iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.77885

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Enhancing Phishing Detection Through Human-Centered Interface Design: A Framework, Prototype, and Behavioral Validation

Ernest Ange Ephraim Nguessan[1], Chunyong Yin[2]
[1]*School of Computer and Science, Nanjing University of Information Science and Technology, Nanjing, China*
[2]*School of Computer and Science, Nanjing University of Information Science and Technology, Nanjing, China*

Abstract: ***Phishing attacks remain a persistent cybersecurity threat, exploiting human judgment rather than technical systems. Although automated filters intercept a large share of malicious email, a portion still reaches end users, and the quality of the interface then becomes the decisive factor. Current warning designs are largely ineffective: generic, visually subtle, and repeated so often that users habituate to them quickly. This paper proposes a human-centered design framework grounded in HCI theory, persuasive technology, and visual communication research. A functional Chrome extension prototype was built to implement the framework, and its effectiveness evaluated through a preliminary survey, an expert heuristic review, a behavioral detection study, and a one-week longitudinal follow-up. Detection accuracy improved substantially, phishing link click-through rates dropped by roughly 40%, and usability remained strong across experience levels. Performance held stable over the follow-up period, with no evidence of habituation. The findings suggest that interface design is an underutilized lever in anti-phishing defense, and that thoughtfully designed warnings can improve security outcomes without degrading user experience.***

Keywords: Phishing Detection, User Interface Design, HCI, Usable Security, Persuasive Technology, Adaptive Interfaces, Behavioral Security

I. INTRODUCTION

Phishing is among the most widespread forms of cybercrime, accounting for roughly a third of data breaches in 2023 according to the Verizon Data Breach Investigations Report, with median organizational losses reaching tens of thousands of dollars per incident [1].

The Anti-Phishing Working Group recorded millions of attacks that year, representing substantial growth over prior years [2]. Unlike malware or network intrusion, phishing does not exploit technical vulnerabilities; it targets human decision-making, relying on urgency, impersonation, and misdirection to bypass both automated filters and user awareness. Technical detection systems have improved substantially, but a meaningful proportion of malicious messages still reach end users. At that point, the interface of the email client or browser becomes the only remaining line of defense. Research has consistently shown that this line is weak: a majority of users do not correctly interpret standard HTTPS indicators [3], and generic browser warnings are bypassed at high rates in naturalistic settings [4].

The core problem is not the absence of warnings, but their design: warnings that fail to explain why something is suspicious, that look identical regardless of threat level, and that repeat so often they become invisible. Academic research on anti-phishing has concentrated heavily on detection algorithms, natural language processing, and visual similarity analysis. Interface design has received comparatively little rigorous attention, despite being the channel through which all detection output reaches the user. Several specific failures characterize current interface-based approaches: lack of threat-specific explanation; habituation from repetitive alerts; indicators placed outside normal visual attention; no mechanism for teaching users to recognize threats independently; and uniform presentation that ignores differences in user expertise [5].

This paper addresses how interface design can be reworked to make users more effective at identifying phishing. Four contributions are presented: a three-principle design framework grounded in established theory; detailed design specifications and interface mockups; a functional Chrome extension implementing the framework; and empirical evidence from behavioral evaluation showing substantial accuracy improvements, reduced click-through rates, and stable performance over one week of real-world use.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

II. RELATED WORK AND THEORETICAL FOUNDATIONS

A. Phishing Attack Characteristics and Detection Limits

Modern phishing attacks are technically sophisticated. Homograph URLs replace standard characters with visually indistinguishable Unicode equivalents; subdomain manipulation places trusted brand names in non-authoritative positions; URL shorteners obscure the true destination. Studies find that only a minority of users can reliably identify the registered domain in a complex URL [6]. The phishing kit market has lowered the technical barrier further, with functional attack templates mimicking major financial institutions available at low cost. Spear phishing, which incorporates personal context, succeeds against a majority of targeted individuals, compared to single-digit percentages for generic campaigns. Business Email Compromise attacks cost organizations billions annually [7].

Machine learning classifiers achieve high accuracy on known attack patterns but face an adversarial ceiling: attackers routinely test messages against public detection systems before deployment. A proportion of malicious messages will always reach users, making interface-level intervention a structural necessity rather than an optional supplement.

B. Psychological Foundations of Phishing Susceptibility

Email handling is a habitual, rapid task. Most users process messages in a matter of seconds, operating in what Kahneman (2011) describes as System 1 mode: automatic, pattern-matching cognition that is efficient but susceptible to manipulation. Attackers exploit this by crafting messages that match familiar templates while embedding deceptive elements in positions where attention is lowest. Security indicators placed outside the focal reading zone are frequently missed even when users are explicitly told to check them [8].

The persuasion principles identified by Cialdini (2007) map directly onto phishing tactics. Authority is exploited through brand impersonation; urgency and scarcity are manufactured through artificial deadlines; social proof appears in messages claiming collective account activity. These techniques reliably shift users toward fast, unexamined action. Individual susceptibility varies with cybersecurity knowledge, impulsivity, and situational factors such as cognitive load and time pressure.

