Email Phishing Prevention: An Explainable Nudging Approach

Mason Wagner[1]
Eller College of Management,
University of Arizona,
Tucson, Arizona, USA

Benjamin M. Ampel
Robinson College of Business,
Georgia State University,
Atlanta, Georgia, USA

Matthew Hashim
Eller College of Management,
University of Arizona,
Tucson, Arizona, USA

Hsinchun Chen
Eller College of Management,
University of Arizona,
Tucson, Arizona, USA

ABSTRACT

Email is a ubiquitous communication tool. This ubiquity has made emails a frequent target of phishing attacks. Organizations often use digital nudges (interventions designed to guide user behavior) to warn users of a phishing email. However, these warnings are often uninterpretable and thus often ignored. Therefore, this study investigates how explainable artificial intelligence (XAI) can be implemented in a digital nudge to enhance user phishing detection and prevention. In this research, we designed an XAI-enabled nudge and conducted a between subjects' experiment comparing it against a popular extant nudge and a control group of no nudge. The results suggested that XAI-enabled nudges significantly improve user phishing detection accuracy by 10.1%. Additionally, XAI-enabled nudges significantly decreased intrinsic and germane cognitive load. These findings contribute to XAI research by demonstrating the value of AI explanations in security contexts and its impact on cognitive processing. These results also have practical implications for email security system integration design.

Keywords: phishing, nudges, user experiment, explainable artificial intelligence, security research

[1] Corresponding author. rmw26@arizona.edu

INTRODUCTION

Email communication is used by 4 billion people to send over 300 billion emails daily (Ceci 2024). However, up to 46% of emails are phishing attempts (Griffiths 2024). Phishing is a method employed by threat actors designed to deceive the recipient into giving up personal, financial, or other information of value to a threat actor (Commission 2022). On average, organizations face a $4.88 million loss per successful phishing attempt (IBM 2024). To combat phishing, organizations are increasingly deploying digital nudges (e.g., warning banners or alerts) to warn users of potential phishing threats. Figure 1 shows an example phishing email impersonating the company Facebook with an organizational digital nudge ("[EXT]" in the subject line, as shown in Figure 1a).

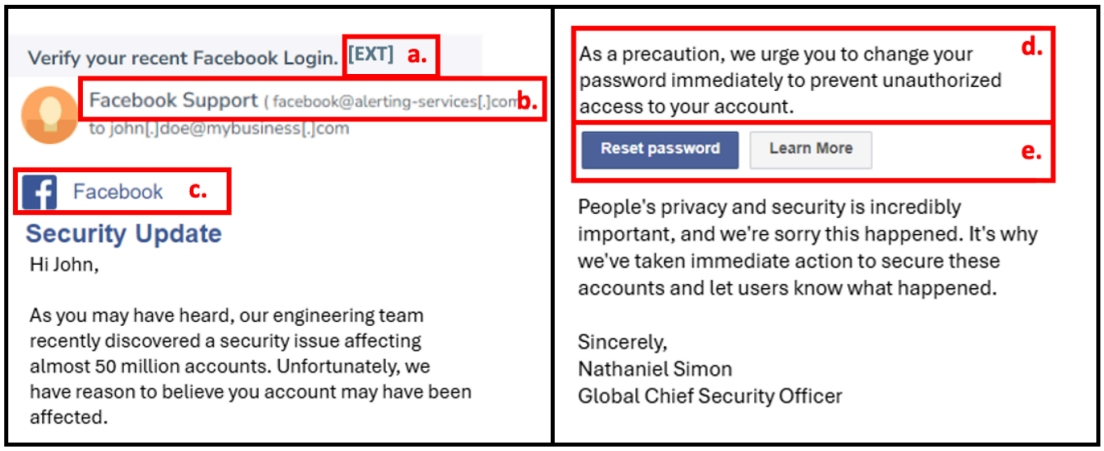

Figure 1. Example of a Phishing Email

Phishing emails often contain four key attributes (Lenaerts-Bergmans 2021). First, malicious actors often spoof realistic looking email addresses (Figure 1b). Second, malicious actors use genuine logos and branding (Figure 1c). Third, phishing emails often contain a persuasion cue (e.g., calls to action, Figure 1d) (Valecha et al. 2021). Finally, phishing emails often disguise a malicious link (Figure 1e) that lead to a credential stealing webpage.

To detect phishing emails, organizations are increasingly implementing Artificial Intelligence (AI)-enabled solutions (e.g., machine learning classifiers) (Ampel et al. 2023). AI-

enabled classifiers often provide their inferences to users in the form of digital nudges (e.g. EXT in Figure 1a.) (Cisa et al. 2023). However, despite training efforts and digital nudges, well-crafted phishing emails can still frequently evade human detection (Ho et al. 2024).

