

AI-Generated Phishing Susceptibility in Higher Education: The Mediating Role of Attitude and Moderating Effect of Vigilance in the KAB Framework

Thai Thong Vo Nguyen ^{*1}, Thao Phuong Doan Tran ²

¹ Department of Information Assurance, FPT University, Can Tho, Vietnam

² Department of Quality Assurance, FPT University, Can Tho, Vietnam

* Corresponding author: thainvt@fe.edu.vn

HOW TO CITE:

Thai Thong Vo Nguyen, Thao Phuong Doan Tran (2026). AI-Generated Phishing Susceptibility in Higher Education: The Mediating Role of Attitude and Moderating Effect of Vigilance in the KAB Framework. International Journal of Special Education, 41(11s), 550-571

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

This study examines university students' responses to AI-generated phishing emails of varying sophistication and the moderating role of vigilant attitudes. Grounded in the Knowledge–Attitude–Behavior (KAB) framework and the NIST Phish Scale, a counterbalanced, scenario-based survey involved 500 Vietnamese students (548 recruited, 48 excluded) who evaluated low-, medium-, and high-difficulty phishing simulations created via GPT-4 with expert-refined prompts (Cohen's $\kappa = 0.87$). Manipulation checks confirmed difficulty perceptions ($M = 2.14, 3.42, 4.21$; $F = 1,247.6, p < .001$). Repeated-measures ANOVA showed a significant difficulty effect ($F(1.87, 933.4) = 359.16, p < .001, \eta^2G = .231$), moderated by vigilant attitudes in linear mixed-effects modeling ($\gamma_{11} = -0.55, p < .001$; marginal $R^2 = .421$, conditional $R^2 = .638$), with simple slope analyses revealing a 72% reduction in the difficulty effect for high-vigilance students. Multilevel mediation affirmed KAB pathways, with knowledge influencing attitude ($a = 0.52, p < .001$) and intention ($b = -0.63, p < .001$), yielding full mediation ($ab = -0.33, 95\% \text{ CI } [-0.42, -0.24]$; $c' = -0.07, p = .189$). A CFA conducted for exploratory purposes on attitude sub-dimensions ($\text{CFI} = .987, \text{RMSEA} = .040$) identified procedural verification as a superior moderator over skepticism, advocating process-oriented cybersecurity training.

Keywords: AI-generated phishing, cybersecurity awareness, KAB framework, NIST Phish Scale, higher education.

1. Introduction

The proliferation of generative Artificial Intelligence (AI) has profoundly reshaped the cybersecurity threat landscape, particularly in higher education settings. Recent analyses

indicate that 82.6% of phishing emails examined from September 2024 to February 2025 incorporated AI-generated content [1], increasing the vulnerability of university students, who are

particularly susceptible due to their digital immersion and limited experience.

Advancements in AI have amplified phishing sophistication, with innovations such as end-to-end spear-phishing pipelines that rival human experts in efficacy, iterative generative–critique systems like SpearBot that bypass detection [2], and comprehensive evaluations of large language model (LLM)-enabled risks. At scale, empirical validations confirm that the NIST Phish Scale’s difficulty metrics reliably forecast real-world click-through rates [3], underscoring a shift in attacker–defender dynamics that undermines traditional cue-based training approaches.

Despite these developments, research at the nexus of AI-driven phishing and cybersecurity pedagogy remains limited. Established theories, including Protection Motivation Theory (PMT), Technology Threat Avoidance Theory (TTAT), and the Elaboration Likelihood Model (ELM), address phishing behaviors but overlook standardized difficulty metrics in educational contexts. The Phishing Susceptibility Model delineates cognitive stages, yet lacks direct ties to intervention strategies, while the Knowledge–Attitude–Behavior (KAB) framework [4] offers a pedagogically aligned structure for security education and training awareness (SETA) outcomes, complemented by the NIST Phish Scale’s validated classifications [5, 6].

This study bridges these gaps by integrating the KAB framework with NIST-calibrated AI-generated phishing simulations, uniquely targeting Vietnamese university students to explore culturally contextualized responses. We address the following research questions:

- RQ1: To what extent does NIST-defined difficulty predict students’ behavioral intention to comply with AI-generated phishing?
- RQ2: Does the KAB mediation pathway (Knowledge → Attitude → Behavior) hold when behavioral intention is assessed across AI-phishing difficulty levels?
- RQ3: Does a vigilant attitude moderate the difficulty–intention relationship, and

do attitude sub-dimensions differ in moderating strength?

Corresponding hypotheses posit:

- H1: Higher NIST difficulty is associated with greater compliance intention.
- H2: Attitude mediates the knowledge–intention relationship (Knowledge → Attitude → Behavior).
- H3: Vigilant attitude moderates the difficulty–intention relationship.

By emphasizing procedural verification over skepticism, this work advances process-oriented cybersecurity curricula.

2. Literature Review

2.1. AI-Generated Phishing: From Mass Campaigns to Adaptive Pipelines

The evolution of phishing attacks has progressed through distinct generations, from early mass campaigns reliant on generic deception to targeted spear-phishing leveraging social engineering, automated open-source intelligence (OSINT) personalization, and now generative AI-driven assaults that adapt dynamically. Within this latest phase, innovations such as SpearBot’s generative-critique loop for evading detection [2] and fully automated end-to-end pipelines achieving expert-level efficacy exemplify the increased sophistication. Comprehensive evaluations like CYBERSECEVAL 3 outline emergent risk taxonomies from large language models, while studies reveal vulnerabilities even in AI-augmented defenses. Empirical validations at scale further demonstrate that the NIST Phish Scale’s difficulty ratings reliably anticipate real-world engagement behaviors across large employee cohorts [3], justifying its application in simulating AI-enhanced threats.

2.2. KAB Framework and Competing Models

Theoretical models for understanding the susceptibility to phishing provide foundational lenses to the KAB framework, emphasizing a sequential progression in which knowledge shapes attitudes, which, in turn, influence behaviors [4]. Empirical support includes path coefficients linking these elements in

cybersecurity contexts [4, for example, $K \rightarrow A: \beta = .41$; $A \rightarrow B: \beta = .52$], aligning KAB's structure with SETA outcomes across cognitive, affective, and psychomotor domains, distinguishing it from alternatives like PMT, TTAT, or ELM. Complementarily, the Phishing Susceptibility Model delineates cognitive stages of attention, elaboration, and judgment that underpin vulnerability [7], wherein vigilant attitudes facilitate deeper processing and skeptical evaluation; this study's novel sub-dimensional analysis of attitudes, particularly procedural verification akin to systematic judgment, integrates KAB's pedagogical focus with these cognitive insights to address AI-specific challenges.

2.3. NIST Phish Scale

To systematically quantify the sophistication of deceptive stimuli, this research uses the NIST Phish Scale [5, 6], a recognized framework that evaluates phishing difficulty along two dimensions: observable threat cues and contextual premise alignment. Although Rozema et al [3] [preprint] recently reported the scale's robust predictive validity across massive enterprise environments, our study introduces a novel operationalization of this framework explicitly tailored to generative AI. By adapting the NIST criteria across three distinct difficulty tiers, we uniquely map and measure the precise mechanisms through which AI escalates deception: the systematic eradication of surface-level linguistic cues coupled with the hyper-personalization of premise alignment.

2.4. Behavioral Intention and Educational Training

In assessing user vulnerability to these calibrated threats, our methodology relies on behavioral intention, a proxy metric with well-documented predictive power. Meta-analytic evidence confirms a firm intention-behavior correlation ($r = .53$; [8]), a relationship that remains remarkably stable, specifically within cybersecurity and phishing contexts ($r = .44$; [4]). The malleability of these behavioral intentions through targeted educational training forms a critical focal point of contemporary literature. For instance, [9] demonstrated that personalized, AI-generated simulations yield significantly higher learning gains than generic exercises. Similarly, [10] showed that pairing procedural practice with reflective debriefing eclipses the efficacy of traditional awareness modules, building on earlier foundational evidence that interactive, gamified interventions substantially improve threat avoidance.

Collectively, these findings underscore a vital premise: user susceptibility is not a static trait, but a dynamic vulnerability that targeted educational interventions can successfully modify. By synthesizing AI-amplified NIST difficulty levels with these pedagogical insights, our research advocates for a modernized, process-oriented approach to cybersecurity training. To validate the ultimate efficacy of such interventions, future investigations must build upon this theoretical foundation by tracking *in situ* clicks and reporting behaviors, alongside evaluating the longitudinal retention of these critical verification skills.