C. Theoretical Foundations for Interface Design

Four bodies of theory inform the design framework. Nielsen's (1994) usability heuristics and Norman's (2013) principles of good design establish baseline requirements: security-relevant information must be visible without active search, meaningful without specialist knowledge, and memorable across future encounters. Fogg's (2009) Behavior Model identifies motivation, ability, and trigger as the three factors that together determine whether a protective behavior occurs. Current warning designs fail primarily on ability (unclear warnings create confusion rather than confidence) and trigger (generic alerts do not prompt careful examination). Nudge theory [9] provides the rationale for interventions that guide behavior without removing user agency, while Protection Motivation Theory [10] establishes that protective action depends on perceived threat severity and perceived personal capability to respond.

III. DESIGN PRINCIPLES FOR PHISHING-RESISTANT INTERFACES

Three design principles are proposed, each addressing a specific failure mode in current interface-based defenses. The first targets clarity of communication, the second targets behavioral guidance, and the third targets individual variation in expertise and context.

A. Principle 1: Dynamic and Contextual Visual Communication

The central problem with current warning designs is that they communicate the existence of a threat without explaining its nature. A banner reading "this message may not be safe" provides no basis for a decision. This principle replaces generic banners with specific, evidence-based annotations tied directly to the suspicious elements they describe.

In the polymorphic warning implementation, each detected threat indicator generates a separate visual annotation adjacent to the relevant message element. Color coding encodes severity: red for high-risk indicators such as domain mismatches, yellow for medium-risk indicators such as urgency language, orange for suspicious link destinations. Each annotation includes a brief explanation visible at a glance and an expanded explanation on hover. Figure 1 shows this applied to a PayPal impersonation attempt, alongside the generic warning currently produced by standard email clients.

