The Impact of LLM Assistance on User Spam Detection

Alexis Morales Flores
University of California San Diego
Computer Science and Engineering
La Jolla, CA, USA
amoralesflores@ucsd.edu

Jay Jhaveri
University of California San Diego
Computer Science and Engineering
La Jolla, CA, USA
jjhaveri@ucsd.edu

Aditya Sharma
University of California San Diego
Computer Science and Engineering
La Jolla, CA, USA
arsharma@ucsd.edu

Sarthak Rege
Independent Researcher
USA
sarthakrege@gmail.com

Imani Munyaka
University of California San Diego
Computer Science and Engineering
La Jolla, CA, USA
imunyaka@ucsd.edu

Abstract

Prior research has extensively examined how everyday users detect spam emails, revealing a consistent need for additional support through features such as automatic detection, secure email tools, and labeling mechanisms. However, with the emergence of large language models (LLMs), it remains unclear how these tools impact users' decision-making in the context of spam detection. In this study, we investigate the role of LLMs as decision-support tools and their impact on users' ability to identify spam. We conduct a user study in which participants (N = 295) respond to a total of four emails—two spam and two legitimate (ham) messages. Our findings suggest that prior educational efforts around spam awareness may have positively influenced user decision-making. However, the results also highlight users' tendency to cognitively offload the task of spam detection onto external tools like LLMs, underscoring the continued need for contextual support to reinforce accurate detection and reduce user vulnerability.

CCS Concepts

• **Security and privacy** → **Spam detection**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → *Natural language processing*.

Keywords

spam detection, large language models, human-AI collaboration, cybersecurity, email security, cognitive offloading, human-computer interaction

ACM Reference Format:
Alexis Morales Flores, Jay Jhaveri, Aditya Sharma, Sarthak Rege, and Imani Munyaka. 2025. The Impact of LLM Assistance on User Spam Detection. In *Proceedings of the 2025 Workshop on Human-Centered AI Privacy and Security (HAIPS '25), October 13–17, 2025, Taipei, Taiwan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3733816.3760754

HAIPS '25, Taipei, Taiwan

ACM ISBN 979-8-4007-1905-9/2025/10
https://doi.org/10.1145/3733816.3760754

1 Introduction

Traditional spam detection systems primarily rely on machine learning techniques, such as convolutional neural networks (CNNs), along with rule-based filters to achieve high classification accuracy in automated settings [12] [44]. These systems support spam detection through both back-end classification and front-end labeling of email characteristics. However, users still play a critical role in identifying spam that permeates current spam filters.

As adversaries continue to improve their tactics and user expectations shift, users' ability to accurately detect spam may be compromised. Attackers frequently exploit user behavior and find new ways to bypass existing security measures [1] [8] [48]. To support users, prior research has explored training interventions and system design approaches aimed at improving users' spam detection capabilities [37]. Building on this work, we investigate the potential of LLMs as a tool to assist users in identifying spam.

Recent advancements in natural language processing have positioned LLMs not only as capable classifiers but also as interactive assistants that can support human decision-making [15]. While most existing research emphasizes the technical performance of automated spam detectors, it is equally important to understand how users interact with LLMs in real-world security contexts. Work in Human-AI Teaming suggests that AI assistance can enhance user decision-making and performance [6]. However, recent findings also raise concerns about cognitive offloading where users rely too heavily on tools like ChatGPT and reduce their own critical thinking [17]. This over-reliance can lead to poor outcomes when users are faced with situations where the tool is unavailable or provides incorrect information [18].

These findings motivate our exploration of how users are affected by LLM support when performing security tasks that require critical thinking. We focus specifically on spam detection because it is a task that is both familiar to users and cognitively demanding.

We examine the use of LLMs as assistive tools in the context of spam detection, conceptualizing them as similar to virtual tutors that can help users reason through a decision. Our goal is to understand how the presence of such a tool affects users' detection strategies and overall accuracy. To explore this, we conducted a survey study with 295 participants. Each participant reviewed four email samples and judged whether they were spam or legitimate. Participants were divided into two groups: a control group (n=155) that reviewed the emails independently, and an experimental group (n=140) that had access to a chatbot assistant during the task. Our results indicate that using the chatbot tool impacted how the user evaluated the system.

We contribute the following:

(1) We find that chatbot assistance may not improve user email classification for knowledgeable and confident users.
(2) We show that even though users have confidence in their abilities to detect spam email, they may still lean into the chatbot to classify the emails and verify their thoughts.
(3) We demonstrate that using a chatbot assistant for spam detection may lead to cognitive offloading which can cause a decrease in critical thinking.

Outline: This paper is organized as follows. We give an overview of related work in Section II describes the related work. Section III is an overview of our methodology. Our results and limitations are presented in Section IV. We discuss our findings and provide solutions in Section V. We conclude in Section VI.

2 Related Work

Despite significant advances in automated email security systems, the end user ability to detect email threat remains a weak point, particularly as attacks become more sophisticated[40][14]. While extensive research exists on both automated spam detection algorithms and human factors in spam detection, limited work has examined how LLM interactive AI assistants like ChatGPT influence user behavior in email security tasks. This section reviews literature in email security, human-computer interaction in cybersecurity, and AI-assisted decision making that informs our investigation.

2.1 Risk factors and vulnerable populations

Research on susceptibility to phishing and spam emails has been conducted extensively, however results vary on which demographics are vulnerable. While some studies find that there are no significant differences between ages, others have found that young adults are more susceptible whereas others observe greater susceptibility among st older adults [16, 20, 38, 39]. Though age may be a factor that has yielded different results, it is clear that spam and phishing tactics not only target software vulnerabilities but also use psychological pressures to elicit an emotional response from their potential victims which could cause them to act impulsively. Studies have shown that successful spam and phishing attacks mostly rely on creating a sense of urgency in the victims in order for them to make rash choices without critical evaluation of the email [13, 45, 48]. Other research has indicated that certain personality traits may increase the likelihood of falling for spam and phishing attacks. User studies indicate that participants with lower impulsivity scores were better at identifying emails and also being more suspicious of links [9] [33].

Additionally, some findings indicate that time constraints can also significantly impact a user's ability to take the necessary time to evaluate an email. One study found that limiting the time participants had to evaluate a spam email negatively impacted their accuracy scores which dropped to nearly 50% accuracy in detecting spam [10]. Time constraints are an important factor in email classification because in professional settings where users are multitasking and working to meet deadlines, they will be unable to take the necessary time to evaluate each email carefully. One study indicates that high volumes of emails received and habits users develop to process those emails can increase their likelihood of falling for a phishing attack [45].

2.2 Backend Spam Email Detection Techniques

Current countermeasures against spam and phishing rely on a combination of techniques, such as Bayesian filtering, rule-based scoring, and reputation checks, to analyze emails on the server and prevent spam from reaching the user's inbox [5]. According to Google, their spam filtering techniques can prevent over 99.9% of spam, phishing, and malware from reaching a user's inbox [19]. While reasonably secure, research indicates that messages generated using LLMs can bypass some of the current spam filters. One study found that using LLMs (ChatGPT-4o) to generate spam emails with varying levels of sophistication, were able to evade detection by filters among popular email service providers for a relatively low cost per generated email [31]. This would indicate that AI-generated emails may bypass safeguards at a higher rate than human generated spam. This problem may be exacerbated further in the future as LLMs become even more prominent and criminals may begin to look for alternative methods to craft spam and phishing emails using LLMs.