Extant literature on phishing nudges has largely focused on evaluating the effectiveness of visual or timing-based interventions (Baki et al. 2024; Lain et al. 2024). However, these nudges do not provide transparent or model-driven explanations to inform user decision-making. Explainable Artificial Intelligence (XAI) presents a potential solution for improving nudges in email systems. XAI can output key factors that lead to a classification result, leading to increased trust and adoption by users (Alufaisan et al. 2021). However, extant work has not explored how XAI nudges can impact user phishing detection. As phishing tactics grow more sophisticated, it is vital that users can quickly comprehend and trust AI-enabled digital nudge warnings.

To address this gap, we introduce an XAI-enabled digital nudge designed to convey the key features that drive a phishing detection model's decision. To determine the efficacy of our XAI-enabled nudge, we conducted a between-subjects behavioral experiment with three conditions: (1) a control group (no nudge), (2) a standard nudge group ([EXT] warning), and (3) an XAI-enabled nudge group. Participants were tasked with evaluating ten emails each (five phishing and five benign) presented in a simulated email client. Results suggested that our XAI-enabled nudge is effective at improving user detection when compared to the other two conditions.

LITERATURE REVIEW

Phishing Email Nudges

To establish a foundation for our study, we reviewed recent and relevant research on phishing-email nudges. Phishing email nudge research has often focused on identifying which nudge effectively prompts users to detect phishing content (e.g., notification banners) and

assessing user behavior under varied nudge conditions (e.g., nudge timings). These studies often implemented behavioral experiments in which participants were presented with email samples and asked to decide whether each message is benign or malicious (Lain et al. 2021, 2024). These studies often found that nudges like notification banners could reduce click-through on phishing links (Cooper et al. 2020; Lain et al. 2021).

Phishing email nudges can be broadly classified into two constructs: (1) Emotional arousal nudges and (2) sensory salience nudges (Cooper et al. 2020; Edwards and Li 2020; House and Raja 2020; Zheng and Becker 2023). First, emotional arousal nudges rely on either fear arousal (i.e., negative consequences) or rewards (i.e., positive reinforcement) (Lain et al. 2024). Second, sensory salience nudges capture users' attention through visual or auditory cues. These nudges often include notification banners inserted at strategic locations (Cooper et al. 2020; Yu et al. 05 2023). Notification banners are the standard for email-based nudges due to their ease of implementation. However, visual warnings may trigger habituation, where users may gradually ignore them over time (Cooper et al. 2020).

However, none of the reviewed studies leveraged XAI-driven explanations to convey why an email is suspected as phishing. One paper highlighted suspicious email elements via simple text-highlighting but did not link them to underlying model confidence or feature importance (Baki et al. 2024). As phishing attacks become increasingly sophisticated, static banners and highlights risk becoming insufficient for informed decision-making. Given this gap, we review XAI literature to identify techniques best suited for integration into phishing nudges.

Explainable Artificial Intelligence (XAI) Techniques

XAI research aims to generate transparent results for AI model decisions (Lundberg & Lee, 2017). There are five main representative XAI approaches: (1) Local Interpretable Model-agnostic

Explanations (LIME), (2) SHapley Additive exPlanations (SHAP), (3) attention mechanisms, (4) saliency maps, and (5) anchors (Lundberg and Lee 2017; Natarajan and Nambiar 2024; Ribeiro et al. 2018; Tjoa and Guan 2023; Vaswani et al. 2017).

First, LIME explains individual model predictions by learning a simple, interpretable surrogate model around the specific instance of interest. For a given email, LIME perturbs the text and observes how these changes affect the phishing classifier's output (Natarajan and Nambiar 2024). Second, SHAP treats each feature as a player in a coalition game, computing Shapley values by averaging marginal contributions across all possible feature subsets. SHAP guarantees that the sum of all token attributions equals the difference between the model's predicted probability for that email and the model's baseline output (Lundberg and Lee 2017). Third, attention mechanisms quantify how much attention the model pays to each token when forming higher-level representations or making a classification (Vaswani et al. 2017). Fourth, saliency maps leverage gradients to compute how small changes in the input would affect the model's output. In image classification, saliency maps highlight pixels that most influence the predicted label (Tjoa and Guan 2023). Finally, anchors generate high-precision, human-readable if-then rules that anchor a model's prediction by identifying a minimal set of features whose presence guarantees (with high probability) that the prediction will not change (Ribeiro et al. 2018).

We believe that SHAP is the best option for our text-based task because it produces a clear, token-level heatmap that directly highlights which words or phrases drive the classification decision. We provide an example of a SHAP visualization in Figure 2.