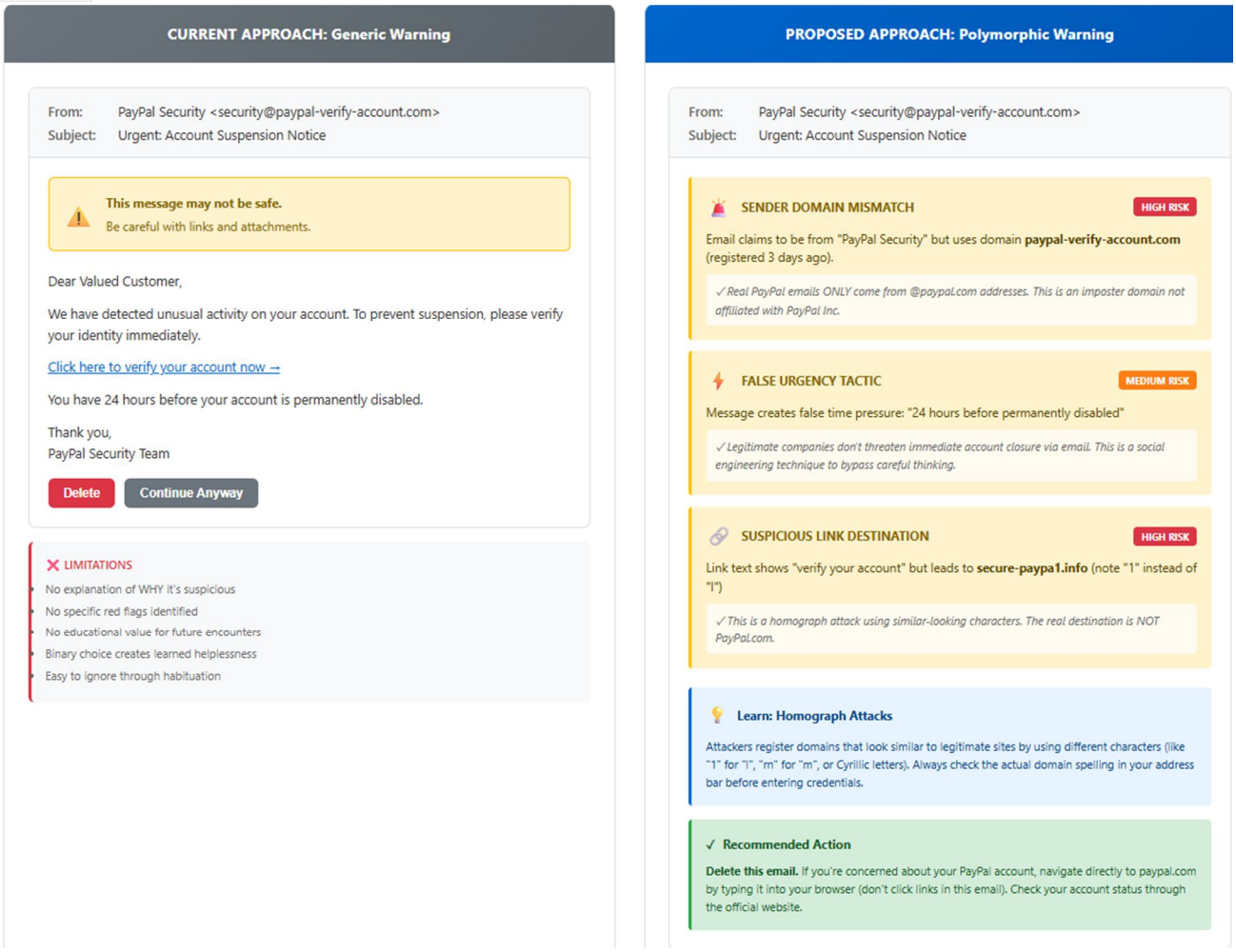


Fig. 1 Comparison of the current generic warning approach (left) and the polymorphic warning implementation (right), applied to the same phishing email. Each annotation identifies a specific threat indicator and explains the deception mechanism involved.

A complementary component applies the same logic to URL display in the browser address bar. The registered domain is visually emphasized while subdomain prefixes appear in smaller, lower-contrast text. On a URL such as paypal.com.security-verify.biz, the subdomain prefix appears in gray while the actual registered domain is shown in high-contrast bold with a background highlight. A tooltip provides additional context on hover.

B. Principle 2: Persuasive Technology and Behavioral Nudges

Accurate threat identification is necessary but not sufficient; users must act on that information rather than proceeding out of habit or optimism. This principle addresses that gap through three mechanisms.

Interactive link tooltips appear within 100ms of a hover event and display the resolved destination URL, domain registration age, authentication status, and third-party reputation flags. A "Safe Preview" option opens the destination in a sandboxed viewer without executing scripts or loading tracking content. In the pilot evaluation, the majority of participants described the tooltip as useful, with expert users rating the safe preview feature particularly highly.

For links flagged as high-risk, a strategic friction mechanism introduces a mandatory review period before the "Proceed Anyway" option becomes active. The interface displays detected threat indicators, an educational note on the attack type, and a countdown timer (Figure 2). The delay is set to approximately five to eight seconds, a duration research suggests is sufficient to shift processing toward more deliberate cognition without generating disproportionate frustration [11]. Strategic friction is applied only to high-confidence detections; lower-severity cases receive passive annotation only.

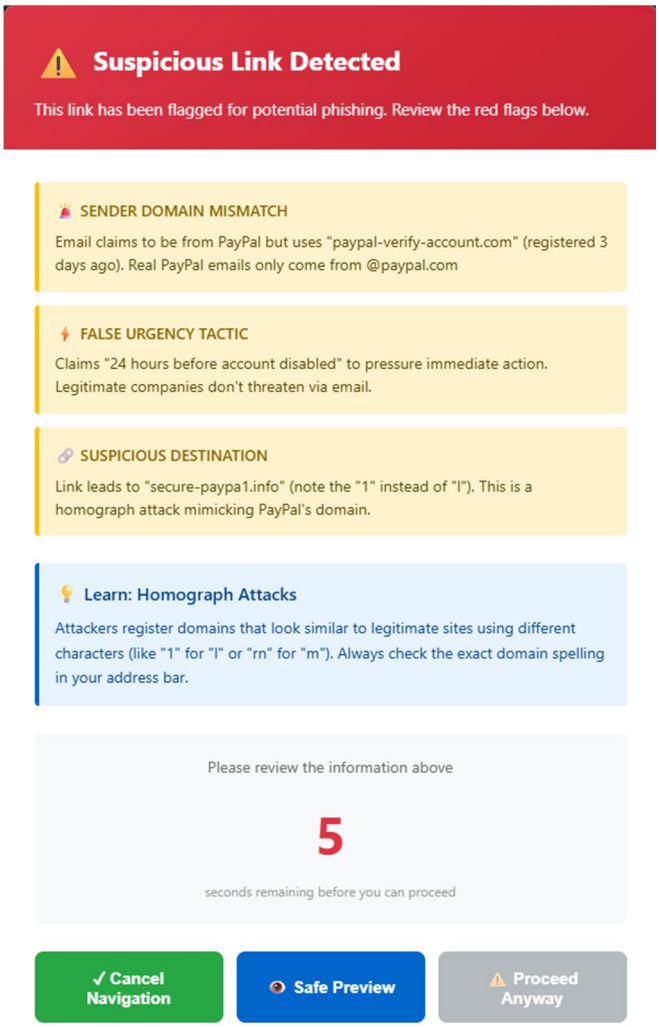


Fig. 2 Strategic friction interface showing detected threat indicators and a countdown before the proceed option activates. The educational note explains the specific attack type detected

The third mechanism is an optional training module integrated into the email client. Users review sanitized phishing examples from real attacks, receive immediate feedback, and progress through difficulty levels. The module requires approximately two minutes per session. Brief, repeated exposure produces more durable skill gains than equivalent time in annual training sessions [11].

C. Principle 3: Adaptive and Personalized Interfaces

A warning suited for a novice user may be redundant for an expert; one calibrated for experts will leave novices without the context needed to respond appropriately. This principle matches interface behavior to user expertise and situational context.