2.3 LLMs for Spam and Phishing Detection

With the increased use of LLMs, many new methods of aligning and or training for specific use cases have been developed. Research into the use of LLMs for phishing and spam detection has shown that a fine-tuned BERT model can classify spam and phishing emails more accurately than baseline models [23]. Other research has shown similar results against traditional methods of spam filtering. Prior work has explored the comparative performance of fine-tuned LLMs against CNN-based methods, indicating that LLMs can outperform traditional approaches in distinguishing spam from legitimate emails [36].

2.4 Untrained LLMs and Prompt-based Approaches

However, the technical knowledge and computational resources that are required to fine-tune models may not be common amongst average users. Therefore free-tier LLMs available online may offer accessible untrained LLMs as a tool to the general public which does not require any of the aforementioned expertise or resources. Recent research has examined zero-shot and prompt-based approaches, which enable LLMs to perform specific tasks without additional training. Researchers [35] have utilized LLMs to summarize and classify emails as spam or not through zero-shot prompting, with results indicating high effectiveness in distinguishing spam content from legitimate emails. Another study [22] found that carefully crafted prompts enabled LLMs to perform tasks like toxicity classification without retraining. This method of leveraging LLMs' inherent language understanding supports the notion that untrained LLMs, when effectively prompted, may help users identify spam emails in real-world settings.

2.5 Human-AI Teaming for Cybersecurity

While prior research has extensively examined technical approaches to spam detection, there remains a significant gap in understanding how AI assistants, particularly LLM influence human decision making in spam detection and cybersecurity contexts. Researchers have explored Human-AI teaming for cybersecurity education, cyber operations, military needs, and training [26] [27] [46] [47] [30]. Prior work on human-AI collaboration for spam detection found that users over-relied on ML models that exhibited high accuracy, even agreeing with the model when it made errors [42]. Their findings showed that providing explanations for model decisions offered no improvement in user accuracy and in some cases, appeared to increase compliance with incorrect predictions. Tariq et al also explored the use of LLMs for phishing and intrusion detection with academics, professionals, and computer science students [43]. The results of that study suggest that LLM response characteristics directly impact prompting strategies, decision revision and user trust.

We further this area of exploration by comparing control and experimental groups to quantify the impact of AI assistance. We include an interactive model to investigate how interacting directly via chatbot (conversational agent built on top of LLM) shapes human reasoning processes in cybersecurity contexts. We extend this area of exploration through our study design and participant recruitment, focusing on users outside traditional STEM contexts to include a diverse range of technical backgrounds. We compare control and experimental groups to quantify the impact of AI assistance in order to investigate how direct interaction with a chatbot shapes human reasoning in cybersecurity contexts.

To address these gaps, this study poses the following research questions:

- RQ1: How does the presence of a chatbot assistant influence users' perceptions of email legitimacy?
- RQ2: How does the use of a chatbot tool impact user's detection strategies?
- RQ3: How does the use of a chatbot tool impact user confidence and risk perception?

3 Methods

We recruited a total of 300 participants from Prolific who were from the United States of America to support our effort to obtain a U.S. representative sample. However, after reviewing responses, we removed 2 participants with incomplete submissions and 3 whose responses to open ended questions were incoherent. After reviewing user submissions, we retained data from a total of 295 participants. Participants were tasked with visiting our webpage and evaluating a set of four emails. In this section we detail the study process and participant demographics.

3.1 Study Design

We developed a web application that participants would use to interact with four emails (See Figure 1). Once logged in, each participant was asked to review the consent form, complete a pre-survey, read study instructions, and then complete the study task. The study task required participants to review a set of emails and evaluate them. Participants were given 5 minutes to engage with an email within the provided interface and then answer a short email evaluation survey.

Each participant was assigned to either the control or experimental group. While both groups completed the same task, the experimental group was asked to use a chatbot, which was embeded into the email interface, to complete the task. The chatbot was a conversational AI system build on a large language model. Once all emails and surveys were completed, they would proceed to the final survey and the experiment would conclude. Each email was presented in a mock Gmail graphical user interface (GUI) to give the user a semi-immersive experience as if they were viewing this email on their personal Gmail account. We opted to use the Gmail interface as it is one of the most common email service providers.

The post survey collected participants' self-reported trust, perceived learning, and prior experience with ChatGPT. We placed the question about prior ChatGPT usage in the post-survey rather than the pre-survey to avoid priming participants or influencing their behavior during the email classification task. Asking this question beforehand could have potentially caused participants in the experimental group to suspect they would be interacting with ChatGPT and impact their interactions with the chatbot.

3.2 Chatbot Design and Capabilities

Our chatbot was powered by the GPT-3.5-turbo model via the OpenAI API and was not a Wizard of Oz simulation. During the study, we referred to the tool as a chatbot to participants and intentionally avoided disclosing that it was powered by GPT-3.5-turbo to prevent introducing potential bias from pre-existing perceptions of ChatGPT specifically. This choice was made in order for participants to evaluate the chatbot solely on its performance during the task. We opted to use GPT3.5-turbo for this task because we wanted to use the same tool that was currently freely available to everyday users. At the time of study, GPT3.5-turbo was the free-tier ChatGPT version that was available by OpenAI.

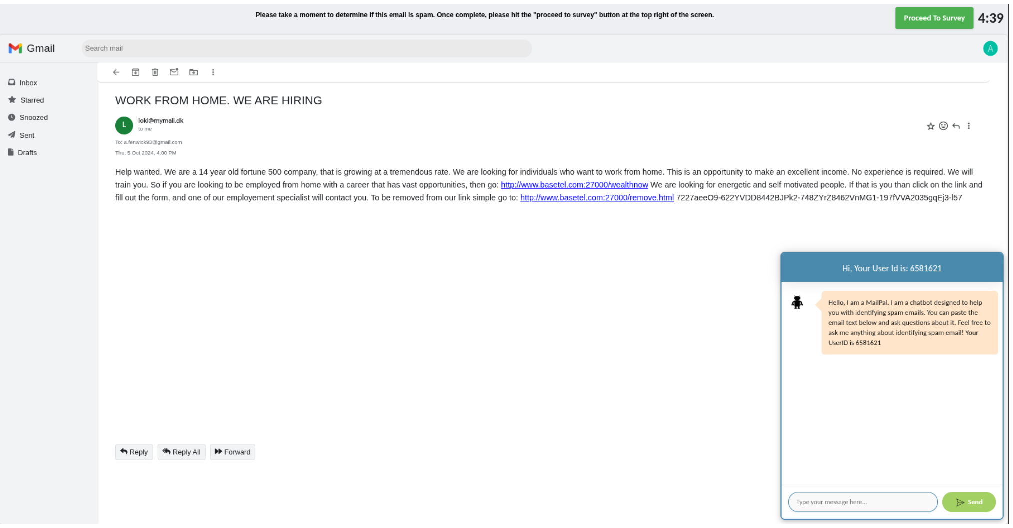
The chatbot's interface was presented as a small window embedded in the bottom right corner of the email page, visually similar to commercial e-commerce assistants. The purpose of this design was to keep the chatbot visible but unobtrusive while allowing participants to view the entire email. The chatbot did not automatically analyze and flag the email because we wanted to focus on how participants naturally interact with a chatbot when requesting assisting with email evaluation. Providing a different design would have shifted the focus toward evaluating a specialized spam detection system, which was outside the scope of this study.