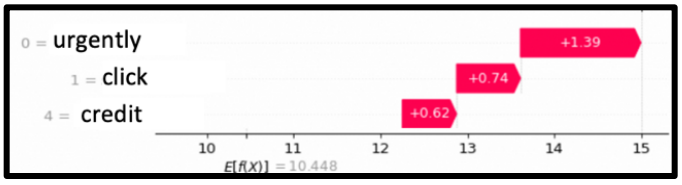

Figure 2. SHAP XAI Technique Visualization

SHAP explanations present a handful of high-impact tokens in a color-coded format. Rather than scanning the entire email content or parsing a verbose textual rule, the user's attention is immediately drawn to "urgently," "click," and "credit," which collectively account for most the model's decision. However, it is not clear if presenting such information will significantly impact user decision making. XAI has been shown to reduce user cognitive effort and facilitate decision-making in complex scenarios (Herm 2023). One framework relevant to this phenomenon is Cognitive Load Theory (CLT), which posits that the way information is presented significantly influences learning and comprehension (Sweller 1988). We review CLT in the following section.

Cognitive Load Theory (CLT)

Cognitive overload can occur when users are confronted with potentially malicious emails such as phishing attempts as users may experience emotions (e.g., stress, anxiety, excitement) that can impair decision-making (Jayatilaka et al. 2024). CLT offers a framework for understanding how information presentation affects users' working memory and overall performance. CLT distinguishes three types of cognitive load (Klepsch et al. 2017; Sweller et al. 1998):

1. **Intrinsic Load**: The mental effort required to understand or perform the task.
2. **Extraneous Load**: The cognitive effort imposed by suboptimal presentation or design.
3. **Germane Load**: The effort required to integrate information into long-term memory.

When confronted with a deceptive email, intrinsic load might be high because users must parse unfamiliar vocabulary or technical jargon. Extraneous load can also increase if the email interface forces users to scan irrelevant sections. Finally, limited germane resources may hinder the integration of new phishing-detection strategies into long-term memory (Klepsch et al. 2017).

RESEARCH GAPS AND HYPOTHESES

We note two key gaps from our literature review. First, despite extensive work on visual and incentive-based phishing nudges, no prior study has empirically evaluated how XAI-enabled

explanations function as digital nudges to improve users' ability to detect phishing emails. Notification banners and behavioral incentives have been shown to reduce click-through rates. However, the literature does not address whether model-level explanations (e.g., token-level attributions from SHAP) can further enhance detection accuracy or user trust without imposing excessive cognitive burden. Second, existing nudge research measures accuracy or click rates but does not assess how different explanation styles influence users' intrinsic, extraneous, and germane cognitive load when scanning for phishing cues. Prior literature suggests that making an automated system's reasoning transparent helps users more reliably distinguish legitimate from malicious digital content and can reduce cognitive burden during decision-making (Herm 2023; Longo et al. 2024). Therefore, we propose the following hypotheses:

- **H1**. *XAI digital nudges will lower phishing susceptibility rates.*
- **H2**. *XAI digital nudges will lower cognitive load during user phishing detection.*

Consistent with CLT, explanations that highlight only the most relevant features can (a) reduce unnecessary processing of irrelevant content (lowering extraneous load), (b) simplify the core decision task by focusing on a few high-impact tokens (lowering intrinsic load), and (c) foster schema formation through repeated exposure to salient phishing patterns (increasing germane load). To unpack these effects, we subdivide H2 as follows:

- **H2a**. *XAI digital nudges will lower intrinsic cognitive load in users.*
- **H2b**. *XAI digital nudges will lower extraneous cognitive load in users.*
- **H2c**. *XAI digital nudges will lower germane cognitive load in users.*

PROPOSED RESEARCH FRAMEWORK

To address the proposed research questions, we developed a framework with four components (Figure 3) to evaluate the effectiveness of XAI-based nudges: (1) Data Collection &

Pre-Processing, (2) Model Architecture, (3) Behavioral Experiment, and (4) Experiment Evaluation. Each component is detailed in the following sub-sections