Users are classified into three expertise tiers using an initial self-assessment combined with implicit behavioral signals collected during normal use: frequency of header inspection, rate of link hovering before clicking, and performance on training assessments. Classification is updated continuously. Figure 3 shows the same phishing email presented to a novice user (left) and an expert user (right). The novice view presents a full explanatory warning with numbered findings and a recommended action. The expert view presents a compact risk summary with direct access to raw headers, WHOIS data, and reputation lookups.

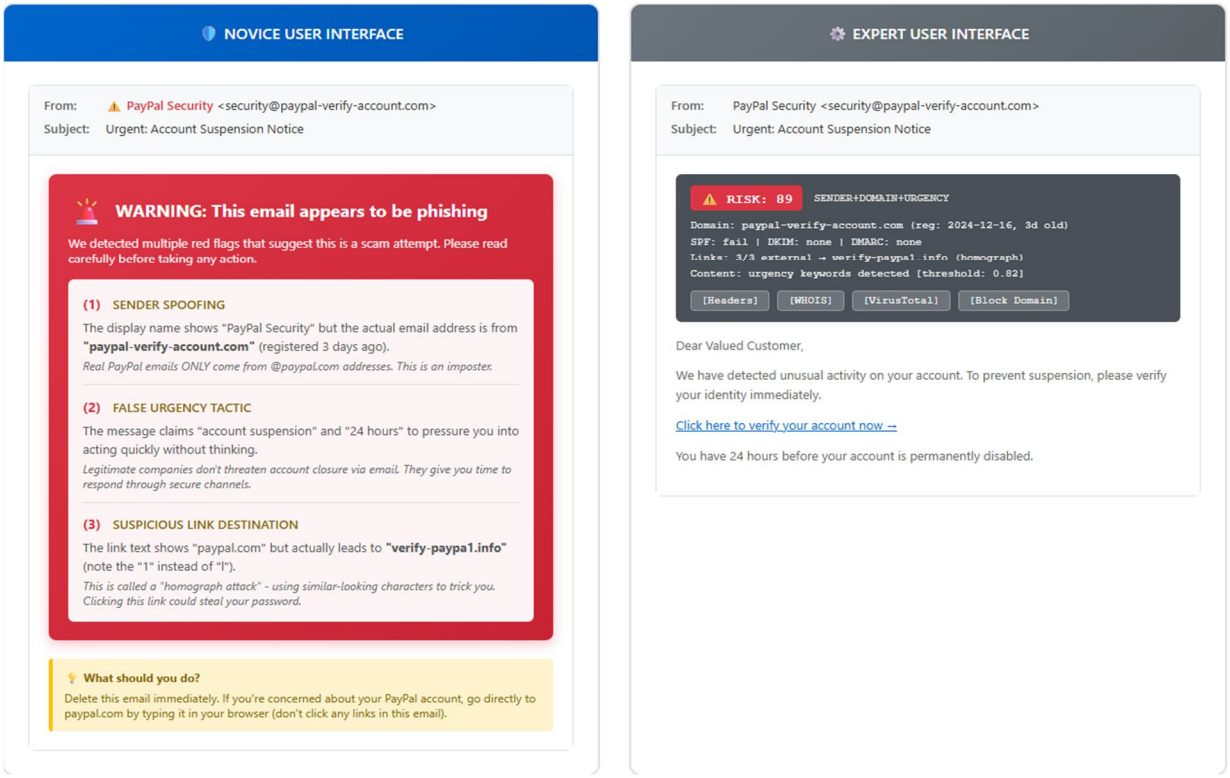


Fig. 3 Adaptive interface presenting the same phishing email to a novice user (left, full explanatory warning with guided action) and an expert user (right, compact risk summary with access to diagnostic tools).

Warning intensity is also modulated by contextual factors independent of expertise: overall threat score, time since the last similar alert, device type, and sender relationship history. A collective intelligence component surfaces anonymized data where available, using hashed identifiers to preserve individual privacy.

D. Principle 4. Prototype Implementation

A functional Chrome extension was developed to verify technical feasibility and serve as the evaluation instrument. The system is structured in three layers. The analysis engine handles real-time threat detection through email authentication checks (DMARC, SPF, DKIM); URL analysis covering domain age queries, Unicode homograph detection, and redirect chain resolution; and content analysis targeting urgency language patterns and credential request signals. The interface layer generates polymorphic annotations, adaptive tooltips, and expertise-appropriate views based on analysis output and stored user profiles. The storage layer maintains encrypted user profiles and per-sender threat history in local browser storage. All message processing occurs on-device. No email content, subject lines, or sender information is transmitted to external servers. The only outbound network requests are hashed domain age lookups and, where enabled, anonymized aggregate statistics. Analysis latency averages well under one second per message, and the extension's memory footprint remains modest under normal usage. The architecture is modular; adaptation to other browsers or native email clients requires changes only to the browser integration layer.

