, and progress tracking (Guardz, 2025; Phished.io, 2024; Phish Firewall, 2024). Although this study did not include a direct comparison with traditional or commercial platforms, the observed improvement in correct-decision performance and the greater consistency of Phase 2 performance across difficulty levels support the practical value of an open-source simulation approach for organizations that may not have access to proprietary platforms.

Conclusion

This study evaluated the behavioral impact of an open-source, AI-supported phishing awareness training approach using GoPhish and LLM-assisted phishing scenario generation. Correct decision rates increased from Phase 1 to Phase 2 among interacted exposures, and supplementary participant-level analyses showed both more correct decisions and fewer repeated incorrect decisions in Phase 2 among eligible participants. Although the effect sizes were modest, the findings indicate measurable behavioral improvement within a controlled simulation design involving an ecologically diverse sample.

Clicking and reporting patterns also suggested meaningful engagement with the simulation, especially during the guided first phase, where responses reflected learning, reflection, and corrective action. Survey findings further indicated generally positive perceptions of usefulness and effectiveness, supported by acceptable internal consistency of the primary perception scale. Together, these results suggest that open source, AI-supported phishing simulation environments can produce meaningful awareness outcomes in resource-constrained settings.

These findings should still be interpreted with caution. The primary behavioral analyses focused on exposures with measurable interaction, which strengthens confidence that the results reflect actual engagement but excludes silent nonresponse and may overrepresent more attentive participants. In addition, the main chi-square analyses were conducted at the exposure level, so full independence among observations may not have been fully met. The purposive sample also came from varied educational and organizational settings, and baseline cybersecurity knowledge, prior training, and familiarity with simulated exercises were not controlled. The short study period, limited spacing between exposures, manual logging of reporting behavior, and usable survey responses from only part of the sample also limit interpretation.

The AI component was also limited in scope. It was used only for LLM-assisted generation of phishing scenarios and did not include adaptive learning algorithms, behavioral profiling, or real-time personalization. The findings, therefore, reflect an LLM-assisted training workflow rather than an autonomous AI-enabled platform. Even so, the study provides applied evidence that an open-source, AI-supported workflow can support behavioral change and offers a practical foundation for further development and evaluation of accessible phishing awareness programs. Future research should examine longer interventions, delayed follow-up measures, more automated behavioral telemetry, and controlled comparative designs that include either traditional awareness training or commercial platforms.

References

Abdykerimov, A., Isaev, R., & Amanov, R. (2025). *An empirical analysis of phishing simulation to mitigate social engineering attacks*. https://doi.org/10.20944/preprints202511.1827.v1

Afolalu, O., & Tsoeu, M. S. (2025). Cybersecurity in higher education institutions: A systematic review of emerging trends, challenges and solutions. *Future Internet*, *17*(12), 575. https://doi.org/10.3390/fi17120575

Alsariera, Y. A., Adeyemo, V. E., Balogun, A. O., & Alazzawi, A. K. (2020). AI meta-learners and extra-trees algorithm for the detection of phishing websites. *IEEE Access*, *8*, 142532–142542. https://doi.org/10.1109/ACCESS.2020.3013699

Ansari, M. F., Sharma, P. K., & Dash, B. (2022). Prevention of phishing attacks using AI-based cybersecurity awareness training. *International Journal of Smart Sensor and Adhoc Network.*, 61–72. https://doi.org/10.47893/IJSSAN.2022.1221

Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., & Walker, K. (2020). Purposive sampling: Complex or simple? Research case examples. *Journal of Research in Nursing*, *25*(8), 652–661. https://doi.org/10.1177/1744987120927206

Chen, Y.-H., & Chen, J.-L. (2021). Intelligent learning architecture with hybrid features for phishing detection. *2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN)*, 225–230. https://doi.org/10.1109/ICUFN49451.2021.9528537

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (Sixth edition). Sage.

CybeReady. (2024). *The Role of AI in Phishing Detection and Employee Training*. Retrieved September 2, 2024, from https://cybeready.com

De Rosa, S., Gringoli, F., & Bellicini, G. (2024). "Hey ChatGPT, Is This Message Phishing?" *2024 22nd Mediterranean Communication and Computer Networking Conference (MedComNet)*, 1–10. https://doi.org/10.1109/MedComNet62012.2024.10578236

Desolda, G., Ferro, L. S., Marrella, A., Catarci, T., & Costabile, M. F. (2022). Human factors in phishing attacks: A systematic literature review. *ACM Computing Surveys*, *54*(8), 1–35. https://doi.org/10.1145/3469886

Dewis, M., & Viana, T. (2022). Phish responder: A hybrid machine learning approach to detect phishing and spam emails. *Applied System Innovation*, *5*(4), 73. https://doi.org/10.3390/asi5040073

Eze, C. S., & Shamir, L. (2024). Analysis and prevention of AI-based phishing email attacks. *Electronics*, *13*(10), 1839. https://doi.org/10.3390/electronics13101839

Goel, S., Williams, K., & Dincelli, E. (2017). Got phished? Internet security and human vulnerability. *Journal of the Association for Information Systems*, *18*(1), 2. https://doi.org/10.17705/1jais.00447

GoPhish. (n.d.). *API documentation. GoPhish Documentation.* Retrieved September 23, 2025, from https://docs.getgophish.com/api-documentation/