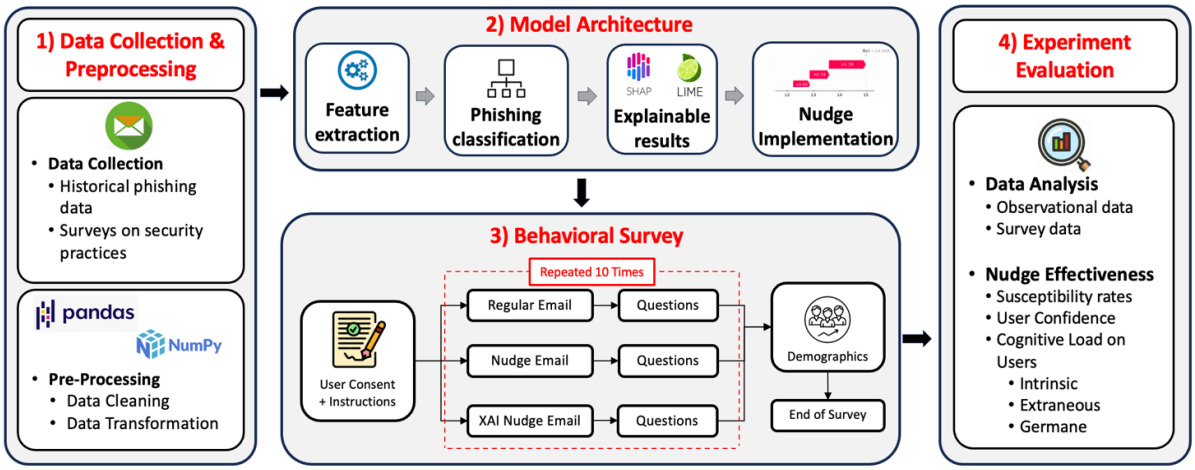

Figure 1. Proposed Research Framework

Data Collection & Pre-Processing

To build an XAI nudge, we required a trained classifier. Therefore, we collected a corpus of 43,000 phishing emails and 39,000 legitimate emails from recent phishing detection literature (Al-Subaiey et al. 2024). We implemented a pre-processing pipeline by removing duplicated records, removing URLs and email addresses (as they are not the focus of this study), stop words, extra spaces, and applying a lemmatizer to each token (Ampel et al. 2023).

Model Architecture

To detect phishing emails with high accuracy and generate interpretable explanations, we adopted a two-stage modeling pipeline. First, we fine-tuned a pre-trained Bidirectional Encoder Representations from Transformers (BERT) classifier on our labeled dataset. BERT is recommended for task classification tasks and was found to be generally robust to unseen data (Ampel et al. 2025; Otieno et al. 2023). Second, we applied SHAP to extract per-token attribution scores for every email instance, which serve as the basis for our XAI-enabled nudge.

Behavioral Experiment

To study how users evaluate emails, we presented them with benign and phishing messages. Each email was designed from a real phishing attempt. We employed a between-subjects experimental design to determine if one treatment (e.g., XAI nudge) is significantly different than another (e.g., no nudge) (*APA Dictionary of Psychology* 2018). Power analysis indicated that at least 51 observations were needed in each treatment group to achieve sufficient statistical power (Bhandari 2021). To ensure the reliability of participant responses, each subject was presented with 10 randomly selected emails of their assigned type (Kost and Correa da Rosa 2018; Revilla and Ochoa 2017). Participants were randomly assigned to one of three conditions (shown in Figure 4), corresponding to different types of phishing email presentations:

1. **Control**: Participants evaluated emails and did not receive a nudge.
2. **Standard Nudge**: Participants received phishing emails with a common notification banner [EXT] (Lain et al. 2024; Zheng and Becker 2023).
3. **XAI-enabled nudge**: Participants group received an XAI-enabled nudge that incorporated SHAP values extracted from our trained BERT classifier.

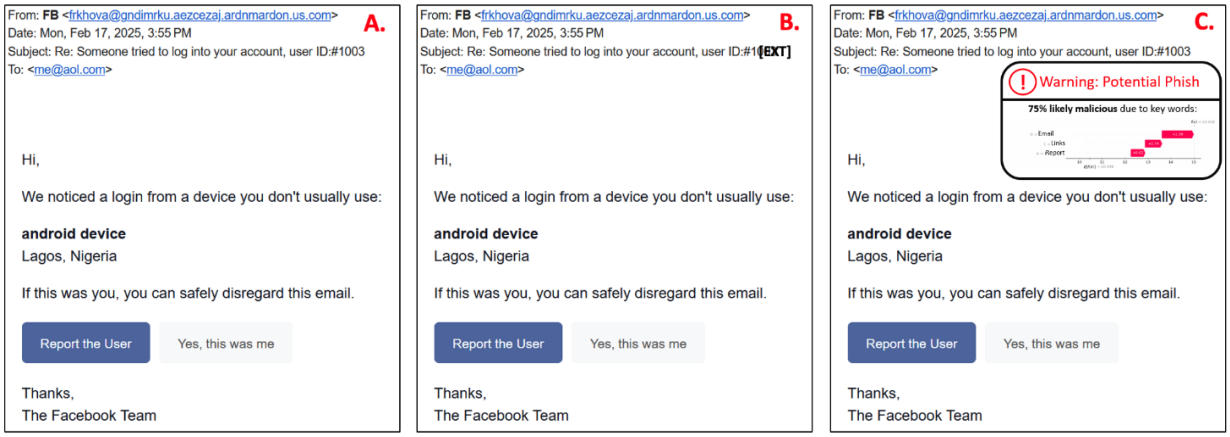

Figure 4. Example Phishing Emails with: (A) no nudge, (B) a standard nudge shown in extant systems, and (C) our proposed XAI-enabled nudge.

Experiment Evaluation

Our survey was deployed using Qualtrics and administered to undergrad business students at a public university. Business-student convenience samples often yield findings comparable to

those from broader consumer or professional panels (Compeau et al. 2012). Participants first reviewed an IRB-approved consent form explaining the study's purpose (evaluating email-judgment processes), confidentiality safeguards, and the voluntary nature of their participation. After consenting, participants read standardized instructions that outlined that they would see a series of email messages, their task was to decide whether each message was legitimate or phishing and rate their confidence and complete a short cognitive load questionnaire after each decision.

The survey platform randomly assigned each participant to a condition to ensure that observed differences in accuracy or cognitive load can be causally attributed to the type of nudge provided (Straub and Gefen 2004). Each participant then completed ten trials in their assigned condition. Participants were not told which emails are phishing, but they saw the same mix regardless of condition. After each trial, the users were asked for the judgement (benign or phishing attempt) and their confidence rating on a 7-point Likert scale (1 = Not at all confident, 7 = Extremely confident). Our survey also measured constructs relating to the three types of cognitive load (intrinsic, extraneous, and germane). We detail each construct in Table 2.

Table 2. Survey Constructs and Associated Questions for our Between-Subject Experiment

Construct	Question
Susceptibility	Which category does this email best fit into?
Confidence	How confident are you in this decision?
Intrinsic	For this task, many things need to be kept in mind simultaneously.
Cognitive Load	This task was very complex.
Extraneous	During this task, it was exhausting to find the important information.
Cognitive Load	The design of this task was very inconvenient for learning.
	During this task, it was difficult to recognize and link crucial information.
Germane	I made an effort, not only to understand several details, but to understand the overall context.
Cognitive Load	My point while dealing with the task was to understand everything correct.
	The learning task consisted of elements supporting my comprehension of the task.

The constructs of susceptibility and confidence were included to determine the direct effects of our XAI nudge on phishing detection performance (Singh et al. 2019). The three types of cognitive load (intrinsic, extraneous, and germane) were measured a using a 7-point Likert scale

(1 = low cognitive load, 7 = high cognitive load) to investigate how each nudge impacted users' cognitive processing (Klepsch et al. 2017; Ouwehand et al. 2021).

After completing the ten email-evaluation trials, participants answered demographics questions such as age, gender, and highest educational attainment. (Podsakoff et al. 2003). The survey contained one instructional manipulation check to verify participants attentiveness. Once demographics were submitted, participants receive debriefing information explaining the study's goals and are thanked for their time.

RESULTS

Participant Data and Response Summary

Our study recruited 367 participants, of which 347 were included in the final analysis after removing 20 responses that were either incomplete or failed attention-checking questions. Table 3 provides brief demographic information about our participants.

Table 3. Participant Summary Statistics (N = 347)

Characteristic	Distribution
Age Range	***18–24***: 339 (97.7%); ***25–34***: 6 (1.7%); ***35–44***: 2 (0.6%)
Gender	***Female***: 157 (45.2%); ***Male***: 188 (54.2%); ***Non-binary/Third gender***: 1 (0.3%); ***Prefer not to say***: 1 (0.3%)
Ethnicity	***White***: 217 (62.5%); ***Multi-ethnic***: 43 (12.4%); ***Asian***: 32 (9.2%); ***Hispanic***: 42 (12.1%); ***African American***: 7 (2.0%); ***Native American***: 3 (0.9%); ***Other***: 3 (0.9%)
Education	***High School Diploma/Equivalent***: 285 (82.1%); ***Associate's***: 44 (12.7%); ***Bachelor's***: 18 (5.2%)

The distribution of participants across conditions was balanced, with 117 in the no nudge condition, 116 in the standard nudge condition, and 114 in the XAI-enabled nudge condition. Table 4 reports how participants in each nudge condition classified the two types of emails (legitimate vs. phishing) and the corresponding misclassification rates.

Table 4. Impacts of Different Nudges on Phishing Classification

Nudge Type	Email Type	Benign	Malicious	Miss %
No Nudge	Legitimate	413	134	24.5%
	Phishing	322	301	51.7%
Standard Nudge	Legitimate	396	157	28.4%
	Phishing	305	302	50.2%
XAI-enabled Nudge	Legitimate	411	123	**23.0%**
	Phishing	206	400	**34.0%**

The XAI-Enabled Nudge condition reduced misclassification of phishing content from 51.7 % to 34.0 % when compared to the no nudge groups and reduction from 50.2 % to 34.0 % when compared to the standard nudge (a 16.2% improvement). By incorporating SHAP explanations, the XAI condition also kept the false-positive rate low at 23.0%, compared to 24.5% under no nudge and 28.4 % under standard nudge. Taken together, these results suggest that token-level explanations via SHAP can reduce false positives and negatives relative to providing no warning or a generic banner.

Statistical Analysis of Phishing Susceptibility

To determine whether the observed difference in phishing detection was statistically significant, we conducted a linear regression analysis with phishing susceptibility as the dependent variable and use the no nudge condition as the baseline. We included dummy variables for the other treatments as shown by the following equation:

$$Susceptibility_{ij} = \beta_0 + \beta_1 StandardNudge_i + \beta_2 XAINudge_i + \varepsilon_{ij}$$

where i indexes participants and j indexes each email trial (10 trials per participant, for a total of 3,470 observations). We clustered robust standard errors (SE) at the participant level (shown in parentheses) to account for within-subject correlation across multiple email judgments. The results are presented in Table 5.

Table 5. Linear Regression for Correct Phishing Email Identification by Treatment

Variable	Model (1)
Standard Nudge	0.009 (0.022)
XAI Nudge	-0.101** (0.021)
Constant	0.390 (0.014)
Observations	3,470
Participants	347
R2	0.011
F Stat	16.22**

*Note: The dependent variable represents susceptibility to phishing. Negative coefficients indicate lower susceptibility. **p < 0.01; *p < 0.05*

The intercept (0.390, $SE = 0.014$) represents the average probability that a phishing email is misclassified in the no nudge condition (i.e., participants with no warning at all misclassified

phishing emails approximately 39.0% of the time). The coefficient on the standard nudge is 0.009 ($SE = 0.022$), which is not statistically different ($p = 0.68$). This suggests that showing a generic [EXT] banner did not significantly change phishing susceptibility relative to no nudge.

Finally, the coefficient on XAI nudge is −0.101 ($SE = 0.021$, significant at $p < 0.01$). A negative coefficient indicates that participants in the XAI condition were on average 10.1% less likely to misclassify a phishing email compared to no nudge (predicted misclassification rate of 28.9%). An F-statistic of 16.22 suggests that the two nudge-condition indicators significantly explain variation in phishing misclassification across the 3,470 email-trial observations. However, an R^2 of 0.011 indicates that only about 1.1% of the total variation in misclassification is explained by nudge type alone. While the XAI nudge has a statistically significant effect, most of the variability in whether a given phishing email is correctly identified depends on factors beyond our experimental conditions and would be useful to explore in future extensions.

Effects on Cognitive Load

We also examined how the different nudge conditions affected participants' cognitive load using an Ordinary Least Squares (OLS) analysis (Marin et al. 2023). OLS in our context can provide interpretable effect sizes for the difference between our proposed XAI-enabled nudge and no nudge. The means for each cognitive load dimension by treatment condition are presented in Table 6.

Table 6. Cognitive Load Panel OLS Parameters by Treatment

Group	Intrinsic	Extraneous	Germane
Standard Nudge	0.0700	0.1485**	0.0239
XAI-enabled Nudge	-0.1411*	0.0807	-0.0914*
R^2	0.0044	0.0019	0.0017
No of Observations	3,470	3,470	3,470

*Note: Baseline is no nudge; **p < 0.01; *p < 0.05*

Participants who received an XAI-enabled nudge saw a significant decrease in intrinsic cognitive load, meaning that users who experienced this type of nudge showed reduced effort in

processing and judging each email's legitimacy when SHAP-based token highlights and a probability estimate were available. In contrast, the standard nudge group displayed a non-statistically significant slight increase in intrinsic cognitive load when compared to the baseline. This indicated that a generic warning banner alone did not simplify the core task of distinguishing phishing from benign messages compared to no nudge.

Extraneous load captures the cognitive effort imposed by how well information is presented. The standard nudge reported the largest impact on extraneous load, indicating that the generic warning banner introduced extra processing overhead. Possible reasons include the nudge's wording [EXT] may have been too ambiguous or it may have created a shift in focus away from email content toward the banner itself. The XAI-enabled nudge saw a slight increase in extraneous cognitive load compared to no nudge but was not statistically significant. Germane load measures the mental resources allocated to constructing and refining mental schemas.

The XAI-enabled nudge significantly decreased germane load, which suggests that participants may have relied more heavily on the model's highlights rather than engaging in deeper schema construction themselves. While this lowered effort, it may also imply less investment in learning to detect phishing independently. Our analyses for the constructs for cognitive load indicate support for H2a and H2c, as XAI explanations have lowered intrinsic load and germane load compared to no nudge. H2b is not supported as the XAI condition was not statistically significant. These results imply that participants may have leaned on the model's tokens rather than investigating effort in schema formation.

CONCLUSION

Our study explored the effect of injecting token-level explanations into email interfaces for phishing detection. Using a between-subjects experiment, we found that participants presented

with an XAI-enabled nudge experienced fewer false positives. Cognitive-load analyses shows that XAI explanations lower intrinsic and germane effort, streamlining the user's decision task without increasing unnecessary mental burden. Our work advances cybersecurity practice by demonstrating how transparent AI can strengthen user vigilance and contributes to XAI theory by illustrating the impact of model-level explanations in security interfaces. We envision several potential extensions for future research. First, email systems can implement optimal explanation granularity, long-term schema formation, and generalizability across diverse user groups and phishing tactics. Second, XAI can be implemented to other prominent forms of phishing (e.g., vishing, websites) (Ampel et al. 2026; Gao et al. 2025). Finally, agentic AI innovations present an interesting opportunity for crafting more advanced and automated phishing explanations. Each direction can provide value for phishing research.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation under Grant No. DGE-2336109.