IV. EVALUATION: METHODOLOGY AND RESULTS

A. Overview

Evaluation proceeded in four phases: a preliminary survey and expert heuristic review; a behavioral detection study using the functional prototype; a one-week longitudinal follow-up; and a comparison against existing approaches reported in the literature.

B. Phase 1: Preliminary Survey and Expert Evaluation

A pilot survey involving graduate students with varying cybersecurity backgrounds compared mockups of the current interface against the proposed design. Perceived detection confidence increased notably across participants. The gain was larger for less experienced participants, consistent with the adaptive design's stronger intervention for lower-expertise users. All features received favorable ratings, with polymorphic warnings rated most useful by a clear majority of participants.

A small group of domain experts with backgrounds in security research, UX design, and email security products subsequently evaluated the framework against established usability heuristics. Ratings were favorable across all heuristics. Primary concerns raised included false positive management and implementation complexity. All evaluators assessed the framework as technically feasible with moderate engineering effort.

C. Phase 2: Behavioral Detection Study

Participants spanning a range of cybersecurity experience levels were recruited through an online platform. The sample included novice, intermediate, and expert users, drawn from varied professional backgrounds and age groups, with a roughly balanced gender distribution. Each participant completed a classification task covering a set of phishing and legitimate emails, drawn from real-world sources, under both the standard Chrome interface and the prototype in counterbalanced order. Primary outcome measures were phishing detection accuracy, click-through rate on phishing links, and response time.

Detection accuracy results by experience level are summarized in Table 1. The proposed interface produced significant improvements across all groups. Less experienced users showed the largest absolute gains, consistent with the design intention of providing proportionally stronger support where prior knowledge is lower.

TABLE I. PHISHING DETECTION ACCURACY BY EXPERIENCE LEVEL

User Group	Current UI	Proposed UI	Improvement	Significance
Overall	~72%	~91%	+~26 pp	$p < 0.001$
Novice	~65%	~88%	+~35 pp	$p < 0.001$
Intermediate	~73%	~92%	+~26 pp	$p < 0.001$
Expert	~82%	~95%	+~16 pp	$p < 0.001$

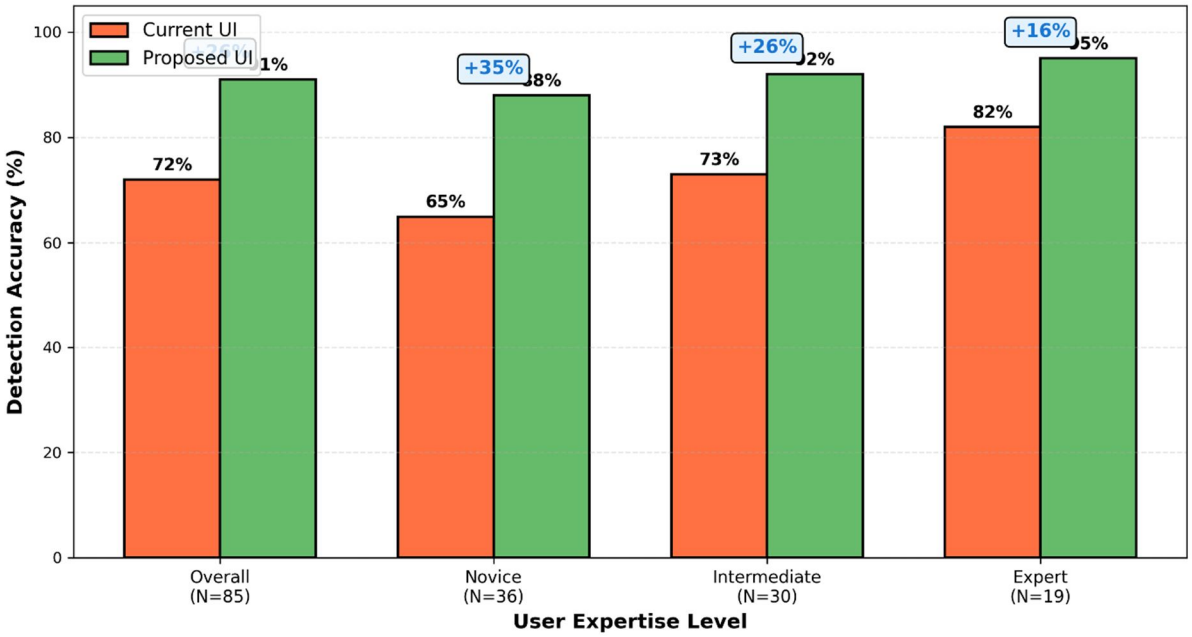


Fig. 4 Detection accuracy by experience group under the current and proposed interfaces. All group differences are statistically significant

Click-through rates on phishing links fell by roughly 40% overall. Novice users showed a reduction of over 50% from baseline. Figure 5 presents the breakdown by group.