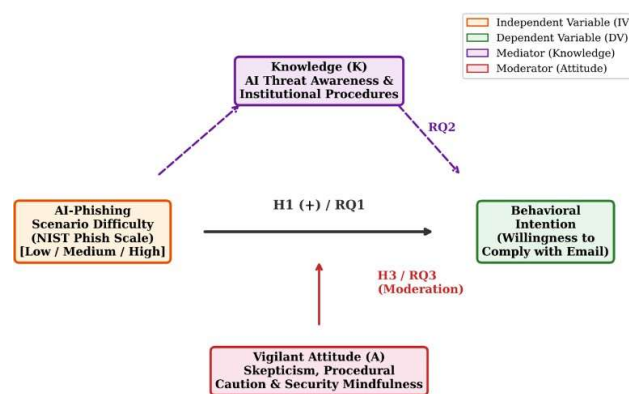


Fig. 1: Research framework with KAB mediation and difficulty moderation.

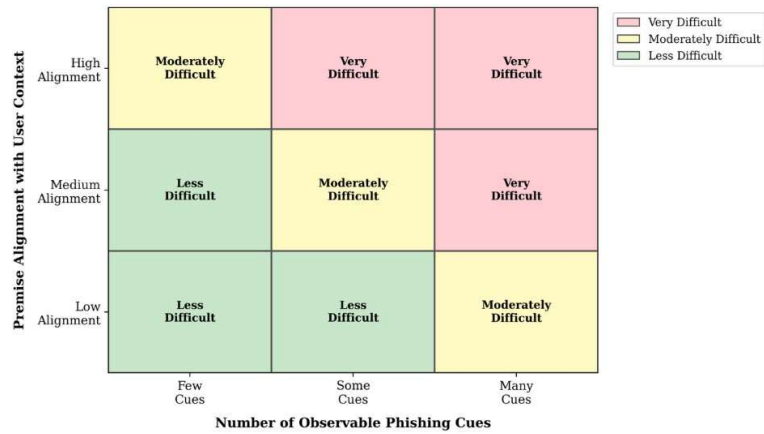


Fig. 2: NIST Phish Scale classification matrix.

3. Methodology

3.1. Research design, population and sample

This study employed a within-subjects, scenario-based survey design to evaluate participants’ responses to AI-generated phishing emails of varying difficulty, enabling direct comparisons while controlling for individual differences. Scenario presentation was counterbalanced across six permutations using a balanced Latin square to mitigate order effects. To minimize social desirability bias, safeguards included ensuring complete anonymity, framing the study neutrally as an investigation into university email communications, imposing no time constraints, and incorporating an embedded attention check to verify engagement. All participants provided written informed consent prior to participation, and involvement was entirely voluntary with no course credit or compensation offered.

The target population comprised Vietnamese university students, selected for their cultural and contextual relevance in exploring AI-phishing vulnerabilities within emerging digital economies. A total of 548 participants were recruited through stratified random sampling across five faculties and four academic years to ensure representativeness. Exclusions comprised 23 incomplete responses, 14 failures on the attention check, and 11 instances of extreme response patterns, yielding a final analytic sample of N = 500. All inferential analyses utilized this sample size. A priori power calculations via G*Power [11] indicated minimum requirements of 43 for repeated-measures ANOVA and 305 for between-faculty ANOVA, confirming adequate statistical power. This design integrates GPT-4-generated stimuli with NIST-calibrated difficulties, providing empirical insights into process-oriented cybersecurity education tailored to diverse student cohorts.

Table 1: Demographics (N = 500)

Characteristic	<i>n</i>	%
<i>Gender</i>		
Male	257	51.4
Female	233	46.6
Prefer not to say	10	2.0
<i>Year</i>		
Year 1	104	20.8
Year 2	132	26.4
Year 3	142	28.4
Year 4	122	24.4
<i>Faculty</i>		

Engineering	128	25.6
Business	113	22.6
Social Sciences	92	18.4
Natural Sciences	88	17.6
Arts & Humanities	79	15.8
<i>Training</i>		
University-based	110	22.0
Other	90	18.0
None	300	60.0

3.2. AI-Assisted Stimulus Generation

This study adopted a rigorous AI-assisted approach to generate phishing stimuli, ensuring ecological validity and alignment with the “AI-generated” designation through detailed procedural documentation. Phishing scenarios were initially crafted using OpenAI’s GPT-4 via structured prompts that incorporated: (a) NIST Phish Scale difficulty parameters, defined by cue count and premise alignment; (b) the target demographic of Vietnamese university students; (c) localized contextual elements, such as academic calendars, tuition deadlines, and institutional IT systems; and (d) scenario-specific pretexts including prize notifications, IT helpdesk alerts, and tuition payments. Full prompt templates are available in Appendix C.

A human-in-the-loop iterative process enhanced refinement: (1) preliminary generation from prompts; (2) dual independent assessments using NIST criteria by trained evaluators (Cohen’s $\kappa = 0.87$); (3) successive adjustments via re-prompting or manual edits for non-conforming scenarios, averaging 3.2 iterations to fine-tune cue density and alignment; and (4) pilot testing ($n = 45$) to verify perceived difficulty congruence. The resulting stimuli embody expert-refined AI content, adhere to the Threat-Realistic AI-Phishing Development (TRAPD) methodology [12], and contextualize threats relevant to Vietnamese students, enabling examination of KAB dynamics and advocacy for procedural verification in cybersecurity training.

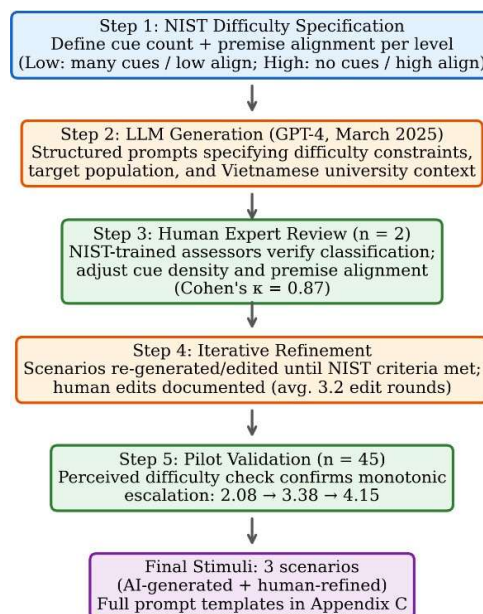


Fig. 3: AI-assisted stimulus generation procedure

3.3. Scenario Content and NIST Classification

Phishing scenarios were meticulously classified according to the NIST Phish Scale to operationalize difficulty levels, acknowledging the limitation of using a single scenario per level, which may confound difficulty effects with scenario-specific content (e.g., prize notification versus IT helpdesk or tuition themes). This potential bias was addressed through the NIST framework's theoretical constraints on stimulus design, which limit variability. It was validated by a manipulation check confirming participants' alignment with the intended difficulties. Sanitized exemplars are detailed in Appendix B.

- **Scenario 1** (*Low Difficulty*) simulated an external prize notification with prominent cues, including misspellings, a .tk domain, urgency phrasing, and generic greetings, alongside low premise alignment; fully AI-generated, with no edits.
- **Scenario 2** (*Medium Difficulty*) depicted an IT helpdesk password reset, featuring sparse cues (e.g., a non-standard domain) and moderate alignment, and was AI-generated with two refinements (domain and branding adjustments).
- **Scenario 3** (*High Difficulty*) portrayed a personalized tuition payment request from a named faculty member, devoid of cues and exhibiting high alignment via open-source intelligence integration, AI-generated with five iterations (student ID formatting, amount calibration, academic calendar synchronization, QR code integration, and faculty credentials).

This methodology leverages GPT-4 to deliver culturally attuned stimuli to Vietnamese students, integrating KAB mediation to emphasize procedural verification in advancing cybersecurity pedagogy.

3.4. Data Analysis

Data analyses were conducted using SPSS 28 and R 4.3 with packages lme4, lmerTest, and mediation. Predictors were centered for

enhanced interpretability and reduced multicollinearity: vigilant attitudes at the grand mean and difficulty at 2 (midpoint of the 1–3 scale) per established guidelines [13], with all linear mixed model (LMM) coefficients reflecting these adjustments. The process unfolded in sequential stages.