REFERENCES

Al-Subaiey, A., Al-Thani, M., Abdullah Alam, N., Antora, K. F., Khandakar, A., and Uz Zaman, S. M. A. 2024. "Novel Interpretable and Robust Web-Based AI Platform for Phishing Email Detection," *Computers & Electrical Engineering: An International Journal* (120:109625), Elsevier BV, p. 109625.

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., and Kantarcioglu, M. 2021. "Does Explainable Artificial Intelligence Improve Human Decision-Making?," *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence* (35:8), Association for the Advancement of Artificial Intelligence (AAAI), pp. 6618–6626.

Ampel, B. M., Gao, Y., Hu, J., Samtani, S., and Chen, H. 2023. "Benchmarking the Robustness of Phishing Email Detection Systems," in *Proceedings of the Americas' Conference on Information Systems (AMCIS)*, pp. 1–10.

Ampel, B. M., Samtani, S., and Chen, H. 2026. "Automatically Detecting Voice Phishing: A Large Audio Model Approach," *MIS Quarterly*, pp. 1–21.

Ampel, B. M., Yang, C.-H., Hu, J., and Chen, H. 2025. "Large Language Models for Conducting Advanced Text Analytics Information Systems Research," *ACM Transactions on Management Information Systems* (16:1), Association for Computing Machinery (ACM), pp. 1–26.

APA Dictionary of Psychology. 2018. dictionary.apa.org. (https://dictionary.apa.org/between-subjects-design).

Baki, S., Qachfar, F. Z., M.Verma, R., Kennedy, R., and N.Jones, D. 2024. "Real-Time, Evidence-Based Alerts for Protection from Phishing Attacks," *IEEE Transactions on Dependable and Secure Computing* (PP:99), Institute of Electrical and Electronics Engineers (IEEE), pp. 1–15.

Bhandari, P. 2021. *Statistical Power and Why It Matters | A Simple Introduction*, Scribbr. (https://www.scribbr.com/statistics/statistical-power/).

Ceci, L. 2024. *Number of E-Mail Users Worldwide 2023 | Statista*, Statista. (https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/).

Cisa, Nsa, Fbi, and Ms-Iasc. 2023. *Stopping the Attack Cycle at Phase One*. (https://www.cisa.gov/resources-tools/resources/phishing-guidance-stopping-attack-cycle-phase-one).

Commission, F. T. 2022. *How To Recognize and Avoid Phishing Scams*, Consumer Information. (https://consumer.ftc.gov/articles/how-recognize-and-avoid-phishing-scams).

Compeau, D., Marcolin, B., Kelley, H., and Higgins, C. 2012. "Research Commentary—Generalizability of Information Systems Research Using Student Subjects—A Reflection

on Our Practices and Recommendations for Future Research," *Information Systems Research : ISR* (23:4), Institute for Operations Research and the Management Sciences (INFORMS), pp. 1093–1109.

Cooper, M., Levy, Y., Wang, L., and Dringus, L. 2020. "Subject Matter Experts' Feedback on a Prototype Development of an Audio, Visual, and Haptic Phishing Email Alert System," *Online Journal of Applied Knowledge Management (OJAKM)* (8:2), pp. 107–121.

Edwards, S. H., and Li, Z. 2020. "A Proposal to Use Gamification Systematically to Nudge Students toward Productive Behaviors," in *Koli Calling '20: Proceedings of the 20th Koli Calling International Conference on Computing Education Research*, New York, NY, USA: ACM, November 19. (https://doi.org/10.1145/3428029.3428057).

Gao, Y., Ampel, B. M., and Samtani, S. 2025. "Examining the Robustness of Machine Learning-Based Phishing Website Detection: Action-Masked Reinforcement Learning for Automated Red Teaming," in *2025 IEEE Security and Privacy Workshops (SPW)*, , June 6, pp. 288–293.

Griffiths, C. 2024. *The Latest Phishing Statistics (Updated January 2023) | AAG IT Support*, aag-it.com. (https://aag-it.com/the-latest-phishing-statistics/).

Herm, L.-V. 2023. "Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study," *ArXiv (Cornell University)*, Cornell University. (https://doi.org/10.48550/arxiv.2304.08861).

Ho, G., Mirian, A., Luo, E., Tong, K., Lee, E., Liu, L., Longhurst, C. A., Dameff, C., Savage, S., and Voelker, G. M. 2024. "Understanding the Efficacy of Phishing Training in Practice," in *2025 IEEE Symposium on Security and Privacy (SP)*, IEEE Computer Society, pp. 76–76.

House, D., and Raja, M. K. 2020. "Phishing: Message Appraisal and the Exploration of Fear and Self-Confidence," *Behaviour & Information Technology* (39:11), Informa UK Limited, pp. 1204–1224.

IBM. 2024. *Cost of a Data Breach 2024*, IBM. (https://www.ibm.com/reports/data-breach).

Jayatilaka, A., Asanka, N., Arachchilage, G., and Babar, M. 2024. "Why People Still Fall for Phishing Emails: An Empirical Investigation into How Users Make Email Response Decisions," *Network and Distributed System Security (NDSS) Symposium*. (https://doi.org/10.14722/usec.2024.23xxx).

Klepsch, M., Schmitz, F., and Seufert, T. 2017. "Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load," *Frontiers in Psychology* (8), p. 1997.

Kost, R. G., and Correa da Rosa, J. 2018. "Impact of Survey Length and Compensation on Validity, Reliability, and Sample Characteristics for Ultrashort-, Short-, and Long-Research

Participant Perception Surveys," *Journal of Clinical and Translational Science* (2), pp. 31–37.

Lain, D., Jost, T., Matetic, S., Kostiainen, K., and Capkun, S. 2024. "Content, Nudges and Incentives: A Study on the Effectiveness and Perception of Embedded Phishing Training," *ArXiv [Cs.CR]*, , September 2. (https://dl.acm.org/doi/10.1145/3658644.3690348).

Lain, D., Kostiainen, K., and Capkun, S. 2021. "Phishing in Organizations: Findings from a Large-Scale and Long-Term Study," *ArXiv [Cs.CR]*, , December 14. (https://ieeexplore.ieee.org/document/9833766/).

Lenaerts-Bergmans, B. 2021. *How to Spot a Phishing Email: 7 Telltale Signs | CrowdStrike*, crowdstrike.com. (https://www.crowdstrike.com/cybersecurity-101/phishing/how-to-spot-a-phishing-email/).

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., and Stumpf, S. 2024. "Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions," *An International Journal on Information Fusion* (106), p. 102301.

Lundberg, S. M., and Lee, S.-I. 2017. "A Unified Approach to Interpreting Model Predictions," *Neural Information Processing Systems* (30), pp. 4765–4774.

Marin, I. A., Burda, P., Zannone, N., and Allodi, L. 2023. "The Influence of Human Factors on the Intention to Report Phishing Emails," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Vol. 7), New York, NY, USA: ACM, April 19, pp. 1–18.

Natarajan, P., and Nambiar, A. 2024. *Underwater SONAR Image Classification and Analysis Using LIME-Based Explainable Artificial Intelligence*, Arxiv.org. (https://arxiv.org/html/2408.12837v1?form=MG0AV3).

Otieno, D. O., Siami Namin, A., and Jones, K. S. 2023. "The Application of the BERT Transformer Model for Phishing Email Classification," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, June, pp. 1303–1310.

Ouwehand, K., Kroef, A. van der, Wong, J., and Paas, F. 2021. "Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales?," *Frontiers in Education* (6), Frontiers Media SA. (https://doi.org/10.3389/feduc.2021.702616).

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. 2003. "Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies," *The Journal of Applied Psychology* (88:5), American Psychological Association (APA), pp. 879–903.

Revilla, M., and Ochoa, C. 2017. "Ideal and Maximum Length for a Web Survey," *International Journal of Market Research* (59), pp. 557–565.

Ribeiro, M. T., Singh, S., and Guestrin, C. 2018. "Anchors: High-Precision Model-Agnostic Explanations," *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence* (32:1), Association for the Advancement of Artificial Intelligence (AAAI). (https://doi.org/10.1609/aaai.v32i1.11491).

Singh, K., Aggarwal, P., Rajivan, P., and Gonzalez, C. 2019. "Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails," *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual Meeting* (63), pp. 453–457.

Straub, D., and Gefen, D. 2004. "Validation Guidelines for IS Positivist Research," *Communications of the Association for Information Systems* (13:1), Association for Information Systems, p. 24.

Sweller, J. 1988. "Cognitive Load during Problem Solving: Effects on Learning," *Cognitive Science* (12), pp. 257–285.

Sweller, J., van Merrienboer, J. J. G., and Paas, F. G. W. C. 1998. "Cognitive Architecture and Instructional Design," *Educational Psychology Review* (10), pp. 251–296.

Tjoa, E., and Guan, C. 2023. "Quantifying Explainability of Saliency Methods in Deep Neural Networks with a Synthetic Dataset," *IEEE Transactions on Artificial Intelligence* (4:4), Institute of Electrical and Electronics Engineers (IEEE), pp. 858–870.

Valecha, R., Mandaokar, P., and Rao, H. R. 2021. "Phishing Email Detection Using Persuasion Cues," *IEEE Transactions on Dependable and Secure Computing* (19:2), Institute of Electrical and Electronics Engineers (IEEE), pp. 1–1.

Vaswani, A., Shazeer, N., and Parmar, N. 2017. *Attention Is All You Need*, proceedings.neurips.cc. (https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html).

Yu, Y., Ashok, S., Kaushik, S., Wang, Y., and Wang, G. 05 2023. *Design and Evaluation of Inclusive Email Security Indicators for People with Visual Impairments*.

Zheng, S., and Becker, I. 2023. "Checking, Nudging or Scoring? Evaluating e-Mail User Security Tools," *Symposium On Usable Privacy and Security*, usenix.org, pp. 57–76.