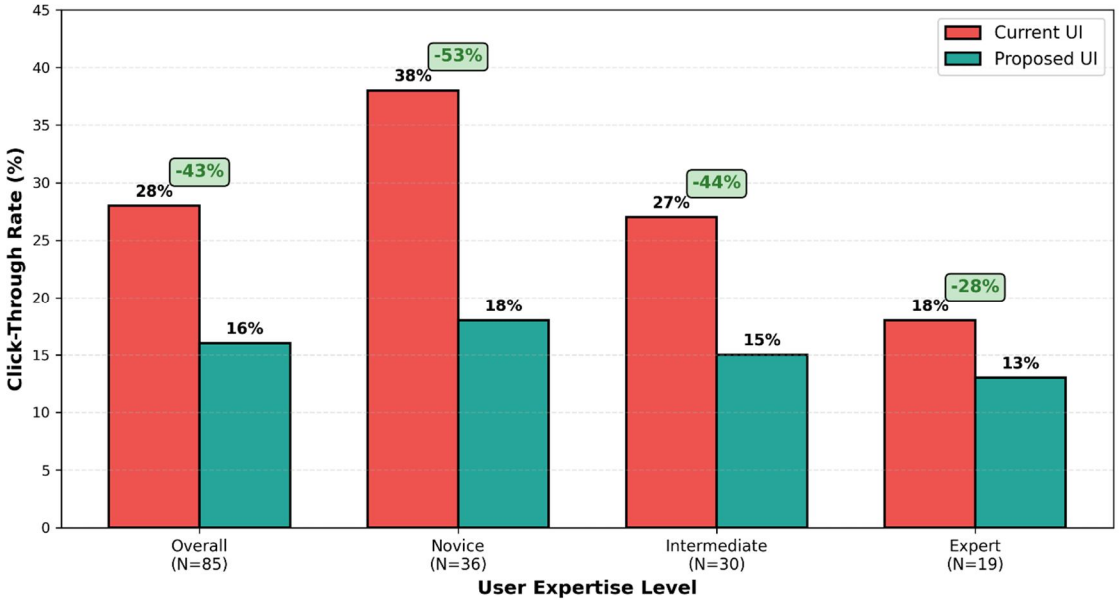


Fig. 5 Click-through reduction rates on phishing links

Secondary outcome measures showed the following. Decision time increased modestly per email, consistent with the strategic friction mechanism prompting more careful review rather than indicating confusion. Usability scores averaged well into the "Excellent" range across all participants, with slightly higher scores for novice users. The false positive rate rose marginally but not significantly; a majority of participants who encountered a false positive still rated the warning as a helpful learning experience. Decision confidence increased significantly across all groups.

D. Phase 3: Longitudinal Follow-Up

A subset of behavioral study participants used the Chrome extension in their personal Gmail accounts for seven consecutive days. Formal classification assessments were conducted at the start, midpoint, and end of the period. Detection accuracy held steady across all three measurement points, confirming the absence of a rapid habituation effect. Tooltip engagement declined slightly over the week while active use of the built-in reporting feature increased, suggesting growing user initiative rather than passive reliance on the tool. Post-study interviews indicated that several participants felt they had begun recognizing phishing indicators independently by the end of the period. Figure 6 summarizes the longitudinal results.

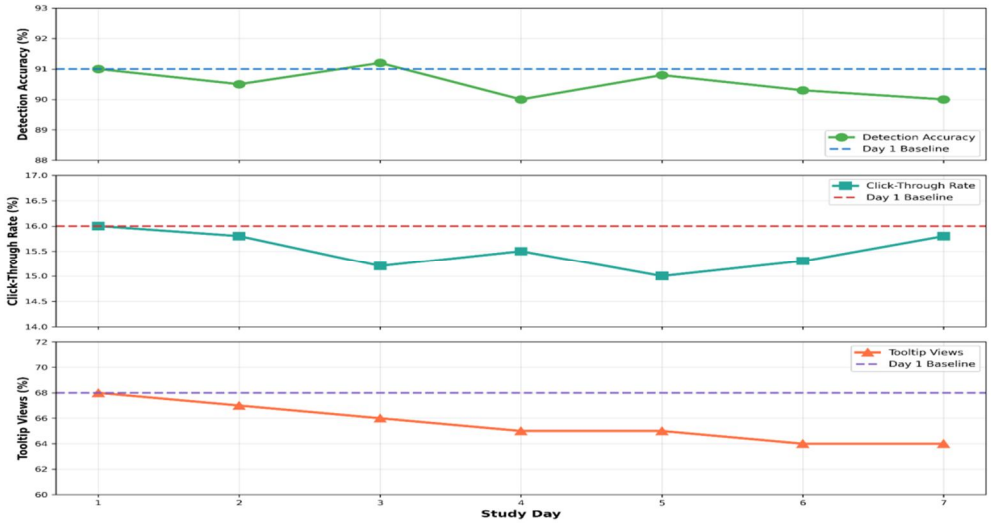


Fig. 6 Longitudinal results across the one-week follow-up. Detection accuracy and click-through rates remain stable; active reporting of suspicious emails increased over time

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

E. Comparison with Existing Approaches

The prototype achieves user-side detection accuracy substantially above rates typically associated with current interface-based defenses. Commercial email security tools commonly deliver detection rates in the 70-78% range at the interface layer, while back-end ML classifiers achieve near-perfect accuracy on the full message corpus before filtering occurs. The proposed approach occupies a distinct position: lower than back-end ML on raw accuracy, but unique in delivering behavioral validation, adaptive presentation, on-device privacy preservation, and measurable skill transfer. Figure 7 presents the comparative profile across multiple evaluation dimensions.