Greco, F., Desolda, G., Tucci, C., Esposito, A., Curci, A., & Piccinno, A. (2025). *Improving phishing resilience with AI-generated training: Evidence on prompting, personalization, and duration*. arXiv. https://doi.org/10.48550/arXiv.2512.01893

Grimes, S., & Kraemer, M. (2023). KnowBe4 data confirms the value of security awareness training (SAT). KnowBe4. Retrieved September 10, 2025, from https://www.knowbe4.com/hubfs/Data-Confirms-Value-of-SAT-WP_EN-us.pdf

Guardz. (2025). *AI-powered phishing simulation*. Retrieved September 3, 2025, from https://guardz.com/solution/phishing-simulation-v1/

Ho, G., Mirian, A., Luo, E., Tong, K., Lee, E., Liu, L., Longhurst, C. A., Dameff, C., Savage, S., & Voelker, G. M. (2025). Understanding the efficacy of phishing training in practice. *2025 IEEE Symposium on Security and Privacy (SP)*, 37–54. https://doi.org/10.1109/SP61157.2025.00076

Horvat, D. M. D., & Radić, D. T. P. (2024). Assessing end-user resilience to phishing: A study on educational interventions and simulated attacks in a Croatian university. *European Journal of Emerging Cybersecurity and Information Protection*, *1*(01), 44–56. https://parthenonfrontiers.com/index.php/ejecip/article/view/83

Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. (2025). An aberrant abundance of Cronbach's alpha values at. 70. *Advances in Methods and Practices in Psychological Science*, *8*(1), 25152459241287123. https://doi.org/10.1177/25152459241287123

Jain, B. (2025, July 2). *How AI-based phishing simulations reduced our attack surface*. ISC2. Retrieved October 18, 2025, from https://www.isc2.org/insights/2025/07/how-ai-based-phishing-simulations-reduced-our-attack-surface

Jayakrishnan, G., Banahatti, V., & Lodha, S. (2022). Pickmail: A serious game for email phishing awareness training. *Proceedings 2022 Symposium on Usable Security*. Symposium on Usable Security. https://dx.doi.org/10.14722/usec.2022.23059

Khan, K. (2024). *Comparison of anti phishing tools*. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1856752&dswid=-1894

King Phisher. (2024). *King Phisher: Phishing Campaign Toolkit*. GitHub. Retrieved September 1, 2024, from https://github.com/securestate/king-phisher

Lim, B., Huerta, R., Sotelo, A., Quintela, A., & Kumar, P. (2025). Explicate: Enhancing phishing detection through explainable AI and LLM-powered interpretability. arXiv. https://doi.org/10.48550/arXiv.2503.20796

Luse, A., & Burkman, J. (2021). Gophish: Implementing a real-world phishing exercise to teach social engineering. Journal of Cybersecurity Education, Research and Practice, 2020(2). https://doi.org/10.62915/2472-2707.1072

Marikyan, D., & Papagiannidis, S. (2023). Protection motivation theory: A review. *TheoryHub Book: This handbook is based on the online theory resource: TheoryHub*, 78-93. https://research-information.bris.ac.uk/ws/portalfiles/portal/376433746/Marikyan_and_Papagiannidis_2023_protection_motivation_theory_Review.pdf

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/BM.2013.018

Mouwers-Singh, C., & Musikavanhu, T. B. (2024). A narrative review on enhancing cybersecurity in higher education institutions: the role of continuous training and awareness. *Expert Journal of Business and Management*, *12*(2). https://business.expertjournals.com/ark:/16759/EJBM_1206mouwers67-73.pdf

O'Brien, H. L., & Toms, E. G. (2010). The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69. https://doi.org/10.1002/asi.21229

Phish Firewall. (2024). *AI-driven Phishing Awareness Training*. Retrieved September 1, 2024, from https://phishfirewall.com

Phished.io. (2024). *Automated Phishing Simulations*. Retrieved September 1, 2024, from https://phished.io

Sameen, M., Han, K., & Hwang, S. O. (2020). Phishhaven—An efficient real-time AI phishing urls detection system. *IEEE Access*, *8*, 83425–83443. https://doi.org/10.1109/ACCESS.2020.2991403

Sumner, A., Yuan, X., Anwar, M., & McBride, M. (2022). Examining factors impacting the effectiveness of anti-phishing trainings. *Journal of Computer Information Systems*, *62*(5), 975–997. https://doi.org/10.1080/08874417.2021.1955638

Taofeek, A. O. (2024). Development of a novel approach to phishing detection using machine learning. *ATBU Journal of Science, Technology and Education*, *12*(2), 336-351. https://www.researchgate.net/publication/381717458_Development_of_a_Novel_Approach_to_Phishing_Detection_using_Machine_Learning

Tam, K. Y., & Ho, S. Y. (2006). Understanding the impact of web personalization on user information processing and decision outcomes1. *MIS Quarterly*, *30*(4), 865–890. https://doi.org/10.2307/25148757

Xue, Y., Spero, E., Koh, Y. S., & Russello, G. (2025). *Multiphishguard: An LLM-based multi-agent system for phishing email detection*. arXiv. https://doi.org/10.48550/arXiv.2505.23803

Young, J., & Farshadkhah, S. (2023). Teaching tip: Hook, line, and sinker – the development of a phishing exercise to enhance cybersecurity awareness. *Journal of Information Systems Education*, *34*(4), 347–359. https://aisel.aisnet.org/jise/vol34/iss4/1