3.3 Email Selection

The four emails used in this study were purposefully selected to represent distinct classification challenges in real-world email filtering. Emails were obtained from three sources: the publicly available SpamAssassin 2004 corpus [4], the Nazario phishing dataset [3], and legitimate ham emails shared by researchers in our lab. The initial dataset consisted of emails categorized as easy ham, hard ham, and spam. Each email was parsed, and its content evaluated

Table 1: Email Classification Summary with Spam/Ham Indicators

Codename	Ground Truth	ChatGPT Label	Key Indicators and Cues
Netflix	Spam	Spam	Spoofed brand identity; urgent language ("We're having some trouble..."); generic greeting; request to verify account; embedded hyperlink not matching official domain.
Amazon	Ham	Spam	Legitimate sender address; transactional and formal tone; no embedded hyperlinks; standard customer service phrasing; lacks deception markers.
Flooded Bathroom	Ham	Ham	Personal narrative; informal tone; no commercial or security language; no links; sender known (academic affiliation); realistic context.
Remote Work	Spam	Ham	Vague job offer; unrealistic income claims; lack of identifiable company name; informal grammar; non-corporate sender address; absence of personalization.



Please take a moment to determine if this email is spam. Once complete, please hit the "proceed to survey" button at the top right of the screen.

WORK FROM HOME. WE ARE HIRING

loki@mymail.dk
to me
To: a.fenwick93@gmail.com
Thu, 5 Oct 2024, 4:00 PM

Help wanted. We are a 14 year old fortune 500 company, that is growing at a tremendous rate. We are looking for individuals who want to work from home. This is an opportunity to make an excellent income. No experience is required. We will train you. So if you are looking to be employed from home with a career that has vast opportunities, then go: http://www.basetel.com:27000/wealthnow We are looking for energetic and self motivated people. If that is you than click on the link and fill out the form, and one of our employement specialist will contact you. To be removed from our link simple go to: http://www.basetel.com:27000/remove.html 7227aeeO9-622YVDD8442BJPk2-748ZYrZ8462VnMG1-197fVVA2035gqEj3-l57

Hi, Your User Id is: 6581621

Hello, I am a MailPal. I am a chatbot designed to help you with identifying spam emails. You can paste the email text below and ask questions about it. Feel free to ask me anything about identifying spam email! Your UserID is 6581621

Figure 1: This is the email interface shown to each participant. The experimental group was shown the interface with the chatbot embedded into it.

using the GPT-3.5-turbo model to yield binary labels: *spam* or *ham*. Our objective was to identify two ham and two spam emails that demonstrated both successful and unsuccessful classifications by the model.

Accordingly, we selected four emails representing four critical scenarios: (1) a phishing email correctly identified as spam, (2) a spam email incorrectly labeled as ham, (3) a legitimate email misclassified as spam, and (4) a legitimate email correctly classified as ham. This allowed us to examine both the strengths and limitations of the model's classifications and explore their impact on human judgment during downstream user interaction. For the remainder of this paper, we use the term *spam* to encompass all emails deemed malicious, deceptive, or unwanted.

Each participant evaluated all four selected emails, which were rendered in a simulated Gmail interface to enhance ecological validity. Each participant evaluated all four selected emails, which were presented in randomized order to minimize potential ordering effects. The emails differed in sender, subject, and body content, and were counterbalanced across participants to distribute any scenario-specific effects evenly. We adopted this within-subjects approach to enable direct comparisons across conditions for the same participant while keeping the required sample size manageable. A between-subjects design, where each participant reviewed only one email, would have required substantially more participants to achieve equivalent statistical power and would have introduced greater variability from individual differences.

Below, we provide our reasoning for the selection of each email, along with a summary table of its linguistic and behavioral indicators.

Netflix (Spam - Correctly Classified). This email contains several characteristics common in spam/phishing emails such as: brand spoofing, urgent messaging, malicious URLs, and non-personal greetings (i.e. Dear customer). Such indicators are widely acknowledged in phishing detection literature [48] [9] [13] [45].

Amazon (Ham - Misclassified as Spam). Although this email is a legitimate communication from Amazon, it requests the receiver to reply with the type of card as well as the last 2 digits of the card in order to process the payment. The email's overlap with phishing templates and account-related language may trigger a false-positive in back-end spam filters, evaluations have been noted in prior work [31]. ChatGPT gave the following reason for its misclassification.

> "This email is likely a phishing attempt because it asks you to reply with credit card details, which Amazon never does, and it uses suspicious email addresses and fear tactics to prompt immediate action. Additionally, it lacks personalization and contains subtle formatting issues typical of phishing scams."

[32]

Flooded Bathroom (Ham - Correctly Classified). This email contains an informal narrative-style message with no commercial or urgent content. Research shows that users and classifiers alike are better at identifying such personal messages as non-threatening [25] [24].

Remote Work (Spam - Misclassified as Ham). This is a work-from-home spam email which contains a fear-of-missing-out component to create interest and urgency in receivers. It contains unrealistic income claims, unspecified work duties or skill requirements. ChatGPT's reason for incorrectly classifying this email as ham was the following:

> "The email uses a professional tone, includes a removal link, and resembles a typical job advertisement. It lacks urgent language, threats, or emotional manipulation commonly used in phishing."

[32]

3.4 Participants

We used the crowdsource platform 'Prolific' to recruit participants for this study. A total of 295 participants completed our study with most participants being female (51%), under 65 years in age (88%), and white (62.03%). In comparison, most people in the United States are female (51%), between 18 and 65 (55.1%), and white (75.3%). Pre-survey analysis for the experimental group's response to the questions "Have you ever interacted with ChatGPT?" and "Have you ever interacted with a chatbot?" indicated that the overwhelming amount of users had used ChatGPT before (96.4%) or another LLM chatbot (99.3%). Most participants (63%) did not have experience having clicked on or replied to spam emails. Similarly, a majority of participants (78%) also expressed not having experienced consequences from responding to a spam email.

Although recruitment was initially balanced between the control and experimental groups, the experimental group had less participants who completed the study, leading to a smaller final sample in that group. Furthermore, there were more experimental group responses which were not coherent and as a result were removed from the data.

3.5 Ethics

Our study was approved by our internal review board. The consent form included information about our data collection methods and participant rights. Each survey response, and chatbot interaction was logged to collect the data necessary for this study. For simplicity and safety, we specifically selected emails that did not contain controversial statements. We also changed the final destination of all embedded links to prevent negative outcomes for participants.

4 Results

This section presents the results from our study examining how ChatGPT assistance impacts users during email spam detection tasks. We analyze participants' classification accuracy and strategies.

4.1 Email Classification Accuracy

Participants were asked to select all labels they believed accurately described the email they reviewed. Participant responses were determined to be accurate as long as they selected at least one correct option. The chi-square test of independence was used to determine if there was a relationship between email classification accuracy and chatbot assistance. The results suggest that there is no relationship ($\chi^2 = 0.45$, $p = 0.502$) between chatbot assistance and email classification accuracy. The accuracy rates between the control and experimental groups were 72.5% and 70.8%, respectively. Thus, receiving assistance from the chatbot tool did not improve overall classification performance (Table 2). Participants in both groups achieved high classification accuracy on the Netflix spam email, remote work spam email, and the flooded bathroom ham email. However, for the Amazon ham email, the participants in both groups were less likely to correctly identify this as legitimate.

Table 2: Email Classification Accuracy by Group

Email Type	Control (%)	Exp. (%)	χ^2	p-value
Netflix (Phishing)	96.4	96.5	0.00	1.000
Amazon (Legitimate)	28.2	19.4	3.25	0.071
Flooded Bathroom (Legit.)	58.5	62.2	0.37	0.541
Remote Work (Spam)	95.5	96.8	0.16	0.686

4.2 Response to the Chatbot

Chatbot Usage: Our instructions asked the experimental group participants to engage with the chatbot to classify the emails. There was no penalty for not engaging and no reward for engaging. We then reviewed the chatbot conversations to determine if they used the chatbot and how they used it (classifier or verifier). If the participant explicitly asked the chatbot to classify the email, we marked that as classifier usage. If the participant used the chatbot to verify information or answer a question about the email that did not explicitly ask the chatbot to classify the email, we marked the conversation as verifier usage. Only 1.45% of the participants (n= 2 out of 140) chose not to use the chatbot in the study. The remaining participants (n=138) used the chatbot as a classifier for each email. That is, they used the chatbot by providing it with the email and asking it to classify the email for them. However, it is important to note that some participants used the tool as a classifier to verify their own beliefs. For example, after receiving a response from the chatbot P5832 said the following:

> "That is what I thought based on the odd spacing and the email address. Thank you very much for helping me confirm it."

Agreement with Chatbot Responses: Participants in the experimental group were asked to select the level to which they "agree with the answers the chatbot gave to my questions" to gauge agreement levels with the chatbot's analysis of the email. Participants agreed with the chatbot assessments across all emails, even the emails misclassified by the chatbot. Agreement levels were 84.6% when combining 'agree' and 'strongly-agree' responses and reached statistical significance (χ^2 = 664.98, p < 0.001)

Chatbot Accuracy: All participants were asked to select the level to which they agree that they "trust that the chatbot will give me accurate and reliable responses". Table 3 results from the Chi-square tests of independence show that the experimental was more likely to trust that the chatbot would provide accurate and reliable responses (Experimental: 86.4% vs. Control: 54.8%, χ^2 = 44.76, p < 0.001).

Chatbot Adoption: All participants were asked to select the level to which they "would use a chatbot to help me detect spam email". Chi-square tests of independence from Table 3 revealed that willingness to use chatbots for spam detection was significantly higher in the experimental group (Experimental: 77.1% vs. Control: 47.7% agreement, χ^2 = 37.93, p < 0.001).

Table 3: Post-Experiment Attitudes Toward Chatbot Usage by Group

Attitude	Control (%)	Exp. (%)	χ^2	p-value
Willingness to use chatbot	47.7	77.1	37.93	< 0.001
Trust in chatbot accuracy	54.8	86.4	44.76	< 0.001

4.3 Participant Attributes

Confidence: Both groups showed similar levels of confidence in their own spam detection abilities across multiple measures. Before and after the email task, participants were asked to select the degree to which they agree that they are "confident in their ability to identify spam emails". Results from the Wilcoxon signed-rank tests show that neither the control group (p = 0.291) nor the experimental group (p = 0.219) had statistically significant changes in confidence levels. Thus, most participants were confident in their ability to identify spam emails.

Table 4: Chatbot Agreement Rates by Email (Experimental Group, n = 140 per email)

Email Type	Agreement (%)	χ^2	p-value
Netflix (Phishing)	82.1	150.57	< 0.001
Amazon (Legitimate)	86.4	142.21	< 0.001
Flooded Bathroom (Legitimate)	81.4	104.29	< 0.001
Remote Work (Spam)	88.6	161.43	< 0.001

Participants were also asked to rate their agreement with the idea that they "would never fall for spam" and "can usually tell the difference between legitimate email and a phishing or spam email". Chi-square tests of independence showed no significant differences in participants' confidence in distinguishing legitimate from spam emails (Experimental: 81.4% vs. Control: 85.2%, χ^2 = 1.65, p = 0.800), beliefs about never falling for spam (Experimental: 57.1% vs. Control: 48.4%, χ^2 = 4.09, p = 0.394), or overall detection confidence (Experimental: 83.6% vs. Control: 83.2%, χ^2 = 1.19, p = 0.880). These results can be seen in Table 6.

Risk Perception: Participants were asked to rate the level of risk associated with opening a spam email and clicking a link in a spam email. Chi-square tests of independence revealed that the experimental group rated opening spam emails as riskier more frequently than the control group (Experimental: 71.4% vs. Control: 59.4% rating it as very/extremely risky, χ^2 = 10.15, p = 0.038). However, both groups showed equally high risk awareness regarding clicking spam links (Experimental: 92.1% vs. Control: 93.5%, χ^2 = 5.28, p = 0.259) and a similar level of agreement about likelihood that a spam email could lead to severe consequence (Experimental: 95.0% vs. Control: 95.5%, χ^2 = 3.86, p = 0.425).

Table 5: Risk Perception of Spam Emails by Group

Risk Factor	Control (%)	Exp. (%)	χ^2	p-value
Opening spam (very/extremely risky)	59.4	71.4	10.15	0.038
Clicking spam links (very/extremely risky)	93.5	92.1	5.28	0.259
Serious consequences (likely/very likely)	95.5	95.0	3.86	0.425

Table 6: Self-Confidence in Spam Detection and Past Experience by Group

Measure	Control (%)	Exp. (%)	χ^2	p-value
Self-Confidence				
Can distinguish legitimate from spam	85.2	81.4	1.65	0.800
Would never fall for spam	48.4	57.1	4.09	0.394
Confident in detection ability	83.2	83.6	1.19	0.880
Past Experience				
Clicked/replied to spam	41.3	32.9	2.50	0.287
Experienced spam consequences	23.9	19.3	= 1.42	0.491

4.4 Self-Reported Learning

The experimental group was more likely to report learning something new about spam detection in each of the email-survey questions, suggesting that the chatbot tool helped participants determine which email characteristics they needed to review in their analysis. Self-reported learning rates were significantly higher for experimental participants across all four emails. When combining responses across all four emails, participants in the experimental group were significantly more likely to report learning something new (37.3%) compared to those in the control group (19.2%), as shown in Table 7. This difference was statistically significant, $\chi^2(1) = 47.28$, $p < 0.001$.

Table 7: Self-Reported Learning from Each Email (by Group)

Email Type	Control (%)	Exp. (%)	χ^2	p-value
Netflix (Phishing)	19.4	32.1	5.69	0.017
Amazon (Legitimate)	26.5	42.1	7.40	0.007
Flooded Bathroom (Legitimate)	12.9	32.1	14.75	< 0.001
Remote Work (Spam)	18.1	42.9	20.43	< 0.001

4.5 Qualitative Analysis of Participant Responses

Participants were asked if they learned anything new when completing the task of analyzing the four emails. We conducted content analysis to highlight the new strategies participants said they learned. To do this, one researcher reviewed 10% of the responses and created a codebook. A second research then coded an additional 10% to update the codebook. Then, both researchers coded the remaining 80% of the responses and discussed any code conflicts. The codes were then combined to create themes that represent email analysis strategies. The strategies mentioned include checking the sender email, providing links, email content, and any request for information that can be used for fraud.

4.5.1 *Check the Sender and Links.*

Some participants noted that they learned to review the sender's email address and check for odd or unusual domains in the link. After reviewing the Netflix spam email, a participant stated:

> "[This experience reinforced] the importance of scrutinizing the sender's email address... and suspicious-looking links. The unusual domain name is a particularly strong red flag in this case." (A3927, Netflix)

> "This interaction highlights that not all unexpected emails are spam. Personal anecdotes and emails from seemingly legitimate, albeit perhaps unfamiliar, email addresses are less likely to be malicious. The content and sender's information are crucial context for evaluation." (D5649, Flooded Bathroom)

4.5.2 *Scrutinize Email Content.*

Other participants highlighted reviewing emails for an urgent tone, grammatical errors, generic message or greeting, false offers, or false information.

> "I learned that legitimate companies rarely...send unprofessional mass emails with grammatical mistakes." (L2078,Remote Work)

> "I learnt that any website that I am a customer with, when sending me an email will always address me by my first Name." (N3168, Netflix)

4.5.3 *Next Steps.*

In response to receiving a suspicious email, some participants expressed that the experience reinforced the importance of reacting in specific ways. This included contacting the company in the email directly, not clicking on a link, and remaining skeptical.

> "This email highlights the deceptive tactic of using a sender address ... It reinforces the importance of always verifying such requests through official channels and never providing sensitive information via email." (R5610, Amazon)

> "This email reinforces the importance of being wary of job offers that seem too good to be true, especially those coming from generic email addresses and lacking specific company details." (E6289, Remote Work)

4.6 Control vs. Experimental Reasoning

We asked participants to explain why they determined an email to be spam or ham. After we used the code book to categorize self-reported learning, we then used the code book to categorize user reasoning. We grouped the codes into five categories but only focused on four. The five categories included the link, sender email address, context of the email (urgent,friendly, etc), request for personal information, and other. We then turned the categories into counts. The goal was to determine if there were any characteristics that the experimental group mentioned that were not mentioned by the control group. The results suggest that in general, the same characteristics were mentioned by participants in both groups. However, differences emerge when we compare group percentages by email. We conduct a statistical z-test of proportions to evaluate the outcomes. We find there is no significant difference in the characteristics participants identified as their reasoning when labeling the ham emails, and the remote work email. However, we find that the control group was more likely to mention the sender's email address and the link in the email, but less likely to mention the content in the email and the request for personal information when providing their reasoning. These results suggest that having the chatbot as support may impact what users focus on. The results can be viewed in Table 8.

4.7 Limitations

Several limitations should be considered when interpreting these findings.

Email Selection: We specifically chose spam emails that use traditional spam techniques to help us focus on the end user interaction with the chatbot tool. More sophisticated or novel attack vectors might yield different results regarding both performance improvements and learning outcomes. The deliberate selection of older-style spam emails, while methodologically sound for isolating chatbot effects, may limit generalizability to contemporary threats.

Table 8: Participant Reasoning Categories by Email Type and Group

Reasoning Category	Amazon (Legitimate)		Netflix (Phishing)		Flooded (Legitimate)		Remote Work (Spam)	
	Exp. (%)	Control (%)	Exp. (%)	Control (%)	Exp. (%)	Control (%)	Exp. (%)	Control (%)
Sender	1.43	5.81	45.00	89.68	3.57	1.29	5.00	19.35
Request	97.14	96.77	38.57	11.61	2.14	4.52	17.14	2.58
Content	2.86	1.94	62.86	20.00	95.71	97.42	94.29	96.13
Link	0.71	3.87	32.14	76.13	4.29	4.52	19.29	17.42

Note: Exp. = Experimental group, Control = Control group. Values represent the percentage of participants who cited each reasoning category for each email.

Generalizability: The recruited participant population through Prolific does not match U.S demographics exactly and includes a population with relatively high technological proficiency. The high accuracy rates for email classification shown in Table 2 suggest our sample may not be representative of less technologically experienced populations. Furthermore, 96.4% of participants in the experimental group had previous experience using ChatGPT while 99.3% had utilized another type of LLM. Prior experience with ChatGPT or similar LLMs may impact our results particularly in terms of willingness to use a chatbot, agreement with chatbot responses, and perceived accuracy of chatbot due to pre-existing biases. Though this is a byproduct of having a participant sample that is more technologically proficient, our results may yield different outcomes on each of the aforementioned measures with a pool of participants that do not have experience using LLMs.

Learning: The self-reported learning measures may not accurately reflect actual knowledge acquisition. Participants' reports of learning new detection strategies could represent recall or reinforcement of existing knowledge rather than genuine skill development.

Interface Design: Our choice to present the chatbot in a separate panel rather than integrating it directly into the email interface helped us observe natural interactions with a general-purpose chatbot. However, this design does not appear in real-world spam detection systems, therefore, our findings reflect participant interactions with an embedded tool without automatic spam detection.

Study Design: To ensure each participant was exposed to all four conditions (spam or legitimate and with or without chatbot assistance), we used different email scenarios across each of condition. Our within-subjects design allowed us to compare each participant's performance across all conditions, using different email scenarios also introduces the possibility that scenario-specific factors influenced judgments beyond the experimental manipulation. It is possible that future work might implement a between-subjects design and gain different results.

5 Discussion

RQ1: How does the presence of a chatbot assistant influence users' perceptions of email legitimacy? The findings of our work indicate that having the chatbot as a tool does not impact user perception. This also suggests that over time, users have become more skeptical and potentially more skilled in their ability to detect spam emails.

RQ2: How does the use of a chatbot tool impact user's detection strategies? The findings of our work indicate that having the chatbot as a tool does impact user detection strategies. It is unknown whether participants truly learned something new. However, user feedback suggests that having a tool may have 1) helped participants recall what characteristics to consider and review when evaluating email or 2) prompted participants to release the cognitive load of the task to the chatbot.

RQ3: How does the use of a chatbot tool impact user confidence and risk perception? The findings of our work indicate that having the chatbot as a tool does not impact user confidence but might impact risk perception. The experimental group's heightened risk perception regarding opening spam emails (71.4% vs. 59.4% rating as very/extremely risky) suggests that chatbot interaction may increase in-time awareness of email-based threats such that simply opening a problematic email is deemed risky. However, the lack of significant differences in clicking spam links and consequence severity perception indicates that both groups already possessed strong awareness of critical email security risks.

5.1 Reinforcement Learning

Experimental participants consistently reported learning significantly more about spam detection across all emails (table 7). However, self-reported learning did not result in better classification reasoning or accuracy. This suggests that the reported learning may represent recall or reinforcement of existing knowledge rather than acquisition of new skills. However, participants may have viewed this type of experience in the same way.

5.2 LLMs as a Resource

Our findings suggest that chatbot-assisted learning tools have the potential to support users before or after cybersecurity trainings. While performance improvements were not observed in our study, the participants reported learning and indicated that they would use a chatbot again for this task. This suggests that, if introduced to everyday users, such tools could be valuable for knowledge reinforcement and explanation of security concepts in novel ways that users will be willing to engage with.

The absence of significant differences in classification accuracy (table 2) suggests that current spam and phishing detection capabilities among users may already be reasonably effective for the types of emails tested in this study. This finding could indicate that existing cybersecurity awareness efforts are working for the type of spam and phishing attempts shown. Thus, the consistent confidence, learning reports, and lack of improvement despite chatbot assistance raises important questions about over-reliance on automated tools.

5.3 Implications for Human-AI Collaboration in Cybersecurity

Since there is no significant difference between how the two groups classified all emails and the characteristics they used to classify the emails was similar, it is unlikely that learning occurred. Instead, our results likely demonstrate cognitive offloading. Cognitive offloading is the act of delegating cognitive tasks to an external resource to reduce one's cognitive load [34]. When offloading, the offloader depends on the external resource to do the task. For example, a person may write a list before going grocery shopping to reduce their cognitive load while at the store.

In our study, it is likely participants were offloading the task of reviewing the email to the chatbot. We believe this happened because users reported learning information that they may have already known, they were more likely to use the chatbot to classify instead of verify their thinking, and in general, there was no difference between the reasoning expressed in each group. When offloading to external tools for tasks that require critical thinking, users may not be engaging in their own cognition, and thus the information provided by the tool may be perceived as new [11]. This may explain why participants in the experimental group who were highly confident in their abilities still expressed learning something new.

We suspect that users with lower technological proficiency may be at a higher risk of heavily relying on an AI assistant for security tasks. While this reliance may be helpful for those without the cognitive ability to evaluate the results provide by technology, it may also have negative effects on critical thinking for those who are capable to make decisions [21].

Our results also highlight trust between the users and AI. Agreement or trust in an AI tool may be rooted in confirmation bias. Research has found that users are more likely to agree with AI tools if their recommendations align with the thoughts the user already had [41]. This may also limit the ability of the user to think critically about their decisions over time. What this study does not reveal is the nature of disagreements between users and AI assistants. Our findings show limited disagreement between users and the chatbot, as participants tended to align with the bot's recommendations.

However, prior research suggests that when users disagree with AI systems, it can lead to trust-related issues and conflicts in the human-AI relationship [29] [7] [28]. Future work should focus on understanding and mitigating the impacts of human-AI conflict, particularly examining how disagreements affect user trust, decision-making processes, and long-term reliance on AI assistance in cybersecurity contexts.

This study demonstrates how the use of LLMs can negatively impact how everyday users make security decisions without resulting in negative outcomes for the task. However, future work should explore how LLM or chatbot usage might impact users in the short and long-term for tasks such as detecting phishing attempts, detecting false information, determining which permissions to allow, and identifying malicious browser extensions.

5.4 Using LLMs to Support User Decision-Making in Security

These results suggest that using LLMs to assist users in security activities that depend on critical thinking can be useful when implemented thoughtfully. If organizations decide to implement these tools, they should be introduced as an additional resource and not a replacement for cybersecurity training. Due to the potential of cognitive loss, LLMs should be introduced as an addition to one's safety arsenal. Additionally, the developers designing these tools, should integrate explainable AI (XAI) into their tool design in order to help users understand how and why the tool arrives at the conclusions that it does. If tools are transparent about their accuracy and suggestions, users with a knowledge imbalance may be able to evaluate suggestions more effectively [2].

6 Conclusion

Our study provides an examination of how ChatGPT assistance impacts users' email security decision-making, offering important insights for the field of human-computer interaction and AI assistance research. While chatbot assistance did not improve classification accuracy, it significantly enhanced self-reported learning and heightened risk perception regarding email threats.

Findings reveal a complex relationship between AI assistance tools and human decision-making in cybersecurity contexts. Discrepancies between self-reported learning and accurate classifications suggest that current cybersecurity awareness efforts may already be effective for straightforward threats, while also highlighting the potential value of AI tools for educational reinforcement rather than performance enhancement.

Future research could build on this work by examining how chatbot assistance supports users with low levels of technological literacy, particularly to explore the performance gains for that user group. In addition, investigating user responses to more complex or emerging attack vectors may help clarify the boundaries of AI support in evolving cybersecurity contexts. Longitudinal studies could also offer insight into how sustained exposure to AI assistance influences users' memory, critical thinking, and classification performance over time. Finally, exploring chatbot interactions in multilingual settings may shed light on how large language models perform across languages and for non-native English speakers.

Acknowledgments

This work was funded in part by the Hellman Fellows Program.

References

[1] 2017. Unpacking Spear Phishing Susceptibility. In *Financial Cryptography and Data Security*. Lecture Notes in Computer Science, Vol. 10323. Springer International Publishing AG, Switzerland, 610–627.

[2] Abdulla Al-Subaiey, Mohammed Al-Thani, Naser Abdullah Alam, Kaniz Fatema Antora, Amith Khandakar, and SM Ashfaq Uz Zaman. 2024. Novel interpretable and robust web-based AI platform for phishing email detection. *Computers & electrical engineering* 120 (2024), 109625–.

[3] Naser Abdullah Alam. 2021. Phishing Email Dataset. https://www.kaggle.com/datasets/naserabdullahalam/phishing-email-dataset. Accessed: 2025-05-30.

[4] Apache Software Foundation. 2004. SpamAssassin Public Corpus. https://spamassassin.apache.org/old/publiccorpus/. Accessed: 2025-05-24.

[5] Apache Software Foundation. 2025. SpamAssassin. https://cwiki.apache.org/confluence/display/SPAMASSASSIN/SpamAssassin. Accessed: 2025-06-02.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[7] Michelle Brachman, Zahra Ashktorab, Michael Desmond, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, and Aabhas Sharma. 2022. Reliance and Automation for Human-AI Collaborative Data Labeling Conflict Resolution. *Proceedings of the ACM on human-computer interaction* 6, CSCW2 (2022), 1–27.

[8] J. Buckley, D. Lottridge, J.G. Murphy, and P.M. Corballis. 2023. Indicators of employee phishing email behaviours: Intuition, elaboration, attention, and email typology. *International journal of human-computer studies* 172 (2023), 102996–.

[9] Marcus Butavicius, Kathryn Parsons, Malcolm Pattinson, and Agata McCormac. 2016. Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails. arXiv:1606.00887 [cs.CY] https://arxiv.org/abs/1606.00887

[10] Marcus Butavicius, Ronnie Taib, and Simon J. Han. 2022. Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. *Computers & security* 123 (2022), 102937–.

[11] Eriona Çela, Mathias Mbu Fonkam, and Rajasekhara Mouly Potluri. 2024. Risks of AI-assisted learning on student critical thinking: a case study of Albania. *International Journal of Risk and Contingency Management (IJRCM)* 12, 1 (2024), 1–19.

[12] Gordon V Cormack et al. 2008. Email spam filtering: A systematic review. *Foundations and Trends® in Information Retrieval* 1, 4 (2008), 335–455.

[13] Marco De Bona and Federica Paci. 2020. A real world study on employees' susceptibility to phishing attacks. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*. ACM, New York, NY, USA, 1–10.

[14] Xun Dong, John A Clark, and Jeremy L Jacob. 2010. Defending the weakest link: phishing websites detection byanalysing user behaviours. *Telecommunication systems* 45, 2-3 (2010), 215–226.

[15] Anthony Favier, Pulkit Verma, Ngoc La, and Julie A Shah. 2025. Leveraging LLMs for Collaborative Human-AI Decision Making. In *Proceedings of the AAAI Symposium Series*, Vol. 5. 60–62.

[16] Alessandro Fedele, Mirco Tonin, and Matteo Valerio. 2024. Phishing attacks: An analysis of the victims' characteristics based on administrative data. *Economics letters* 237 (2024), 111663–3.

[17] Michael Gerlich. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15, 1 (2025), 6.

[18] Catalina Gomez, Mathias Unberath, and Chien-Ming Huang. 2023. Mitigating knowledge imbalance in AI-advised decision-making through collaborative user involvement. *International journal of human-computer studies* 172 (2023), 102977–.

[19] Google. 2024. *Gmail security – How Gmail keeps your emails safe.* https://safety.google/gmail/ Accessed: 2025-05-27.

[20] Galen A. Grimes, Michelle G. Hough, and Margaret L. Signorella. 2007. Email end users and spam: relations of gender and age group to attitudes and actions. *Computers in human behavior* 23, 1 (2007), 318–332.

[21] Sandra Grinschgl, Frank Papenmeier, and Hauke S Meyerhoff. 2021. Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology* 74, 9 (2021), 1477–1496.

[22] Xinlei He, Savvas Zannettou, Yun Shen, and Yang Zhang. 2024. You Only Prompt Once: On the Capabilities of Prompt Learning on Large Language Models to Tackle Toxic Content. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 770–787.

[23] Suhaima Jamal, Hayden Wimmer, and Iqbal H. Sarker. 2024. An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach. *Security and privacy* 7, 5 (2024).

[24] Olumide Babatope LONGE, Stella Chinye CHIEMEKE, Olufade F. Williams ONIFADE, and Folake Adunni LONGE. 2009. Camouflages and Token Manipulations-The Changing Faces of the Nigerian Fraudulent 419 Spammers. *African journal of information and communication technology* 4, 3 (2009).

[25] Daniel Lowd and Christopher Meek. 2005. Good Word Attacks on Statistical Spam Filters.. In *CEAS*, Vol. 2005.

[26] Clara Maathuis. 2024. Human-Centred AI in Military Cyber Operations. In *19th International Conference on Cyber Warfare and Security: ICCWS 2024*. Academic Conferences and publishing limited.

[27] Kaie Maennel and Olaf M Maennel. 2024. Human-AI Collaboration and Cyber Security Training: Learning Analytics Opportunities and Challenges. In *2024 17th International Conference on Security of Information and Networks (SIN)*. 01–08. https://doi.org/10.1109/SIN63213.2024.10871610

[28] Gerald Matthews, Ryon Cumings, James Casey, April Rose Panganiban, Antonio Chella, Arianna Pipitone, Jinchao Lin, and Mustapha Mouloua. 2024. Compromise in Human-Robot Collaboration for Threat Assessment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 68, 1 (2024), 329–335.

[29] Elham Nasarian, Roohallah Alizadehsani, U.Rajendra Acharya, and Kwok-Leung Tsui. 2024. Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework. *Information fusion* 108 (2024), 102412–.

[30] Rita Olla, Emily Hand, Sushil J. Louis, Ramona Houmanfar, and Shamik Sengupta. 2024. A Cybersecurity Game to Probe Human-AI Teaming. In *2024 IEEE Conference on Games (CoG)*. 1–5. https://doi.org/10.1109/CoG60054.2024.10645666

[31] Chidimma Opara, Paolo Modesti, and Lewis Golightly. 2025. Evaluating spam filters and Stylometric Detection of AI-generated phishing emails. *Expert systems with applications* 276 (2025), 127044–.

[32] OpenAI ChatGPT. 2025. ChatGPT: GPT-4 model [Large language model]. https://chat.openai.com. Accessed: 2025-06-03.

[33] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, Cate Jerram, Lech J. Janczewski, Henry B. Wolfe, and Sujeet Shenoi. 2013. Phishing for the Truth: A Scenario-Based Experiment of Users' Behavioural Response to Emails. In *IFIP Advances in Information and Communication Technology*. IFIP Advances in Information and Communication Technology, Vol. AICT-405. Springer Berlin Heidelberg, Berlin, Heidelberg, 366–378.

[34] Evan F Risko and Sam J Gilbert. 2016. Cognitive offloading. *Trends in cognitive sciences* 20, 9 (2016), 676–688.

[35] Sergio Rojas-Galeano. 2024. Zero-Shot Spam Email Classification Using Pretrained Large Language Models. *arXiv.org* (2024).

[36] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas, and Dimitrios K. Nasiopoulos. 2024. Next-Generation Spam Filtering: Comparative Fine-Tuning of LLMs, NLPs, and CNN Models for Email Spam Classification. *Electronics (Basel)* 13, 11 (2024), 2034–.

[37] Scott Ruoti, Jeff Andersen, Tyler Monson, Daniel Zappala, and Kent Seamons. 2018. A comparative usability study of key management in secure email. In *Fourteenth symposium on usable privacy and security (SOUPS 2018)*. 375–394.

[38] Dawn M. Sarno, Joanna E. Lewis, Corey J. Bohil, and Mark B. Neider. 2020. Which Phish Is on the Hook? Phishing Vulnerability for Older Versus Younger Adults. *Human factors* 62, 5 (2020), 704–717.

[39] Dawn M. Sarno, Joanna E. Lewis, Corey J. Bohil, Mindy K. Shoss, and Mark B. Neider. 2017. Who are Phishers luring?: A Demographic Analysis of Those Susceptible to Fake Emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1 (2017), 1735–1739.

[40] MA Sasse, S Brostoff, and D Weirich. 2001. Transforming the 'weakest link' - a human/computer interaction approach to usable and effective security. *BT technology journal* 19, 3 (2001), 122–131.

[41] Friso Selten, Marcel Robeer, and Stephan Grimmelikhuijsen. 2023. 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public administration review* 83, 2 (2023), 263–278.

[42] Mallory C. Stites, Megan Nyre-Yu, Blake Moss, Charles Smutz, Michael R. Smith, Stavroula Ntoa, and Helmut Degen. 2021. Sage Advice? The Impacts of Explanations for Machine Learning Models on Human Decision-Making in Spam Detection. In *Artificial Intelligence in HCI*. Lecture Notes in Computer Science, Vol. 12797. Springer International Publishing AG, Switzerland, 269–284.

[43] Shahroz Tariq, Ronal Singh, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2025. Bridging Expertise Gaps: The Role of LLMs in Human-AI Collaboration for Cybersecurity. (2025).

[44] Ekramul Haque Tusher, Mohd Arfian Ismail, Md Arafatur Rahman, Ali H Alenezi, and Mueen Uddin. 2024. Email spam: A comprehensive review of optimize detection methods, challenges, and open research problems. *IEEE Access* (2024).

[45] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H. Raghav Rao. 2011. Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *DECISION SUPPORT SYSTEMS* 51, 3 (2011), 576–586.

[46] Karina Vold. 2024. Human-AI cognitive teaming: using AI to support state-level decision making on the resort to force. *Australian journal of international affairs* 78, 2 (2024), 229–236.

[47] Reda Yaich, Alexandre Balondrade, Antoine Sicard, Christelle Fouquiau, Guillaume Giraud, Kahina Amokrane-Ferka, and Emmanuel Arbaretier. 2025. Symbiotic Human-AI Collaboration For Augmented Cybersecurity Operations. In *AAAI 2025 Summer Symposium Series*.

[48] Yue Zhang, Serge Egelman, Lorrie Cranor, and Jason Hong. 2006. *Phinding Phish: Evaluating Anti-Phishing Tools.* Technical Report. Carnegie Mellon University. https://doi.org/10.1184/R1/6470321.v1

A Appendix

A.1 Codebook

This codebook documents the coding scheme used for analyzing participant responses to phishing email scenarios.

Suspicious Sender: Participants identified the sender email address or domain as unknown, unusual, or not matching the claimed organization or person.

Suspicious Links: Participants identified links in the email that appeared to have mismatched destinations or lacked the 's' in https.

Urgent Tone: Participants noted some emails creating a sense of urgency or pressure to resolve an issue that may negatively impact them if not dealt with as soon as possible.

Conversational Tone/Grammatical Errors: Participants identified poor grammar, spelling mistakes, or unprofessional language inconsistent with legitimate business communications.

Contact Company Directly: Participants indicated that contacting the organization through official channels found on their webpage or by logging into their accounts to verify if there is indeed an issue with their accounts or payment.

Request Sensitive Information Via Email: Participants recognized that the email requests financial and or login information to be sent via email by replying.

Unsolicited Offers/Requests: Participants noted that the email was unexpected and offered or requested information.

Not to click Email Links: Participants were aware that they should not click links in the email.

Generic greeting/message: Participants identified the use of generic greetings and email body as a tactic used by spammers or malicious actors.

Too Good to be True Offers: Participants recognized offers, prizes, or opportunities that seem unrealistic as a tactic used by spammers or malicious actors.

False Purchase Details: Participants noted that they first needed to verify that the purchase or transactions in the email were fabricated.

Incorrect Terminology: Participants identified that the emails used wrong technical terms or industry-specific language.

A.2 Survey Questions

The following contains our survey questions asked to each participant. The pre-surveys were administered before participants saw an email or chatbot and was used to get an understanding of their familiarity with technology, one survey is for the experimental group and the other is for the control group. Email-survey questions were administered to participants after they evaluated the spam email scenarios. Two versions were used: one for the experimental group (with chatbot assistance) and one for the control group (without chatbot assistance). The final survey was issued to both groups after the email tasks and was used to understand what differences would appear (if any) between the control and experimental group.

Pre-Survey (Control Group)

1. Which email provider do you use?
(Open text box)

2. How often do you use your email?

- Daily
- Weekly
- Monthly
- Yearly
- Seldom
- Not at all

3. Is English the first language you learned? (Your answer will not impact your ability to participate in this study. If you choose "I am a bot" you will be disqualified from participating.)

- Yes
- No
- I am a bot

4. Please check all phrases that apply to you.

- Delete emails
- Add an attachment to my email
- Create folders for email organization
- Detect spam emails without assistance
- Send emails
- Send emails with public key
- Detect spam emails with assistance

5. Please select the degree to which you agree or disagree with the following statement: "I am confident in my ability to identify spam emails".

- Strongly Agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly Disagree

Pre-Survey (Experimental Group)

1. Which email provider do you use?
(Open text box)

2. How often do you use your email?

- Daily
- Weekly
- Monthly
- Yearly
- Seldom
- Not at all

3. Is English the first language you learned? (Your answer will not impact your ability to participate in this study. If you choose "I am a bot" you will be disqualified from participating.)

- Yes
- No
- I am a bot

4. Have you ever interacted with ChatGPT?

- Yes
- No
- I am not sure

5. Have you ever interacted with a chatbot? (This could be for banking, shopping, general Q&A, or anything like ChatGPT, Google's Gemini, Siri, or Alexa)?

- Yes
- No
- I am not sure

6. Please check all phrases that apply to you.

- Delete emails
- Add an attachment to my email
- Create folders for email organization

- Detect spam emails without assistance
- Send emails
- Send emails with public key
- Detect spam emails with assistance

7. Please select the degree to which you agree or disagree with the following statement: "I am confident in my ability to identify spam emails".

- Strongly Agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly Disagree

Email Survey (Control Group)

1. How would you advise they respond to the email shown? (Please select all that apply)

- Keep, save, or archive the email
- Click on link or links provided (if any)
- Forward the email to someone else
- Reply to the email
- Delete the email
- Other (Please specify)

1.1 If you selected 'Other' as your response to the previous question, please elaborate on what action you would advise them to take, otherwise type 'N/A'.

(Open text box)

2. How would you categorize this type of email? (Please select all that apply)

- Spam
- Malicious Email
- Email likely sent to wrong person
- Marketing
- Not spam
- Potentially spam

2.1 Please explain why you selected the choice/choices in Q2.

(Open text box)

3. Did you learn anything new about detecting spam emails while viewing the email?

- Yes
- No

3.1 If your previous answer was 'Yes', please briefly describe what you learned from this interaction. Otherwise type N/A.

(Open text box)

4. If you used any external resources other than your own ability, mark the ones you used. (Please select all that apply)

- Google Searches
- Asked a trusted person (i.e. friend or relative)
- Online tools for spam detection
- Other (Please specify)
- None

4.1 If you selected 'Other' as your response to the previous question, please elaborate on what resources you utilized, otherwise type 'N/A'

(Open text box)

Email Survey (Experimental Group)

1. How would you advise they respond to the email shown? (Please select all that apply)

- Keep, save, or archive the email
- Click on link or links provided (if any)
- Forward the email to someone else
- Reply to the email
- Delete the email
- Other (Please specify)

1.1 If you selected 'Other' as your response to the previous question, please elaborate on what action you would advise them to take, otherwise type 'N/A'

(Open text box)

2. Please select the degree to which you agree or disagree with the following statement: "I agree with the answers the chatbot gave to my questions".

- Strongly Agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly Disagree

3. How would you categorize this type of email? (Please select all that apply)

- Spam
- Malicious Email
- Email likely sent to wrong person
- Marketing
- Not spam
- Potentially spam

3.1 Please explain why you selected the choice/choices in Q3.

(Open text box)

4. Did you learn anything new about detecting spam emails during your interaction with the chatbot?

- Yes
- No

4.1 If your previous answer was 'Yes', please briefly describe what you learned from this interaction. Otherwise type N/A

(Open text box)

5. If you used any external resources other than your own ability, mark the ones you used. (Please select all that apply)

- Google Searches
- Asked a trusted person (i.e. friend or relative)
- Online tools for spam detection
- Other (Please specify)
- None

5.1 If you selected 'Other' as your response to the previous question, please elaborate on what resources you utilized, otherwise type 'N/A'

(Open text box)

Final Survey (All Participants)

1. Please select the degree to which you agree or disagree with the statement: "I would use a chatbot to help me detect spam email".

- Strongly Agree
- Agree
- Neither agree or disagree
- Disagree

- Strongly Disagree

2. "I trust that the chatbot will give me accurate and reliable responses".

- Strongly Agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly Disagree

3. "Opening a spam email is risky".

- Not risky at all
- Slightly risky
- Moderately risky
- Very risky
- Extremely risky

4. "Clicking a link in a spam email is risky".

- Not risky at all
- Slightly risky
- Moderately risky
- Very risky
- Extremely risky

5. "A spam email could lead to serious consequences (e.g., financial loss, identity theft, malware)".

- Very unlikely
- Unlikely
- Neutral
- Likely
- Very likely

6. "I can usually tell the difference between a legitimate email and a phishing or spam email".

- Strongly Agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly Disagree

7. "I believe I would never fall for a spam or phishing email".

- Strongly Agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly Disagree

8. "I am confident in my ability to identify spam emails".

- Strongly Agree
- Agree
- Neither agree or disagree
- Disagree
- Strongly Disagree

9. Have you ever mistakenly clicked a link or replied to a spam or phishing email?

- Yes
- No
- I am not sure

10. Have you ever experienced consequences from a spam email (e.g., malware, account hacked, lost money)?

- Yes
- No
- I am not sure

11. Please define 'Spam' in your own words.
(Open text box)
12. Please enter your Prolific ID.
(Open text box)