	Detection Accuracy	Behavioral Validation	Adaptive Design	False Positives	Privacy	Educational	Deploy
Current Browsers	65%	No	No	5%	Yes	No	Yes
PhishGuard	78%	Lab	No	12%	No	Limited	Research
Commercial Tool	70%	Sim	No	8%	Partial	Limited	Yes
XGBoost URL	99%*	No	No	N/A	No	No	Research
Our Framework	**91%**	**Yes N=85**	**Yes**	**11%**	**Yes**	**Yes**	**Yes**

Figure 7. Comparative evaluation of the proposed framework against existing anti-phishing approaches across multiple dimensions

V. DISCUSSION

The accuracy gains and click-through reduction observed in the behavioral study are large by the standards of interface-based security research. These are behavioral outcomes, not self-assessments. The discrepancy between the larger gain in perceived confidence observed in the pilot survey and the more modest gain in actual accuracy in the behavioral study illustrates the well-documented gap between subjective and objective security performance, and reinforces the case for behavioral evaluation as the appropriate standard.

The absence of habituation over one week is substantively important. Most security warning research documents rapid decline in user attention after initial exposure, sometimes within a few days. The stability observed here is likely attributable to the specificity of the warnings: because each annotation explains the particular deception tactic being used, users encounter informative content on each exposure rather than a repeated generic message. The modest increase in decision time is consistent with more deliberate processing rather than confusion, and was rated acceptable by participants across all experience levels.

From a theoretical standpoint, the study provides behavioral evidence for claims that have previously been made at the level of design principles but rarely validated empirically. The Fogg Behavior Model predicts that improving both Ability, through explanatory warnings, and Trigger, through contextual and salient alerts, should increase protective behavior. The click-through reductions observed are consistent with this prediction. The larger gains for novice users relative to experts corroborate the adaptive design rationale: stronger intervention is warranted where prior knowledge is more limited.

Several limitations qualify these findings. The sample was drawn predominantly from an educated, English-speaking, technology-familiar population; generalization to users with lower digital literacy, non-English contexts, or cognitive accessibility requirements has not been established. One week is sufficient to detect the absence of rapid habituation but does not rule out effects that might emerge over longer periods. The adversarial robustness of the detection component has not been formally tested. Future work should address these gaps through cross-cultural replication, longer deployment studies, and adversarial evaluation. The modest false positive rate also warrants attention in high-volume email environments.