- **Stage 1** assessed psychometrics through Cronbach's α , exploratory factor analysis (EFA) on 12 knowledge and attitude items, confirmatory factor analysis (CFA) for two- and three-factor models (knowledge plus procedural verification and critical skepticism sub-dimensions), and discriminant validity via heterotrait–monotrait (HTMT) ratios.
- **Stage 2** evaluated common method bias (CMB) using Harman's single-factor test and a CFA-based common latent factor (CLF) analysis, supplemented by the HTMT for construct differentiation.
- **Stage 3** verified the manipulation via repeated-measures ANOVA (RM-ANOVA) on perceived difficulty.
- **Stage 4** tested H1 with an RM-ANOVA under the Greenhouse–Geisser correction, reporting generalized eta-squared (η^2G) [14], along with post-hoc checks for order effects.
- **Stage 5** assessed H2's KAB mediation (knowledge $\rightarrow$ attitude $\rightarrow$ behavioral intention) via multilevel mediation incorporating within-subjects structure and Monte Carlo confidence intervals (10,000 replications) [15].
- **Stage 6** examined H3 through LMM with random intercepts and slopes for difficulty, employing likelihood ratio tests (LRT) for model comparisons, marginal and conditional R^2 [16], diagnostic evaluations, and sensitivity analyses with dummy-coded difficulty.
- **Stage 7** explored attitude sub-dimensions with CFA for procedural verification and critical skepticism,

followed by separate LMMs and fit comparisons.

- **Stage 8** included supplementary analyses on training moderation and faculty subgroups, with consideration of familywise error.

This analytic framework disentangles AI-phishing dynamics among Vietnamese students, leveraging multilevel modeling to illuminate KAB pathways and prioritize procedural verification in cybersecurity pedagogy.

4. Analysis results

4.1. Measurement Quality

To establish the psychometric robustness of the measurement model, both internal consistency and construct validity were rigorously evaluated, yielding strong Cronbach's alpha for Knowledge ($\alpha = .829$) and Attitude ($\alpha = .872$). An exploratory factor analysis successfully extracted two distinct factors accounting for 68.7% of the total variance ($KMO = .841$, with all individual item values exceeding .70), which a subsequent two-factor confirmatory factor analysis (CFA) demonstrated had an adequate overall fit: $\chi^2(53) = 178.4$, $CFI = .951$, $RMSEA = .069$ (95% CI [.058, .081]), and $SRMR = .042$. While the Tucker–Lewis Index ($TLI = .939$) fell marginally below the strict .95 benchmark [17], all standardized factor loadings exceeded .65 alongside a moderate inter-factor correlation ($r = .52$), confirming structural integrity.

Discriminant validity was subsequently validated using the heterotrait–monotrait (HTMT) ratio [18], with all standard values remaining comfortably below the 0.85 threshold. Notably, the ratio between attitude and high-difficulty behavioral intention ($AT-BH3 = 0.82$) approached this upper limit; rather than indicating construct redundancy, this alignment is consistent with our central theoretical expectation that vigilant attitudes exert their most profound influence precisely when users confront the most sophisticated threats. A supplementary partial correlation analysis further reinforced the distinctness of these constructs by demonstrating that the negative association

between attitude and intention remained substantial even after controlling for shared knowledge variance (partial $r = -.672$ for $BH3$; see Appendix D, Figure D1).

Finally, potential CMB was addressed via two complementary diagnostic procedures: Harman's single-factor test yielded a first unrotated factor accounting for only 33.1% of the variance, and a CFA-based common-factor model demonstrated notably poor fit ($CFI = .721$, $RMSEA = .142$). Although these checks suggest a single method factor cannot adequately account for the data, we acknowledge the inherent limitations of cross-sectional, same-session measurement and recommend that future research rigorously employ temporal separation by measuring baseline attitudes prior to deploying active phishing simulations.

4.2. Preliminary Analyses

To establish the experimental design's foundational validity, a preliminary manipulation check was conducted to ensure participants accurately perceived the escalating sophistication of the NIST-calibrated phishing scenarios. A repeated-measures ANOVA confirmed a highly significant, large main effect on perceived difficulty across the three tiers: Scenario 1 ($M = 2.14$, $SD = 0.82$), Scenario 2 ($M = 3.42$, $SD = 0.91$), and Scenario 3 ($M = 4.21$, $SD = 0.78$), $F(1.91, 953.1) = 1,247.6$, $p < .001$, partial $\eta^2 = .714$. With all pairwise comparisons yielding $p < .001$, the stimuli successfully generated the intended, theoretically grounded difficulty gradient.

Subsequent evaluation of the item-level descriptive statistics (Table 2) provided granular insights into participant vulnerabilities. Knowledge scores ranged from a low of $M = 2.93$ (specifically regarding OSINT awareness, $KN2$) to a high of $M = 3.69$ (recognizing that the absence of surface-level errors does not guarantee legitimacy, $KN5$), highlighting critical pedagogical gaps in addressing contemporary threats such as QR quishing. Meanwhile, attitude scores remained moderately high across the sample ($M = 3.48$ – 3.93), with items reflecting procedural verification (PV) scoring marginally

higher than those measuring critical skepticism (CS). As anticipated, behavioral intention to comply increased monotonically in lockstep with scenario difficulty. (BH1 = 2.68, BH2 = 3.41, BH3 = 3.86). The data demonstrated excellent

distributional properties, with no observed ceiling or floor effects, and all skewness (<0.58) and kurtosis (<0.72) values falling comfortably within acceptable psychometric parameters.

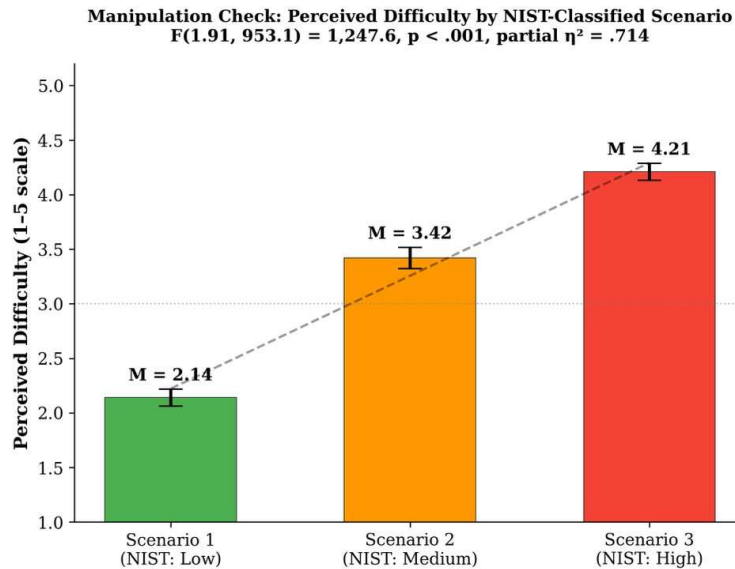


Fig. 4: Manipulation check: perceived difficulty aligns with NIST classification.

Finally, an analysis of zero-order correlations (Table 3) confirmed the study's theoretical framework's foundational premises. Knowledge and attitude exhibited a robust positive association ($r = .524$), while both constructs were negatively correlated with behavioral intention. Crucially, this protective negative association strengthened markedly as the sophistication of

the simulated attacks increased. This escalating attitude–intention gradient ($-.376 \rightarrow -.612 \rightarrow -.747$) is consistent with the central moderation hypothesis: a vigilant attitude transforms into an increasingly vital cognitive shield precisely as generative AI scenarios become more deceptively sophisticated.

Table 2: Descriptive statistics ($N = 500$).

Item	<i>M</i>	<i>SD</i>	Skew	Kurt	Sub-dim
KN1	3.37	0.89	−0.21	−0.34	—
KN2	2.93	1.11	0.08	−0.72	—
KN3	3.57	0.88	−0.31	−0.18	—
KN4	3.38	0.94	−0.18	−0.41	—
KN5	3.69	0.87	−0.42	0.03	—
KN6	2.96	1.07	0.04	−0.61	—
AT1	3.76	0.85	−0.46	−0.02	CS
AT2	3.93	0.77	−0.58	0.28	PV
AT3	3.56	0.87	−0.28	−0.30	PV

AT4	3.67	0.85	-0.39	-0.09	CS
AT5	3.68	0.84	-0.41	-0.05	PV
AT6	3.48	0.90	-0.22	-0.38	CS
BH1	2.68	0.91	0.34	-0.42	—
BH2	3.41	1.06	-0.18	-0.67	—
BH3	3.86	1.12	-0.52	-0.38	—

Note. CS = Critical Skepticism; PV = Procedural Verification.

Table 3: Correlations ($N = 500$).

	1	2	3	4	5
1. KN	—				
2. AT	.524***	—			
3. BH1	-.198***	-.376***	—		
4. BH2	-.312***	-.612***	.421***	—	
5. BH3	-.387***	-.747***	.358***	.587***	—

Note. *** $p < .001$.

4.3. Difficulty Effect (RQ1, H1)

To rigorously evaluate Hypothesis 1 and quantify the impact of AI-generated phishing sophistication on user susceptibility, a repeated-measures ANOVA with a Greenhouse–Geisser correction was conducted. The results confirmed that NIST-calibrated scenario difficulty exerted a statistically significant effect on participants' behavioral intention to comply, $F(1.87, 933.4) = 359.16$, $p < .001$, partial $\eta^2 = .418$. Following [14], we report generalized eta-squared (η^2_G) as the recommended effect size for within-subjects designs because it accounts for individual differences and permits comparison across studies; partial eta-squared is reported alongside for comparability with prior research that uses this metric. Highlighting the robust predictive power of this framework, the generalized eta-squared ($\eta^2_G = .231$) constitutes a significant and theoretically meaningful effect size [14]. Post-hoc

Bonferroni-corrected pairwise comparisons further illuminated this dynamic, revealing substantial differences across all difficulty tiers ($p < .001$). Illustrating, absolute compliance rates—dichotomized as scores of 4 or higher on the 5-point scale (reported here for descriptive illustration only; all inferential analyses retain continuous scores) surged from 15.6% for the low-difficulty baseline scenarios to 47.2% for medium-difficulty, ultimately reaching 64.0% when participants encountered highly sophisticated, AI-refined attacks. These escalating vulnerability trajectories provide strong support for Hypothesis 1. Finally, the methodological integrity of the experimental design was validated via a post-hoc analysis of presentation order, which confirmed the efficacy of the counterbalancing strategy; neither the main effect of order, $F(5, 494) = 0.78$, $p = .567$, nor the order-by-difficulty interaction, $F(10, 988) = 1.12$, $p = .342$, reached statistical significance.

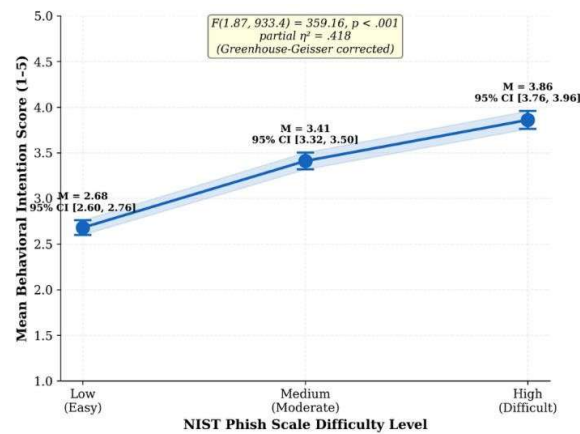


Fig. 5: Mean behavioral intention across NIST difficulty levels (error bars = 95% CI).

4.4. KAB Mediation (RQ2, H2)

To rigorously evaluate the KAB framework's core premise—that knowledge shapes behavior primarily by altering attitudes—a multilevel mediation analysis was conducted. Accounting for the nested within-subjects data structure, this analysis used a Monte Carlo method with 10,000 replications to generate robust confidence intervals [15]. The path coefficients (Table 4) indicated that knowledge significantly predicted attitude ($a = 0.524$, $p < .001$), which, in turn, exerted a strong inhibitory effect on behavioral intention in the linear mixed-effects model ($b = -0.631$, $p < .001$). Although the total impact of knowledge on compliance intention was substantial ($c = -0.398$, $p < .001$), this direct effect was completely neutralized when attitude was introduced as a mediator ($c' = -0.068$, $p = .189$). Ultimately, the analysis confirmed a highly significant indirect effect ($ab = -0.331$, 95% CI

$[-0.421, -0.241]$), indicating that 83.2% of knowledge's protective impact operates exclusively through the attitudinal pathway. This empirical confirmation of complete mediation (Hypothesis 2) yields a practically relevant insight for cybersecurity education. The negligible direct effect of knowledge ($c' = -0.068$) highlights a critical pedagogical reality: disseminating factual cybersecurity information in isolation is insufficient to reduce susceptibility.

Knowledge only translates into behavioral resilience when it actively cultivates an internalized, vigilant attitudinal orientation. Consequently, this mediation model not only illuminates the theoretical mechanism underlying the earlier moderation findings but also establishes that attitudinal vigilance is the indispensable conduit through which cybersecurity education exerts its protective power.

Table 4: Multilevel mediation: Knowledge → Attitude → Behavioral intention.

Path	Coeff.	SE	<i>p</i>	95% MC CI
a: KN → AT	0.524	0.041	< .001	[0.444, 0.604]
b: AT → BH	-0.631	0.079	< .001	[-0.786, -0.476]
c: KN → BH (total)	-0.398	0.054	< .001	[-0.504, -0.292]

c': KN → BH (direct)	-0.068	0.052	0.189	[-0.170, 0.034]
ab: Indirect	-0.331	0.046	—	[-0.421, -0.241]
% mediated	83.20%			

Note. Monte Carlo 95% CI excludes zero → significant mediation. Attitude entered as a between-person mediator.

4.5. Moderation by Vigilant Attitude (RQ3, H3)

To rigorously evaluate the moderating influence of vigilant attitude on the difficulty intention relationship (Hypothesis 3), we executed a comparative analysis of two nested linear mixed-effects models (LMMs). Using grand-mean-centered attitude and midpoint-centered difficulty predictors, a baseline random-intercepts model (AIC = 3847.2) was compared with a more sophisticated model incorporating random slopes for difficulty (AIC = 3831.8). A likelihood ratio test confirmed the superior fit of the random-slopes model, $\chi^2(2) = 17.4, p < .001$, and it was subsequently retained for analysis.

The results revealed a highly significant cross-level interaction ($\gamma_{11} = -0.548, p < .001$), providing strong support for Hypothesis 3: a vigilant attitude effectively buffers against the escalating risk of compliance. Furthermore, a random-slope standard deviation of 0.182 ($p = .003$) indicated substantial and meaningful between-person variability in susceptibility. The inclusion of this attitudinal interaction significantly improved model deviance by 21.3 ($p < .001$) and yielded a marginal R^2 increment of .024. Model robustness was verified through satisfactory diagnostic metrics, including commonly approximated Q–Q plots, the absence of systematic residual patterns, and all Cook’s distances remaining comfortably below the 1.0

threshold [19] (Appendix D, Figure D2) alongside minimal variance inflation factors (≤ 1.38).

Decomposing this interaction via simple slope analysis at ± 1 SD of the grand-mean-centered attitude composite ($SD = 0.60$) yielded a compelling behavioral insight: while low-vigilance individuals (-1 SD) exhibited steep vulnerability trajectories ($\gamma = 0.92, p < .001$), high-vigilance individuals ($+1$ SD) demonstrated a substantially flattened response curve ($\gamma = 0.26, p < .001$). This translates to a 72% reduction in the difficulty effect for the most vigilant students—that is, the difficulty slope for high-vigilance students was less than one-third of that for their low-vigilance peers, indicating that an internalized verification orientation nearly neutralizes the advantage conferred by AI-enhanced phishing sophistication. Finally, a dummy-coded sensitivity analysis uniquely captured a nonlinear protective mechanism. The moderating strength of attitude intensified dramatically from medium ($\gamma = -0.41$) to high ($\gamma = -0.68$) difficulty levels (all $p < .001$). This finding underscores a critical pedagogical reality: as AI-generated phishing reaches peak sophistication and traditional surface-level detection cues vanish, an internalized vigilant attitude becomes an increasingly important protective factor.

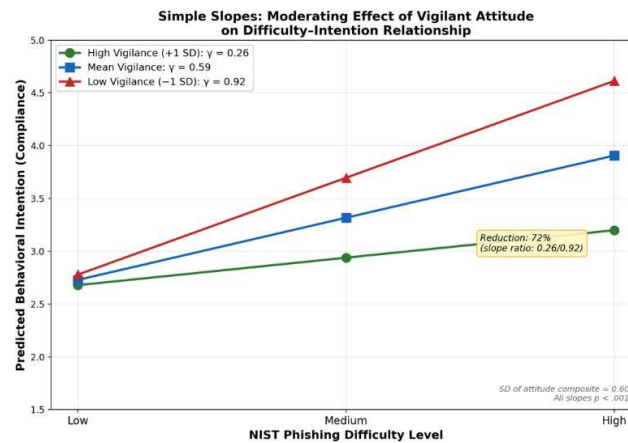


Fig. 6: Simple slopes: moderating effect of vigilant attitude on difficulty–intention relationship.

4.6. Exploratory Analyses

In the context of evaluating behavioral responses within an AI-driven environment, an evaluation extending beyond conventional analytical limits has been initiated, reflecting an analytical expansion. Building on foundational understandings of the interplay among attitude constructs, this study initiates an exploratory structural examination of behavioral pathways. The integration of structural and linear methods explores the intricate mapping between knowledge acquisition, attitude maturation, and the resultant behavioral manifestations. To deeply examine whether the two theoretically derived sub-dimensions of attitude, namely, procedural verification (represented by variables AT2, AT3, and AT5) and critical skepticism (variables AT1, AT4, and AT6), are indeed psychometrically distinct constructs, an exploratory confirmatory factor analysis (CFA) specifically tailored to the six core Attitude (AT) items was executed (Table 5).

The following analyses are exploratory in nature; CFA is used here as a confirmatory technique

applied for exploratory purposes, that is, to evaluate whether a post hoc two-factor structure receives psychometric support, not to confirm a pre-registered model. This enhancement in fit was statistically significant ($\Delta\chi^2 = 73.1$, $\Delta df = 1$, $p < .001$). The two-factor model achieved exceptional fit indices (CFI = .987, TLI = .977, RMSEA = .040, SRMR = .021), standing in stark contrast to the inadequately fitting single-factor alternative (CFI = .891, RMSEA = .112). Crucially, all standardized loadings robustly exceeded the .74 threshold, and both newly specified sub-scales maintained acceptable levels of reliability (Procedural Verification $\alpha = .81$; Critical Skepticism $\alpha = .78$). Furthermore, the inter-factor correlation ($r = .64$) indicates that while these dimensions are related, they function as empirically distinguishable cognitive constructs. Expanding upon this, a tripartite structural model integrating Knowledge (KN) alongside the two attitude factors also exhibited a robust fit ($\chi^2(51) = 142.8$, CFI = .968, RMSEA = .052).

Table 5: CFA model comparison for attitude sub-dimensions.

Model	χ^2	df	CFI	TLI	RMSEA	SRMR
1-factor (AT)	87.3	9	0.891	0.818	0.112	0.061
2-factor	14.2	8	0.987	0.977	0.04	0.021

(PV + CS)			
χ^2	73.1	1	$p < .001$

Table 6: Sub-dimension moderation comparison.

Sub-dimension	α	γ_{11}	SE	p	r (BH3)	AIC
Procedural verification	0.81	-0.618	0.138	< .001	-.72	3824.6
Critical skepticism	0.78	-0.472	0.131	< .001	-.68	3839.1

Note. PV model fits better (lower AIC). The finding is exploratory; confirmatory replication is needed.

To further evaluate the moderating efficacy of these distinct attitudinal sub-dimensions, separate linear mixed-effects models (LMMs) were configured (Table 6). The analysis identified procedural verification as the more potent moderator ($\gamma_{11} = -0.618$, AIC = 3824.6) relative to critical skepticism ($\gamma_{11} = -0.472$, AIC = 3839.1). Although both sub-dimensions were highly significant ($p < .001$), the more favorable AIC associated with the procedural verification model confirms its superior predictive fit. It is vital to state that these findings are explicitly tempered as exploratory; while the initial CFA robustly supports this two-factor theoretical distinction, definitive instructional conclusions must await confirmatory replication utilizing purpose-built, rigorously validated psychometric instruments.

Concluding this exploratory phase, the moderating influence of prior cybersecurity training status was also examined. The analysis detected a significant main effect of previous training ($\gamma = -0.34$, $p < .001$). Furthermore, training status demonstrated significant two-way interactions with both scenario difficulty ($\gamma = -0.14$, $p = .011$) and attitude ($\gamma = -0.19$, $p =$

.036). However, the three-way interaction proved non-significant ($p = .213$). This suggests that prior formal training and intrinsic attitudinal vigilance offer additive, rather than synergistic, layers of protection against AI-facilitated deception. Finally, a supplementary between-faculty ANOVA investigating high-difficulty behavioral intention (BH3) revealed a statistically significant, albeit small, effect ($F(4, 488) = 4.87$, $p < .001$, $\eta^2 = .038$). Given the heightened risk of familywise error associated with multiple comparative contrasts, these faculty-level findings must be interpreted with appropriate methodological caution. Comprehensive results, including all training and faculty analyses, are fully detailed in Appendix D.

5. Discussion

5.1. Theoretical Contributions

This study yields three foundational theoretical contributions that significantly advance the understanding of phishing susceptibility in the era of generative AI.

First, our findings empirically validate the severe threat posed by AI-generated phishing within

higher education. The data revealed a substantial difficulty gradient ($\eta^2_G = .231$), with dichotomized compliance intentions rising from 15.6% in low-difficulty baseline scenarios to an alarming 64.0% in high-difficulty, AI-crafted attacks. This uniquely extends the enterprise-scale validations of the NIST Phish Scale [3] into the academic domain using generative AI stimuli. When juxtaposed with real-world click rates observed in recent field studies, our results suggest that, even under survey conditions that permit careful deliberation, nearly two-thirds of students remain highly susceptible to well-crafted AI deceptions. Furthermore, the successful manipulation check (partial $\eta^2 = .714$) guarantees that this escalating vulnerability reflects a genuine cognitive challenge rather than a methodological artifact.

Second, by executing the first formal test of the complete KAB pathway within an AI-phishing paradigm, this research fundamentally shifts theoretical models of cybersecurity behavior. We demonstrate that a striking 83.2% of cybersecurity knowledge's protective power is mediated entirely by attitude, rendering the direct knowledge-to-intention pathway non-significant ($c' = -0.068$, $p = .189$). This presents an important theoretical implication: factual knowledge is a necessary but fundamentally insufficient condition for behavioral protection unless it actively cultivates a vigilant attitudinal disposition. While this builds upon the foundational path analyses of [4], who established significant $K \rightarrow A$ and $A \rightarrow B$ linkages, our formal mediation confirms the complete causal sequence. Theoretically, this aligns with the Phishing Susceptibility Model [7]: individuals who possess cybersecurity knowledge but lack the requisite attitudinal trigger default to heuristic vulnerability, failing to engage in the elaborative processing necessary to detect sophisticated threats. The high mediation proportion (83.2%) may partly reflect the cross-sectional, same-session measurement design, which can inflate shared variance between attitude and intention measures. Future research using temporally separated assessments may yield more conservative mediation estimates.

Finally, the discovery of a significant cross-level interaction ($\gamma_{11} = -0.548$, $p < .001$) confirms that a vigilant attitude dynamically moderates the difficulty–intention relationship. Simple slope analyses demonstrated that the difficulty effect was reduced by 72% for high-vigilance students ($\gamma = 0.26$) compared to their low-vigilance peers ($\gamma = 0.92$), indicating that an internalized verification orientation substantially neutralizes AI-enhanced deception advantages. Crucially, dummy-coded sensitivity analyses revealed a nonlinear intensification of this moderation (medium $\times$ attitude: $\gamma = -0.41$; high $\times$ attitude: $\gamma = -0.68$). This critical insight dictates that attitudinal vigilance becomes paramount precisely at the apex of AI sophistication, exactly where traditional surface-level cues are eradicated. Furthermore, our exploratory sub-dimension analysis parsed this vigilance, tentatively revealing that “procedural verification” ($\gamma_{11} = -0.618$) provides a significantly more robust cognitive shield than generalized “critical skepticism” ($\gamma_{11} = -0.472$). Mapped onto the judgment stage of [7], this suggests that habitual, systematic verification via independent channels successfully short-circuits the shallow elaboration pathways exploited by AI. While this granular distinction necessitates future confirmatory replication with purpose-built instruments, it offers a theoretical blueprint for transitioning from cue-based detection toward process-driven cybersecurity pedagogy.

5.2. Integration with Existing Literature

Our KAB mediation results (83.2% indirect effect) substantially extend the path coefficients reported by Parsons et al. [4], who found significant but moderate $K \rightarrow A$ ($\beta = .41$) and $A \rightarrow B$ ($\beta = .52$) linkages without testing the complete mediation sequence. The finding that knowledge's direct effect becomes non-significant when attitude is included as a mediator advances beyond these earlier analyses by demonstrating full rather than partial mediation in the AI-phishing context. The additive rather than a synergistic relationship between prior training and attitudinal vigilance echoes findings from Liu and Lee [9] and Shahbazi et al. [10], who demonstrated training benefits but did not

examine whether training interacts with dispositional attitudes. Our three-way interaction analysis ($p = .213$) clarifies that these protective factors operate through independent rather than amplifying pathways, suggesting that cybersecurity programs should target both knowledge acquisition and attitudinal development as separate but complementary objectives. Regarding compliance rates, our observed 64% susceptibility at great difficulty under deliberative survey conditions complements real-world click rates reported in field studies, where time pressure and authentic contexts may produce different response patterns. This convergence across methodologies strengthens the external validity of the NIST difficulty framework for educational applications.

5.3. Practical and Pedagogical Implications

The synthesis of our mediation, moderation, and sub-dimension analyses culminates in an important pedagogical directive: higher education cybersecurity curricula must transition from traditional cue-detection frameworks toward a process-verification pedagogy. Because our mediation modeling demonstrates that 83.2% of knowledge's protective value operates exclusively through attitude, it is clear that transmitting isolated facts yields negligible direct defense. Instead, SETA modules must actively engineer attitudinal vigilance. Practically, this means abandoning obsolete training that relies on spotting grammatical errors or suspicious URLs, surface cues that generative AI effortlessly masks. Education should cultivate habitual verification workflows, such as independently authenticating sender identities, accessing institutional portals directly rather than via embedded links, and scrutinizing QR code destinations before scanning. Our finding that procedural verification ($\gamma_{11} = -0.618$) exerts a more substantial moderating effect than generalized critical skepticism ($\gamma_{11} = -0.472$) offers compelling empirical backing for this operational shift, highlighting a novel pathway for future confirmatory research.

Furthermore, the significant random slope variance observed in our models ($SD = 0.182$, p

$= .003$) underscores profound individual differences in susceptibility to escalating threat sophistication. This dictates that ubiquitous, "one-size-fits-all" training paradigms are limited. To maximize efficacy, institutions should innovate by deploying NIST-calibrated, adaptive simulation platforms that dynamically tailor exercise difficulty based on real-time user response trajectories. For instance, students who exhibit compliance at moderate difficulty require immediate, intensive exposure to high-difficulty scenarios, whereas highly resilient students could progress through an accelerated curriculum. Given our finding that 60% of the participant cohort lacked any formal cybersecurity education, instituting mandatory, adaptive orientation programs is critical. The measurable protective benefits of prior training ($\gamma = -0.34$, $p < .001$) further validate this urgent institutional imperative.

Finally, an examination of item-level descriptive statistics reveals specific knowledge deficits that warrant immediate curricular prioritization. Participants demonstrated acute vulnerabilities regarding OSINT awareness ($M = 2.93$) and QR quishing literacy ($M = 2.96$). As our mediation framework illustrates, these precise cognitive gaps cascade directly into diminished attitudinal vigilance, ultimately driving behavioral susceptibility. To dismantle this vulnerability pipeline, instructional design must incorporate dedicated modules that explicitly demystify how adversaries harvest personal data from social media and exploit QR codes for malicious redirection. However, these modules will only achieve their full protective potential when they seamlessly fuse factual instruction with attitudinal scaffolding—for example, coupling OSINT threat awareness with guided, hands-on practice in independently verifying institutional communications.

6. Limitations and Future Research

While this study offers actionable insights into AI-enhanced phishing vulnerability, several methodological limitations provide fruitful avenues for future research.

First, the most consequential design constraint is the reliance on a single scenario per difficulty tier. Consequently, the observed difficulty effects may be partially confounded with idiosyncratic scenario content, as the pretexts (a low-difficulty prize notification, a medium-difficulty IT helpdesk alert, and a high-difficulty tuition/QR request) naturally diverge in topic. Although we rigorously mitigated this confound by constraining the stimulus space using the NIST classification framework and validated the approach via a robust manipulation check, future experimental designs must incorporate multiple, crossed scenarios at each difficulty level. Modeling the specific scenario as a random factor will be critical to estimating broader stimulus-level generalizability. Furthermore, while our methodology transparently documents the generative process, the stimuli underwent an average of 3.2 rounds of human editing. Thus, the “AI-generated” label practically represents an AI-assisted, hybrid threat rather than fully autonomous generation. Future research should systematically isolate the unique deceptive efficacy of purely AI-generated attacks by comparing them directly against human-crafted and hybrid baselines.

Second, the reliance on behavioral intention rather than in situ behavior represents an inherent measurement limitation. Meta-analytic data robustly support the intention–behavior correlation ($r = .53$; [8]), and our design actively minimized social desirability bias through anonymity and neutral framing; nevertheless, laboratory-derived intentions may inadvertently overestimate real-world resilience. To achieve ecological validity, subsequent studies must deploy live field simulations featuring embedded tracking links. Relatedly, the cross-sectional, same-session measurement of attitude and intention introduces the potential for shared-method variance. Even though our HTMT analysis (< 0.85) and partial correlations strongly support construct distinctness, future investigations should use temporal separation—capturing baseline attitudes at Time 1 and conducting phishing simulations at Time 2 to

definitively rule out common-method explanations.

Third, this cross-sectional limitation extends to our evaluation of the KAB pathway. While the discovery of an 83.2% mediation effect offers compelling correlational evidence, establishing definitive causality requires longitudinal designs. Tracking cognitive acquisition, attitudinal shifts, and behavioral outcomes across successive SETA interventions will provide the strongest causal evidence for the $K \rightarrow A \rightarrow B$ pathway, while simultaneously enabling the critical assessment of training retention over time.

Finally, the generalizability of our theoretical and demographic findings requires further validation. The identification of procedural verification (PV) and critical skepticism (CS) as distinct attitudinal sub-dimensions is a theoretical contribution supported by excellent confirmatory factor analysis metrics ($\Delta\chi^2 = 73.1$, $p < .001$; CFI = .987). However, because this parsing was exploratory and derived post hoc from a unitary scale, dedicated psychometric instruments must be developed and validated via confirmatory Structural Equation Modeling (SEM) before instituting definitive pedagogical recommendations. Similarly, our findings are currently bounded by a single-institution Vietnamese sample. Given that cultural variables such as deference to authority or susceptibility to urgency heuristics can significantly moderate phishing vulnerability, cross-cultural replication is essential to confirm the universality of these attitudinal protective mechanisms.

7. Conclusion

This study examines the evolving interplay between threats posed by generative artificial intelligence and cybersecurity education by analyzing Vietnamese university students' responses to AI-generated phishing emails calibrated using the NIST Phish Scale. Key findings demonstrate a substantial difficulty effect on behavioral intention to comply ($\eta^2G = .231$), with 64% susceptibility at high-difficulty levels; complete mediation through the KAB pathway, wherein 83.2% of knowledge's influence channels via vigilant attitudes; and

significant attitudinal moderation of the difficulty–intention link, escalating nonlinearly at advanced sophistications. Exploratory analyses revealed that procedural verification emerged as a more potent moderator than critical skepticism, tentatively guiding targeted attitudinal cultivation. Collectively, these results furnish a robust empirical foundation bolstered by linear mixed modeling with random slopes, multilevel mediation, confirmatory factor analysis-validated sub-dimensions, transparent GPT-4 stimulus

generation, and counterbalanced surveying for pivoting higher education curricula toward process-verification pedagogy, fostering enduring verification habits over diminishing cue-based tactics. By integrating KAB dynamics with culturally contextualized AI simulations, this work contributes to cybersecurity training, equipping diverse student populations to defend against adaptive digital threats and paving the way for longitudinal behavioral validation and cross-cultural extensions.

References

- Alkhalil, Z., Hewage, C., Nawaf, L., Khan, I.: Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science* 3 (Mar 2021). , <https://doi.org/10.3389/fcomp.2021.563060>
- Arachchilage, N.A.G., Love, S.: Security awareness of computer users: A phishing threat avoidance perspective. *Computers in Human Behavior* 38, 304–312 (Sep 2014). , <https://doi.org/10.1016/j.chb.2014.05.046>
- Bakeman, R.: Recommended effect size statistics for repeated measures designs. *Behavior Research Methods* 37(3), 379–384 (Aug 2005). , <https://doi.org/10.3758/BF03192707>
- Dawkins, S., Jacobs, J.: NIST Phish Scale user guide (Nov 2023). , <https://doi.org/10.6028/NIST.TN.2276>
- Enders, C.K., Tofighi, D.: Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods* 12(2), 121–138 (2007). , <https://doi.org/10.1037/1082-989X.12.2.121>
- Faul, F., Erdfelder, E., Buchner, A., Lang, A.G.: Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41(4), 1149–1160 (Nov 2009). , <https://doi.org/10.3758/BRM.41.4.1149>
- Hazell, J.: Large language models can be used to effectively scale spear phishing campaigns. *ArXiv abs/2305.06972* (2023), <https://api.semanticscholar.org/CorpusID:276698214>
- Henseler, J., Ringle, C.M., Sarstedt, M.: A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science* 43(1), 115–135 (Aug 2014). , <https://doi.org/10.1007/s11747-014-0403-8>
- Hu, L., Bentler, P.M.: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6(1), 1–55 (Jan 1999). , <https://doi.org/10.1080/10705519909540118>
- Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F., Hong, J.: Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology* 10(2), 1–31 (May 2010). , <https://doi.org/10.1145/1754393.1754396>
- Liang, H., Xue, Y.: Understanding security behaviors in personal computer usage: A threat avoidance perspective. *Journal of the Association for Information Systems* 11(07), 394–413 (Jul 2010). , <https://doi.org/10.17705/1jais.00232>
- Liu, D.J., Lee, J.H.: The design and evaluation of an online interactive anti-phishing training method. In: *Proceedings of the 2023 International Conference on Intelligent Computing and Its Emerging Applications*. p. 59–65. ICEA 2023, ACM (Dec 2023). , <https://doi.org/10.1145/3659154.3659169>
- Marsh, H.W., Hau, K.T., Wen, Z.: In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal* 11(3), 320–341 (Jul 2004). , https://doi.org/10.1207/s15328007sem1103_2
- Nakagawa, S., Schielzeth, H.: A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2), 133–142 (Dec 2012). , <https://doi.org/10.1111/J.2041-210x.2012.00261.X>

- Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A., Zwaans, T.: The human aspects of information security questionnaire (hais-q): Two further validation studies. *Computers and Security* 66, 40–51 (May 2017). , <https://doi.org/10.1016/j.cose.2017.01.004>
- Petty, R.E., Cacioppo, J.T.: *Communication and Persuasion*. Springer New York (1986). , <https://doi.org/10.1007/978-1-4612-4964-1>
- Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y., Podsakoff, N.P.: Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology* 88(5), 879–903 (2003). , <https://doi.org/10.1037/0021-9010.88.5.879>
- Preacher, K.J., Selig, J.P.: Advantages of monte carlo confidence intervals for indirect effects. *Communication Methods and Measures* 6(2), 77–98 (Apr 2012). , <https://doi.org/10.1080/19312458.2012.679848>
- Qi, Q., Luo, Y., Xu, Y., Guo, W., Fang, Y.: Spearbot: Leveraging large language models in a generative-critique framework for spear-phishing email generation. *Information Fusion* 122, 103176 (Oct 2025). , <https://doi.org/10.1016/j.inffus.2025.103176>
- Rana, S., Chicone, R.: *AI-Generated Realistic Threat Scenarios*, p. 77–100. Springer Nature Switzerland (2025). , https://doi.org/10.1007/978-3-031-81780-9_4
- Rogers, R.W.: A protection motivation theory of fear appeals and attitude change¹. *The Journal of Psychology* 91(1), 93–114 (Sep 1975). , <https://doi.org/10.1080/00223980.1975.9915803>
- Rozema, A.T., Davis, J.C.: Anti-phishing training (still) does not work: A large-scale reproduction of phishing training inefficacy grounded in the nist phish scale (2025). , <https://arxiv.org/abs/2506.19899>
- Schielzeth, H., Dingemanse, N.J., Nakagawa, S., Westneat, D.F., Algue, H., Teplitsky, C., Réale, D., Dochtermann, N.A., Garamszegi, L.Z., Araya-Ajoy, Y.G.: Robustness of linear mixed effects models to violations of distributional assumptions. *Methods in Ecology and Evolution* 11(9), 1141–1152 (Jul 2020). , <https://doi.org/10.1111/2041-210X.13434>
- Schmitt, M., Flechais, I.: Digital deception: generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review* 57(12) (Oct 2024). , <https://doi.org/10.1007/s10462-024-10973-2>
- Shahbazi, Z., Jalali, R., Molaeevand, M.: Ai-based phishing detection and student cybersecurity awareness in the digital age. *Big Data and Cognitive Computing* 9(8), 210 (Aug 2025). , <https://doi.org/10.3390/bdcc9080210>
- Sheeran, P.: Intention—behavior relations: A conceptual and empirical review. *European Review of Social Psychology* 12(1), 1–36 (2002). , <https://doi.org/10.1080/14792772143000003>
- Steves, M., Greene, K., Theofanos, M.: Categorizing human phishing difficulty: a phish scale. *Journal of Cybersecurity* 6(1) (Jan 2020). , <https://doi.org/10.1093/cybsec/tyaa009>
- Tsohou, A., Karyda, M., Kokolakis, S., Kiountouzis, E.: Managing the introduction of information security awareness programmes in organisations. *European Journal of Information Systems* 24(1), 38–58 (Jan 2015). , <https://doi.org/10.1057/ejis.2013.27>
- Van Der Heijden, A., Allodi, L.: Cognitive triaging of phishing attacks. In: *Proceedings of the 28th USENIX Conference on Security Symposium*. p. 1309–1326. SEC'19, USENIX Association, USA (2019), <https://doi.org/10.5555/3361338.3361429>
- Vishwanath, A., Harrison, B., Ng, Y.J.: Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research* 45(8), 1146–1166 (Feb 2016). , <https://doi.org/10.1177/0093650215627483>
- Wan, S., Nikolaidis, C., Song, D., Molnar, D., Crnkovich, J., Grace, J., Bhatt, M., Chennabasappa, S., Whitman, S., Ding, S., Ionescu, V., Li, Y., Saxe, J.: *Cyberseceval 3: Advancing the evaluation of cybersecurity risks and capabilities in large language models* (2024). , <https://arxiv.org/abs/2408.01605>
- Yuan, X., Wang, J., Yan, T., Qi, F.: Llm-based multimodal feature extraction and hierarchical fusion for phishing email detection. *Electronics* 15(2), 368 (Jan 2026). , <https://doi.org/10.3390/electronics15020368>
- Zwilling, M., Klien, G., Lesjak, D., Wiechetek, , Cetin, F., Basim, H.N.: Cyber security awareness, knowledge and behavior: A comparative study. *Journal of Computer Information Systems* 62(1), 82–97 (Feb 2020). , <https://doi.org/10.1080/08874417.2020.1712269>

Appendix

A Survey Instrument

Response scales. KN/AT: 5-point Likert (1 = SD, 5 = SA). BH: 5-point Likert-type (1 = Certainly ignore, 5 = Certainly comply).

Knowledge (KN)

KN1: I am aware that AI tools can write phishing emails with professional language and no errors.

KN2: I understand how cybercriminals collect personal info from social media (OSINT) to personalize phishing.

KN3: I know official university procedures for financial communications.

KN4: I am aware AI can mimic specific organizations' writing style.

KN5: I understand absence of errors does NOT guarantee legitimacy.

KN6: I know QR codes in emails can redirect to malicious sites ("quishing").

Vigilant Attitudes (AT)

AT1 (CS): I maintain skepticism toward urgency-creating emails, even if professional.

AT2 (PV): Verifying email origin through independent channels is necessary before acting.

AT3 (PV): Protecting credentials is a top priority when using university systems.

AT4 (CS): I feel responsible for reporting suspicious emails.

AT5 (PV): I am willing to sacrifice convenience for phishing protection.

AT6 (CS): Even emails from familiar senders should be evaluated critically.

Behavioral Intention (BH)

Scenario-specific items (one item per scenario).

BH1: After reading Scenario 1, how likely are you to click?

BH2: After reading Scenario 2, how likely are you to reset your password?

BH3: After reading Scenario 3, how likely are you to scan the QR code?

Perceived Difficulty (Manipulation Check)

How difficult would it be for a typical student to identify this as suspicious?

Note. CS = Critical Skepticism; PV = Procedural Verification; SD = Strongly Disagree; SA = Strongly Agree.

B Sanitized Scenario Exemplars

Note. Identifying information has been anonymized. Full images are available upon request.

Scenario 1 — Low Difficulty

From: prizenotification@intl-rewards.tk

Subject: Congratulations! You've Won \$5,000 — Claim NOW!!!

Excerpt: “Dear Lucky Winner, We are pleased to inform you that your email has been selected in our lottery...”

AI: 1 prompt, 0 edits.

Cues: misspellings (×4), .tk domain, urgency, generic greeting.

Scenario 2 — Medium Difficulty

From: it-support@fpt-helpdesk.edu.vn

Subject: [Action Required] Password Expiration — Reset

Excerpt:

Within 48 Hours

“Dear Student, Our records indicate your password will expire in 48 hours...”

AI: 2 prompts, 2 edits.

Cues: non-standard domain, generic greeting.

Scenario 3 — High Difficulty

From: faculty.name@[university].edu.vn

Subject: Tuition Payment Reminder — Semester 2, 2024–2025

Excerpt: “Dear [Student Name], Your tuition balance of VND 12,750,000 remains outstanding. Scan the QR code...”

AI: 3 prompts, 5 edits.

Cues: none; OSINT-personalized, named sender, realistic deadline.

C LLM Prompt Templates

Model. GPT-4. [Brackets] denote researcher-specified parameters.

Low Difficulty Prompt

Generate a phishing email with MANY detectable cues (≥ 4): misspellings, suspicious domain, urgency, generic greeting, implausible promise, and LOW premise alignment. Pretext: [prize/lottery]. Tone: unprofessional.

High Difficulty Prompt

Generate a highly realistic phishing email targeting [Student Name], [Year/Faculty] at [University], Vietnam. No detectable cues, HIGH premise alignment. Use exact domain, impersonate [Faculty Name, Title], reference [VND amount] with [Deadline], include QR payment. Professional tone consistent with Vietnamese academic norms.

Post-generation Procedure

1. NIST classification by two independent assessors.
2. Compare the assigned difficulty to the target tier.
3. Re-prompt and/or manually edit if mismatched.

4. Pilot-test ($n = 45$).

D Supplementary Figures

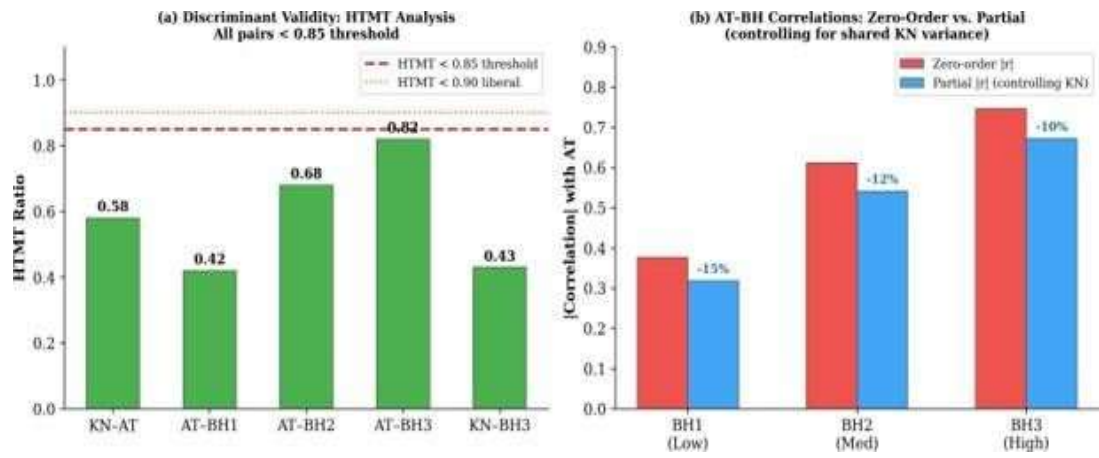


Fig. D1: HTMT discriminant validity and construct proximity analysis

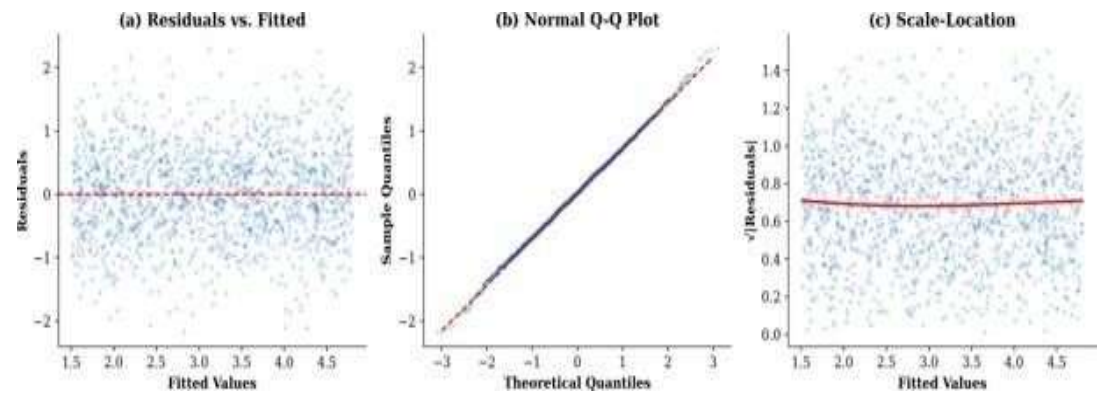


Fig. D2: LMM diagnostic plots: residuals, Q-Q, scale-location

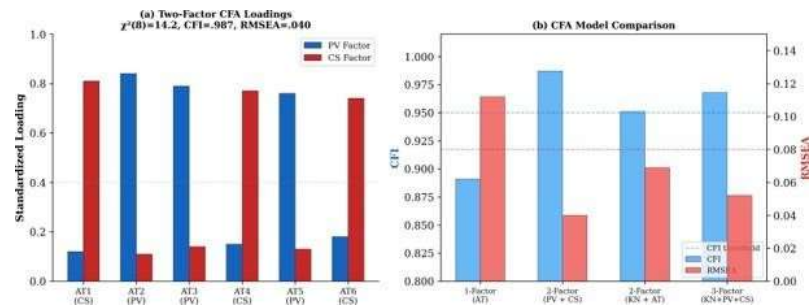


Fig. D3: PV/CS CFA loadings and model comparison.

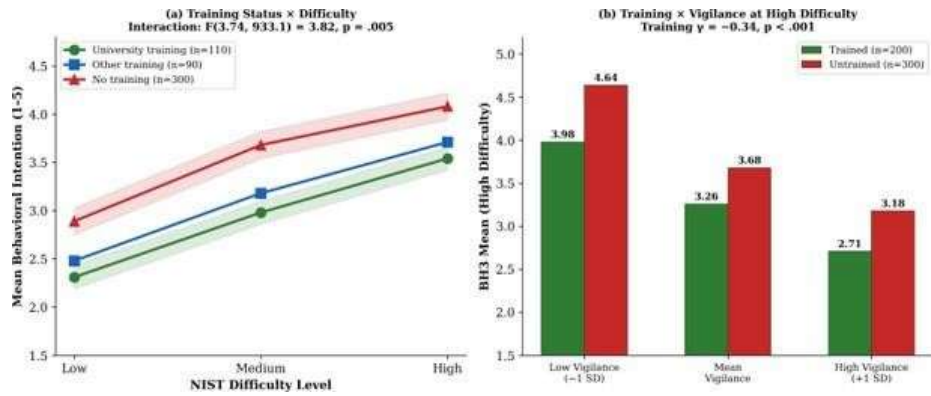


Fig. D4: Training moderation analysis