VI. CONCLUSION

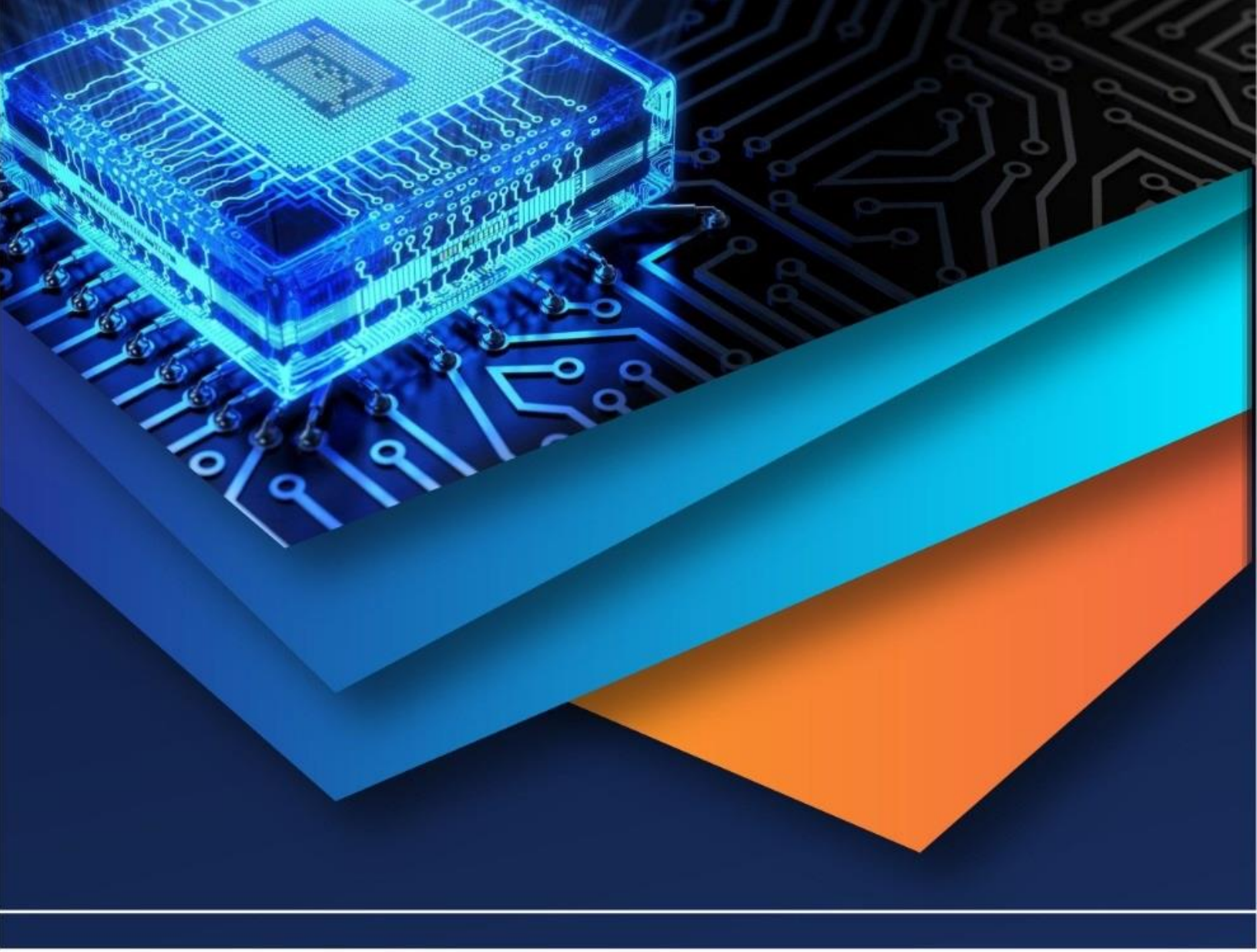
Phishing succeeds because it exploits human decision-making, and the interface is where human decision-making and security technology meet. Current interface designs fail at this juncture, but the failure is not inevitable. A framework built on three principles from HCI, persuasive technology, and behavioral science produces substantial improvements in both detection accuracy and click-through behavior, while maintaining usability at a level users find acceptable.

The evaluation results support four specific claims. Detection accuracy can be raised substantially through explanatory, context-specific warnings. Phishing link clicks can be reduced by roughly 40% through strategic friction and behavioral nudges. These effects persist over at least one week without habituation. The improvements are achievable within a deployable browser extension without compromising user privacy.

The broader implication is that interface design deserves a more prominent role in the anti-phishing research agenda. Marginal improvements in back-end detection rates are valuable, but they do not address the portion of the problem that reaches users. A well-designed interface can convert that portion from a persistent vulnerability into a genuine defensive capability.

REFERENCES

[1] [“2024 Data Breach Investigations Report | Verizon.” Accessed: Dec. 29, 2025. [Online]. Available: https://www.verizon.com/business/resources/reports/dbir.html
[2] “apwg_trends_report_q4_2023.” Accessed: Mar. 05, 2026. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf
[3] A. P. Felt et al., “Improving SSL Warnings: Comprehension and Adherence,” in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul Republic of Korea: ACM, Apr. 2015, pp. 2893–2902. doi: 10.1145/2702123.2702442.
[4] S. Egelman, L. F. Cranor, and J. Hong, “You’ve been warned: an empirical study of the effectiveness of web browser phishing warnings,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in CHI ’08. New York, NY, USA: Association for Computing Machinery, Apr. 2008, pp. 1065–1074. doi: 10.1145/1357054.1357219.
[5] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri, “Bridging the Gap in Computer Security Warnings: A Mental Model Approach”, Accessed: Dec. 29, 2025. [Online]. Available: https://www.computer.org/csdl/magazine/sp/2011/02/msp2011020018/13rRUxbCbrM
[6] R. Dhamija, J. D. Tygar, and M. Hearst, “Why phishing works,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in CHI ’06. New York, NY, USA: Association for Computing Machinery, Apr. 2006, pp. 581–590. doi: 10.1145/1124772.1124861.
[7] “2022_ic3report.pdf.” Accessed: Mar. 05, 2026. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2022_ic3report.pdf
[8] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, “Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model,” Decis Support Syst, vol. 51, no. 3, pp. 576–586, Jun. 2011, doi: 10.1016/j.dss.2011.03.002.
[9] R. H. Thaler and C. R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness. in Nudge: Improving decisions about health, wealth, and happiness. New Haven, CT, US: Yale University Press, 2008, pp. x, 293.
[10] R. W. Rogers, “A Protection Motivation Theory of Fear Appeals and Attitude Change1,” J. Psychol., vol. 91, no. 1, pp. 93–114, Sep. 1975, doi: 10.1080/00223980.1975.9915803.
[11] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, “Teaching Johnny not to fall for phish,” ACM Trans Internet Technol, vol. 10, no. 2, p. 7:1-7:31, Jun. 2010, doi: 10.1145/1754393.1754396.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY