From "Be Careful" to "Here's Why": Investigating User Reasoning with Context-Specific SMS Scam Warnings

Elijah Bouma-Sims, Enze Liu, Alexandra Xinran Li and Lorrie Faith Cranor
Carnegie Mellon University
Pittsburgh, PA 15213
Email: {eboumasi, enzeliu, xinranl2, lorrie}@andrew.cmu.edu

***Abstract*—SMS-based scams continue to pose a persistent security threat, yet today's mobile warning interfaces often provide generic alerts that users may overlook. In other domains, context-specific explanations have improved users' ability to evaluate malicious content, and advances in generative AI (GAI) make it feasible to generate such explanations at scale. While service providers are now exploring similar approaches for SMS, it remains unclear how to best present contextual information so that users can act on it appropriately. We conducted a task-based interview study ($n = 20$) with US-based Android users in which participants assessed SMS messages using an inbox-style interface. Participants viewed both current Google Messages warnings and hypothetical contextual warnings. Among other results, our findings indicate that users value the concrete direction and evidence provided by context-specific warnings, which helped them reason about the legitimacy of the message. However, participants had differing preferences on the level of detail necessary or the extent to which AI involvement should be disclosed. We conclude by discussing design implications of our results for integrating context-specific warnings into mobile messaging interfaces, with relevance for both SMS and other scam domains.**

2026 IEEE Symposium on Security and Privacy (SP) | 979-8-3315-6065-2/26/$31.00 ©2026 IEEE | DOI: 10.1109/SP63933.2026.00200

1. Introduction

Despite its long history, Short Message Service, or SMS, remains a popular venue for fraudsters to spread scams. According to the US Federal Trade Commission, text-message-based scams caused a reported $470 million of losses in 2024, which was a nearly 500% increase from 2020 [1]. A typical pipeline to combat these scams is to first automatically detect them, and then either block them from delivery or flag them with a warning. These systems (notably, Google Messages[1]) have opted for simple warnings that do not provide much detail about *why* a message is flagged. This approach runs contrary to best practices, which recommend that warnings provide users with the evidence necessary to help them make their decision [2], [3].

Historically, incorporating context-specific explanations has been challenging, due to the black-box nature of AI reasoning [4]. Generative AI (GAI) offers new opportunities to overcome these barriers. Modern language models can extract rich contextual cues, such as the named sender, brand mentions, or suspicious URLs, and articulate human-like explanations. With appropriate prompting, these models can highlight linguistic signals (e.g., urgency, mismatched URLs) and generate concise, context-specific warnings that may help users make better decisions. Indeed, Google has recently announced plans to utilize GAI to generate tailored scam warnings [5].

Recent studies on explainable phishing detectors have shown that LLM explanations can improve task performance [6], [7], [8], [9]. However, these efforts have focused primarily on the explanations themselves, evaluating their accuracy and linguistic quality, rather than on how such explanations function when embedded within security warnings. Yet in practice, explanations reach users through warning interfaces that must balance brevity, visual salience, and trust. How these design factors shape users' reasoning and behavior remains poorly understood in the context of SMS. We address this gap by asking:

1) How do users reason about instant messages from unknown senders in the presence of warnings?
2) How do different warning design choices, especially the integration of context-specific explanations of varying lengths, affect users' reasoning?

To answer these questions, we present a task-based interview study ($n = 20$) using high-fidelity mockups of the Google Messages app to examine how people interpret and respond to scam-related SMS warnings. Across interviews, participants primarily reasoned based on message features, such as URLs and linguistic errors, their prior experience, and expectations of "normal" communications. Warnings supported this reasoning in multiple ways: sometimes serving as a decisive signal, other times offering reassurance or highlighting specific cues and actions. Participants valued the context-specific direction and evidence generated by GAI, but differed in their preferred level of detail and reactions to AI disclaimers. Visual and interactive elements such as red iconography and one-tap "Report & Block" options further influenced perceptions.

This paper makes two main contributions. Empirically, it presents the first in-depth qualitative study of how people reason about SMS messages in an inbox-style interface

1. https://messages.google.com

DOI 10.1109/SP63933.2026.00200

when exposed to multiple, realistic warning designs, including false positives. It shows how warnings influence decision-making in multiple overlapping ways. From a design perspective, it reveals that users prefer context-specific, action-oriented warnings, while highlighting how design parameters, including length, tone, AI disclaimers, and visual styling, can either support or undermine effective responses.

2. Related Work

Our work builds on prior research studying instant messaging-based scams, the design of security warnings, and the development of explainable security warnings. We discuss each of these areas in turn.

2.1. Instant Messaging-based Scams

The computer security literature on scam SMS and instant messaging has developed along three dominant themes: characterizing scam messages and their infrastructure, developing automated detection systems, and understanding how users respond to messages. The first line of work focuses on characterizing the content and distribution patterns of scam messages. Researchers have collected scam SMS messages using diverse methods, including measurements of messages received through public gateways [10], analyses of user-reported datasets [11], [12], and longitudinal studies of messages collected from participants' phones [13]. These efforts have shed light on both the technical infrastructure and the social engineering tactics underlying SMS-based scams. Scammers frequently impersonate well-known brands or government agencies, mimic legitimate security alerts (e.g., two-factor authentication codes or delivery updates), employ shortened or obfuscated URLs, and use urgency or fear appeals to prompt quick action. Together, these studies provide a detailed picture of the evolving scam ecosystem and inform the selection of stimuli in our own study (Section 3.3).

Another line of related work has investigated the detection of scam messages using a variety of automated techniques, such as machine learning (e.g., [14], [15], [16]). This line of work focuses heavily on detection accuracy, paying less attention to the explainability of the results. Our work assumes that the underlying detection algorithm is accurate (except where false positives are deliberately introduced) and explains the decision result using contextual information that users can understand (e.g., the presence of a phishing URL). As we demonstrate later in Section 4, contextual warnings can positively influence users' reasoning and potentially encourage them to take safer actions.

The instant messaging scam research most closely related to ours examines how users perceive and evaluate SMS messages [17], [18], [19], [20]. Tabassum et al. conducted an interview study ($n = 29$) investigating participants' experiences with scams and their approach to instant messaging processing, showing that users rely on contextual reasoning and content-based cues, such as URL features or message formatting [19]. Katsarakes et al.'s eye-tracking study ($n = 25$) similarly found that users focus attention on the SMS body when making judgments [18]. Timko et al.'s survey study ($n = 187$) identified demographic correlates of scam susceptibility and found that attention to specific message features (particularly sender information and URLs) predicts correct classification [17]. Finally, Siadati et al. demonstrated the effectiveness of SMS-based attacks targeting 2FA codes and suggested text changes as a mitigation [20]. Our work builds on these insights by demonstrating how warnings affect SMS processing. Additionally, we explore the design space of warnings by varying their levels of contextual detail. Prior studies focus primarily on cases where no warnings are present [17], [18], [20] or include little exploration of their influence [19].

2.2. Security Warning Design

Our work builds upon prior research on general warning design, including studies on risk communication from outside the computing field. The Communication-Human Information Processing (C-HIP) model, developed by Wogalter et al., is particularly influential in this context [21]. C-HIP models how a warning produces behavioral change through a series of stages, requiring the delivery of the warning from a source to a receiver via a communication channel. To prompt action from the receiver, an effective warning must capture and maintain attention, support comprehension through clear language and appropriate visuals, and ultimately motivate safe behavior. While C–HIP highlights that failures at any stage can prevent compliance, our study focuses on how warning content shapes later stages of processing, especially comprehension and intended behavior.

Security researchers have examined how these general principles apply to digital interfaces [22]. One prominent line of work focuses on browser warnings (e.g., those deployed in Chrome and Firefox). While early warnings were more effective than subtle security indicators [23], they remained largely ineffective at discouraging dangerous behavior in practice due to their passive nature [24], [25]. Subsequent research has demonstrated ways to improve warning efficacy, such as making them require interaction to proceed [26], [27] and increasing the friction required to take dangerous actions [28], [29], [30]. Subsequently, longitudinal field studies have shown that browser warnings applying these principles are effective at discouraging unsafe behavior [31], [32].

Another major thread investigates phishing email warnings. Studies in this area have similarly shown that the effectiveness of such warnings can be enhanced by making them more active (e.g., briefly stopping users from clicking on the link [33]), by improving the timing and placement of the warning (e.g., displaying a warning next to the link when a user clicks on it [34]), and by improving the content (e.g., warnings that contain explanations [35] or nudges that alert users [36], [37]).

These studies provide important insights into the design of phishing warnings; however, they fail to explore the SMS context. While similar to email phishing, SMS scams present

unique challenges due to limited screen real estate, shorter message length, and a device form factor that increases the likelihood of divided attention. To our knowledge, only Zare et al. have explored warning design in the instant messaging context, focusing on how icons can signal whether a message is trustworthy or untrustworthy [38]. They found that participants favored icons that were familiar from everyday life or social media, such as traffic lights or profile verification badges. Our study offers a more comprehensive examination of warning design, with a particular emphasis on contextual explanations.

2.3. Explainable Security Warnings

A sub-area of warning design research has focused on creating explainable security warnings. Prior work has consistently demonstrated that contextual warnings improve users' comprehension and decision-making across different contexts, such as phishing emails [34], [35], phishing websites [39], Android malware [40], and phishing URLs [41]. Before LLMs, such warnings typically involved manually preparing messages or message templates or presenting important decision-making features. With the rise of LLMs, researchers have applied them to produce more user-friendly warnings. Studies to date have applied LLMs to phishing URLs, emails, and websites. For instance, Desolda et al. [7], Cau et al. [8], and Lim et al. [9] examined LLM-generated phishing email warnings, Rashid et al. [42] analyzed phishing URL explanations, and Roy et al. [43] addressed phishing website warnings. These studies have confirmed that LLM-generated contextual warnings are helpful to users.

Within the SMS context, Uddin et al. [44] illustrate how language models can highlight words that influence spam classification, and Lee et al. [45] show that LLMs can produce contextual warnings for smishing messages. However, both studies stop short of evaluating users' perception of such contextual information. The work closest to ours is Wang et al., which suggests that users find context-specific explanations helpful; however, they only briefly investigate why these explanations are helpful or which elements of the explanations are most effective [6]. Our work, by contrast, focuses on warnings, which are designed to capture users' attention and encourage users to adopt safer behaviors. Consequently, warnings must account for a wide range of factors, including the content presented, visual design, conciseness, and clarity, all of which influence user behavior. We extend this line of inquiry by conducting an in-depth analysis of which components of warning messages users consider most useful.

3. Methods

We conducted an online interview study with 20 US-based participants to understand their reasoning about SMS messages in the presence of warnings. Below, we describe our recruitment process, interview protocol, selection of benign and phishing messages, warning design choices, data analysis, and limitations.

3.1. Recruitment

We primarily recruited participants from Pittsburgh, PA. To avoid alerting participants to our focus on warning design, the recruitment material specified that our study aimed to understand how users respond to messages from unknown senders. We posted physical posters on the campus of Carnegie Mellon University, at local libraries, and on utility poles around the city. To reach older adult participants, we distributed a digital version of the poster on the mailing lists of lifelong learning centers at Carnegie Mellon University and the University of Pittsburgh. We also distributed the digital poster to personal social networks and Slack channels. People were encouraged to share the poster with others who may be interested in the study, even if they were not eligible themselves. Finally, we placed posters on the campus of the University of California, San Diego.

Recruitment materials directed potential participants to complete an informed consent form via a link on the poster. This form explained the procedure of the study and required potential participants to affirm their eligibility for the study (18 years of age or older, use an Android phone to send and receive instant messages, and able to scan a QR code with their phone to view images). Eligible participants were directed to a survey to collect demographic and contact information for scheduling the interview. The demographic survey is available in the extended Appendix [46].

Not all those who filled out the survey were invited for an interview. We purposively selected participants based on demographic characteristics. We sought to ensure that our sample included individuals from a diverse range of ages and levels of computing education or work experience. Age is a frequently investigated demographic variable in studies on phishing vulnerability, although it is unclear to what extent it affects users' susceptibility [47], [48]. In total, 53 people completed the demographic survey, 26 people were invited to participate in an interview, and 20 completed the interview. At the conclusion of the study, we discarded the demographic and contact information of those who did not participate in an interview. Consistent with similar qualitative studies [19], [37], we planned to recruit 20 to 30 participants. The research team regularly met to discuss emerging themes and whether the most recent interviews had generated new insights. Upon reaching the twentieth interview, the team agreed that the recent interviews had not yielded new ideas, indicating that theoretical saturation had been reached and that additional interviews would likely not yield new insights [49].

3.2. Interview Procedure

All interviews were conducted over Zoom by a single author, who also took notes during the sessions. The interviews were semi-structured, meaning the interviewer asked a set of standard questions, along with follow-up questions tailored to participants' responses. The interviewer conducted four pilot interviews to practice the interview script and ensure

that questions elicited informative responses. The pilot interviews are not included in our final analysis.

The final interview guide is available in Appendix A. Before each interview session, the interviewer briefly discussed the consent form to ensure that the participant agreed to being audio-recorded and understood the voluntary nature of their participation. The interviewer then asked a series of background questions (Q1 – Q5) about participants' mobile phone use.

The bulk of the interview time was dedicated to a message-processing task, using a within-subjects design. Each participant viewed a total of 12 text messages (six benign and six fraudulent) and was exposed to five different warning designs. The message stimuli and warnings are described in greater detail in subsequent subsections. The first two messages in each inbox were intended to warm up participants to the task: they were always one benign message and one scam message, with no warnings. The remaining messages were randomly ordered to control for order effects. The interviewer instructed participants that they were viewing messages received by a mobile phone user named Bob and gave some details relevant to the task (i.e., where he lives, the name of his bank, and the delivery services he uses). Messages were presented through a Figma[2] mockup of the Google Messages app (examples in Figure 1), which is the most widely used messaging app on Android. For each message, participants were prompted to tap on the message and review it. The interviewer then asked them to recommend an action for each message and explain why (Q6 – Q9). To avoid priming for deception, we solicited open-ended recommendations rather than legitimacy ratings. Participants were not restricted to a limited set of options and were encouraged to recommend any action they might normally take. For messages presented with a warning, the interviewer prompted the participant to explain whether the warning affected their reasoning and to provide feedback on the warning.

To approximate real-world conditions, most participants viewed the inbox on their personal mobile devices. Due to technical issues with her phone, P20 reviewed the messages through a screen share of the mockup, which the interviewer controlled. P3 viewed all but one message on their personal device. The final message in the inbox had to be viewed via a screen share due to an error in the implementation of the mockup, which resulted in a duplicate message being included.

After the message-processing task, participants were asked to provide general feedback on the warnings they saw and to share any other suggestions for improvement (Q10 – Q12). As the intended length of the interview was an hour, these questions were asked only in 14 interviews in which an hour had not elapsed before the conclusion of the messaging task. The average length of the recorded portion of interviews was 58 minutes and 4 seconds.

2. https://www.figma.com

3.3. Message Stimuli

The complete details of all message stimuli can be found in Appendix B. We selected fraudulent stimuli from the SmishTank dataset [11], a crowd-sourced smishing dataset that has been used in prior studies [17]. The version of the dataset used contained messages received between March 2022 and November 2023. We chose messages to reflect a diversity of common scams and different social engineering tactics that users encounter, as reported by Tabassum et al. and Timko et al. [11], [19]:

- **Fake Amazon Account Alert**: Claims the user's Amazon account was locked after a login from Turkey and urges them to click a link. Sent from an email to SMS gateway. Contains grammatical errors.
- **Fake Bank of America Alert**: Claims the user's account was locked and directs them to a shortened link. Contains a minor grammatical error.
- **Fake Delivery Alert**: Claims that a delivery is on hold due to invalid shipping information. Directs the user to click a link impersonating USPS.
- **Employment Scam**: Offers a part-time Amazon sales job paying $30–$200 per day for one hour of work. Directs users to reply via Telegram or a UK-based WhatsApp number.
- **Fake T-Mobile Giveaway**: Claims to offer a T-Mobile gift and links to a shortened URL. Contains a small typographical error (missing dash in "T-Mobile").
- **IRS Arrest Threat**: Claims to be from the IRS and threatens arrest unless the user calls a southeastern US number. Includes spelling errors and awkward phrasing.

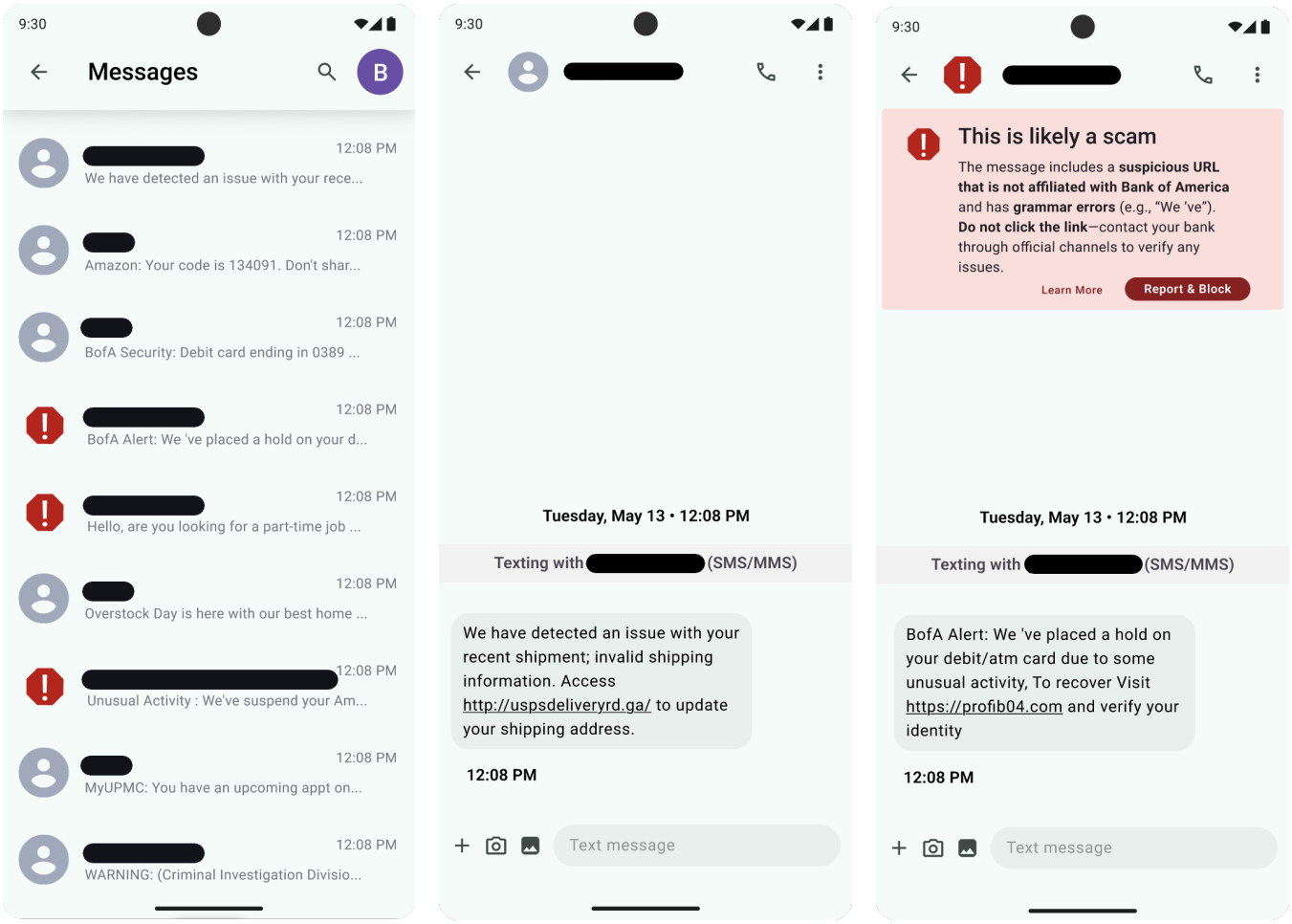
We selected benign messages that were legitimate counterparts to the fraudulent messages or messages that prior work suggests users may have difficulty identifying as legitimate, such as political campaign messages [19] and patient-doctor communications [50].

- **FedEx Notification**: Confirms a package delivery and links to an image of the shipment.
- **Healthcare Reminder**: Alerts the recipient to an upcoming appointment with a link to details.
- **Amazon OTP**: Provides a one-time passcode and includes a link for users who did not request it.
- **Overstock.com Promotion**: Advertises a 20% off sale with a shortened link to the promotion.
- **Bank of America PIN Alert**: Notifies the user that a card transaction was declined for missing a PIN and advises retrying.
- **Political Solicitation**: Urges support for a local candidate via a personalized message and link.

The doctor's appointment reminder and political solicitation were selected because they were relevant to Pittsburgh, PA, where the majority of participants were recruited.

3.4. Warning Designs

We evaluated five warning designs within our interview script. The first two warnings are employed by Google



(a) Inbox viewed by P8

(b) *Fake delivery notification*

(c) *Fake Bank of America alert* & long/confident contextual warning

Figure 1: Example screens displayed in the study. Redactions were not present during the study and are used to hide phone numbers.

Messages and were selected to represent the status quo of generic SMS warnings. The last three warnings followed the same visual design with varying levels of contextual details and confidence. While other dimensions of warning design exist, we chose to focus on the level of contextual detail and confidence based on the results of Wang et al. [6]'s study of LLM-based SMS classification. Explanation length emerged as a common concern among participants in their work, with some feeling that the "short" explanation they presented was too long. By varying the level of contextual detail, we sought to directly investigate how different lengths of explanation affect a user's experience. Furthermore, we investigated different confidence levels to probe how framing affects users' trust calibration. This was motivated by Wang et al.'s findings on over-reliance on AI-generated explanations, in which 18 of 24 participants reversed their correct initial classification of phishing after reading an incorrect AI explanation. We sought to investigate whether more cautious framing could mitigate this over-reliance by signaling the system's uncertainty and encouraging verification. Examples of each notice are shown in Figure 2:

- ***Save contact notice***: Default Google Messages notice for unknown senders, offering options to save the number or report as spam.
- ***Spam warning***: Google Messages notice for messages flagged as spam, stating the message resembles other spam messages and offering an option to mark it as not spam.
- ***Generic scam warning***: Prototype warning stating a message is likely a scam, with no explanation but options to "Report & Block" or "Learn more."
- ***Contextual scam warning (short, uncertain framing)***: Prototype warning that a message may be a scam, with one AI-generated sentence of advice, a disclaimer stating that the message was generated by AI and may not be accurate, and options to "Report & Block" or "Learn more."
- ***Contextual scam warning (long, confident framing)***:

Prototype warning that a message is likely a scam, with two to three AI-generated sentences of advice and options to "Report & Block" or "Learn more."

The design of the hypothetical warnings was based on best practices in risk communication [2], [51], emphasizing attention-grabbing color and iconography, clear hazard statements, and explicit recommended actions. Specifically, the warnings used a red stop-sign icon, headlines that explicitly included scam as a "signal word" [51], and GAI-generated explanation text paired with a one-tap "Report & Block" option to convey both the nature of the threat and the recommended response. To ensure visual consistency with the simulated Google Messages interface, the design was implemented using the Material 3 Design Kit[3] and a color palette derived from the existing application. The prompts used to generate the contextual-warning text are available in the extended Appendix [46].

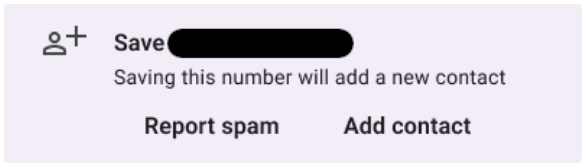
Apart from the first scam message that each participant saw, all scam messages were presented with one of the above warnings. The specific warning applied to each scam was randomly selected. One randomly selected benign message was presented with a false positive warning in the form of the *contextual scam warning (long, confident framing)*. The false positive was included to evaluate whether participants could identify benign content even when it was incorrectly labeled as fraud.

3.5. Analysis

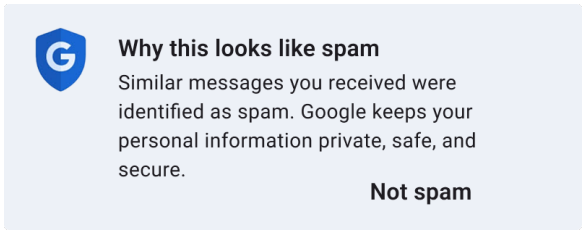
Interviews were audio-recorded and automatically transcribed via Zoom. The interviewer and another author reviewed the transcript to correct mistakes. The interviewer then divided the transcript into analysis units corresponding to sets of related questions and follow-ups. For each instant message, the analysis units were the action recommended by the participant, their reasoning for the action, the effect of the warning (if applicable), feedback on the warning (if applicable), their level of confidence, and whether they had received similar messages before. Analysis units were not always mutually exclusive, since interviewees often discussed topics across different parts of the interviews.

Analysis followed an iterative, inductive coding approach to identify themes. The resulting codebook is in the extended Appendix [46]. The interviewer served as the lead coder and developed an initial codebook based on notes from all the interviews and an analysis of three interviews. Two other authors then reviewed the codebook and the coding of these three interviews, providing feedback and refining the codebook. This process was repeated with six additional interviews in groups of three until the codebook stabilized, with no major changes made. The lead coder analyzed the remaining 11 interviews independently. The secondary coders then reviewed the codes and provided feedback. As with previous rounds, this feedback was discussed and integrated to determine the final set of themes.

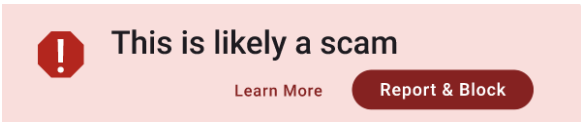
3. https://m3.material.io/



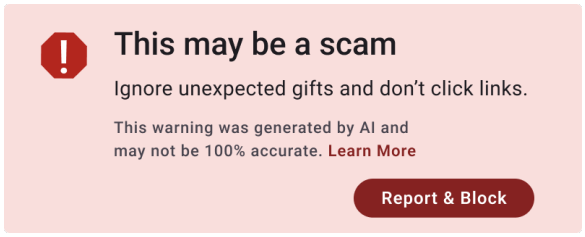
(a) Save contact notice



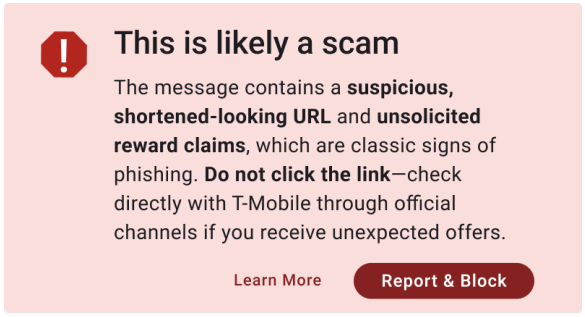
(b) Spam warning



(c) Generic scam warning



(d) Contextual warning (short, uncertain framing)



(e) Contextual scam warning (long, confident framing)

Figure 2: Examples of the warnings presented in the study on the T-Mobile Giveaway Scam

Since coding was a collaborative process where we attempted to reach consensus, we cannot calculate inter-rater reliability [52].

3.6. Limitations

Our study offers in-depth qualitative insights into users' instant messaging processing behavior and is not designed to draw quantitative conclusions. We provide code frequencies to contextualize the results; however, these should not be

interpreted as statements about the frequency of particular sentiments in the general population. While we attempted to recruit a diverse sample through purposive sampling, we do not claim a representative sample (see section 4.1). Moreover, most participants came from a single city in the United States. This may limit the generalizability of our findings, particularly for users outside Western cultures or those with lower educational attainment. Our study was conducted in a lab environment, so participants may not have reacted to the stimuli as they would in real life. Users' reactions to warnings are also subject to long-term habituation [53], [54] or other effects that may only manifest after repeated exposure. Due to time constraints, participants could view only a limited number of scam scenarios and warnings. Different scam scenarios may elicit different reactions and reasoning, especially considering that the contextual warning text would differ. Finally, while stimuli were randomized to control for potential order effects, participants' perspectives on later stimuli were likely influenced by their experience of earlier stimuli. Participants may also have developed cognitive fatigue due to the study's repetitive evaluations.

4. Results

In this section, we explain the results of our interview study. We start by describing our participants. Next, we discuss how participants reason about messages from unknown senders in the presence of different types of warnings. We conclude with an analysis of how various elements of a warning influence users' reasoning.

4.1. Participants

Table 1 shows the demographic characteristics of participants. Ten participants were between 18 and 34 years old, four were between 35 and 64, and six were 65 or older. In general, our sample was highly educated, with only four participants lacking a college degree. Prior work does not show a link between educational attainment and secure behavior [47], although differences may still exist. This skew toward higher education may impact the study's results, as above-average reading comprehension could affect the processing of warning information. Moreover, six participants reported having work experience or education in computing. This greater level of technical experience could result in an above-average ability to identify the scams within this study. Finally, all participants were white (15) or Asian (4), with two reporting Hispanic or Latino ethnicity, indicating limited racial/ethnic diversity in our sample. People of different racial and ethnic backgrounds may have different experiences with scams, which could affect participants' approach to the messages presented within the study. For example, immigrant communities are targeted by specialized immigration and government-imposter scams [55].

We also asked about participants' use of instant messaging. All participants discussed using instant messaging for *social* purposes, such as "catching up with friends and family" (P8). Most also had experience with the types of messages presented in the study. For example, all participants except P13 reported receiving One-Time Passwords (OTPs) via instant messaging, either for authentication or phone number verification. All except P20, P7, and P5 received *finance* related messages, such as transaction alerts from their bank or receipts from businesses. All except P1, P11, and P5 had experience receiving messages from a *healthcare* provider, like appointment reminders or requests for online check-in. As discussed in the following subsection, these participant characteristics and prior messaging experiences are particularly important, as many participants referenced their personal experiences when reasoning about unsolicited messages.

4.2. Reasoning about Unsolicited SMS Messages

In this section, we discuss how participants reasoned about messages during the instant message review task. We first provide a brief overview of participants' assessments of the different messages and describe the actions they recommended in various scenarios. These results provide important context for understanding how the warnings impacted participants' decision-making. We then describe the features participants used to reach their decisions, with a focus on warning-related reasoning. Our results show that participants engaged in a layered reasoning process when evaluating unsolicited SMS messages. Scam messages elicited decisive defensive actions, whereas legitimate messages prompted conditional reasoning dependent on context and prior experience. Across both, participants reported that message features were the primary cues, while warnings influenced decisions by serving as shortcuts, confirmations, or action prompts.

4.2.1. Overview of Message Assessments. Table 2 summarizes participants' judgments of each message they saw while explaining their recommended actions. Participants were coded as identifying a message as *legitimate* when they clearly indicated that they believed a message was genuine (e.g., "I would say, 10 out of 10 this is... not a scam. They can be annoying. But it's legitimate.") In contrast, participants were coded as identifying a message as *suspicious* when their response suggested that they believed the message was fraudulent (e.g., "As soon as I read this I'm aware that this is a scam..."). Participants' evaluations were coded as *unclear* when they didn't make an unambiguous statement about whether a message was real or when their reasoning was unrelated to whether the message was genuine.

In general, participants performed poorly at correctly identifying legitimate messages as clearly safe to interact with. This result likely stems from the presence of false positive warnings and the artificial nature of the study environment, where message analysis was a primary task. Some participants noted the lab effect. As P2 observed while reviewing the Amazon OTP message "I'm in a place where... I kind of view these messages with [a] more critical lens." This type of issue is common among usable

TABLE 1: Participant demographics. A dash indicates that a participant opted not to respond to a particular question. Question text and categories are abbreviated for space. P17 and P18 additionally selected the option to self-describe, but the information provided was not relevant to race.

ID	Age	Gender	Race	Hispanic or Latino	Computing Edu./Exp.	Education	Income
P1	35 to 44	Woman	White	No	No / No	Bachelor's	$50k to $100k
P2	18 to 24	Man	White	No	Yes / No	Bachelor's	> $150k
P3	25 to 34	Woman	Asian	No	No / No	Bachelor's	$25k to $50k
P4	35 to 44	Woman	White	No	Yes / Yes	Bachelor's	$100k to $150k
P5	18 to 24	Woman	Asian	No	No / No	Bachelor's	<$25k
P6	65 or older	Man	White	No	Yes / Yes	Graduate	-
P7	65 or older	Woman	White	No	No / No	Graduate	$100k to $150k
P8	25 to 34	Non-binary	White	No	Yes / Yes	Bachelor's	$25k to $50k
P9	25 to 34	Woman	Asian	No	No / No	Graduate	<$25k
P10	65 or older	Woman	Asian	No	No / No	Graduate	$50k to $100k
P11	18 to 24	Woman	White	No	No / No	High school	-
P12	35 to 44	Man	White	No	Unsure / No	Graduate	<$25k
P13	65 or older	Man	White	No	No / No	Some college	<$25k
P14	65 or older	Woman	White	No	No / No	Graduate	$100k to $150k
P15	45 to 54	Man	White	No	Yes / Yes	Some college	<$25k
P16	25 to 34	Woman	White	Yes	Yes / No	Bachelor's	$25k to $50k
P17	18 to 24	Woman	-	Yes	No / No	Bachelor's	<$25k
P18	18 to 24	Woman	White	No	No / No	Some college	$25k to $50k
P19	25 to 34	Man	White	No	No / No	Bachelor's degree	$25k to $50k
P20	65 or older	Woman	White	No	No / No	Graduate	$50k to $100k

security lab studies [56], [57], and means that the absolute distributions of participants' judgments or recommended actions should not be interpreted as representative of real-world accuracy. However, the cues participants attended to, and their qualitative reflections on the warnings themselves, remain valuable indicators of how people interpret and make sense of warning information in context.

Participants' action recommendations often defied mutually exclusive categorization. Many responded with multiple, conditional, or even contradictory actions. For example, P13 initially laughed at the Employment Scam, stating "I would block this." While explaining his action recommendation, however, he added that he might actually click the link within the message or attempt to find the job posting on the company's official website, "if I was looking... to work at some job like this." To capture the full breadth of decision-making, action frequencies reported throughout this section include all actions mentioned by participants.

Despite this complexity, qualitative differences emerged by message type. In line with their general assessment of scam messages as suspicious, participants typically recommended defensive actions. In the majority of cases, participants recommended reporting a scam message (76 occurrences of *report* out of 120 observations of scam messages) or blocking the sender (63 occurrences of *block* on scam messages out of 120 observations of scam messages). This result contrasts findings from Tabassum et al. [19], where over half of the participants stated that they generally would not report scam messages. As discussed later, this difference likely results from the presence of warnings that highlight the report option. Two exceptions are the Fake Bank of America Alert and the Fake Delivery Alert, in which participants often recommended contacting the relevant service externally to confirm there was no issue, even when they were confident the message was a scam. For example, upon viewing the Fake Bank of America Alert, P1 stated, "I would probably think that it is fake, but if I was concerned I would either go to my app or website and log in to verify or pick up the phone and call the bank." This indicates a preference for actions that guarantee safety, even when the deception is apparent.

For legitimate messages, participants frequently indicated that their responses would depend on external context not provided in the message, regardless of whether they believed the message was legitimate (76 occurrences of *conditional* out of 120 observations of legitimate messages). For instance, they might *ignore* a message or *interact* with a message based on the situation they were in. For the FedEx Notification, the majority of participants indicated that their decision would depend on whether they were actually expecting a package. P13 stated, "If Bob had actually ordered this package and was expecting it, I guess I would click through just to see... But if I hadn't ordered it... I wouldn't click." Similarly, most participants indicated that their response to the Amazon OTP message would depend on whether they were trying to access Amazon at the time. P19 stated, "if Bob is on amazon.com attempting to log in just type [the code]." When asked what they would do if Bob wasn't logging in at the time, they said, "I think I'd probably just delete this text because I would be concerned about clicking the link." For the Political Solicitation and Overstock.com Promotion, actions were frequently conditioned on individual interest or on whether Bob had signed up to receive these messages. P9 concluded about the Political Solicitation, "I don't think this warrants a reply, but if I wanted to read what they're trying to do and participate, that would be it." Tabassum et al. similarly observed that user decision-making about SMS messages was highly con-

TABLE 2: The frequency of the different judgment types, the number of action recommendations that were conditional, and the most common features for each message. The numbers in parentheses show the frequency that each message feature was mentioned, regardless of whether the feature increased or decreased suspicion.

	User Judgement of Message				
Message	**Legitimate**	**Unclear**	**Suspicious**	**Condl.**	**Message features frequently mentioned**
Fake Amazon Account Alert	0	0	20	3	url features (15), sender information (10), linguistic features (8)
Fake Bank of America Alert	0	1	19	3	url features (16), request for action (11), linguistic features (10)
Fake Delivery Alert	0	2	18	6	url features (18), personal experience (7), communication channel (5)
Employment Scam	0	2	18	3	communication channel (15), absurd offer (13), sender information (7)
Fake T-Mobile Giveaway	0	2	18	2	absurd offer (12), url features (10), linguistic features (5)
IRS Arrest Threat	0	0	20	3	communication channel (14), linguistic features (11), sender information (9)
FedEx Notification	6	10	4	14	url features (11), personal experience (8), linguistic features (7)
Healthcare Reminder	10	4	6	9	personal experience (13), url features (11), linguistic features (9)
Amazon OTP	8	7	5	14	personal experience (17), request for action (8), url features (8)
Overstock.com Promotion	1	9	10	11	url features (18), unfamiliar organization (6), personal experience (5)
Bank of America PIN Alert	8	3	9	16	linguistic features (10), personal experience (8), request for action (7)
Political Solicitation	11	9	0	12	url features (12), personal experience (8), sender information (6)

textual, with particular life situations making messages more or less credible [19]. The frequency of conditional actions for all messages is shown in Table 2.

4.2.2. Cues to Legitimacy or Deception. Table 2 lists the most frequent message features cited by participants, many of which overlap with those emphasized in the contextual warnings (e.g., URLs and language features). While the presence of contextual warnings may have increased the salience of these cues, participants identified them spontaneously, including before encountering any contextual warnings. These reasoning patterns have also been documented in prior studies [17], [19]. Below, we summarize the main themes. Where relevant, we indicate which warning condition was shown to contextualize user responses. Benign stimuli were viewed without warnings unless otherwise specified.

URL features were the most common indicator of deception across all messages. Participants frequently focused on mismatched or shortened links. For example, in the Fake Amazon Alert and Bank of America Scam, many participants noted that the links did not include any references to the companies they were supposedly associated with. As P9 said when viewing the message with the save contact notice, "there's the website that they've given to unlock the Amazon account. It's again not an Amazon website so super suspicious." The vast majority of participants (17) found the shortened URL in the benign Overstock.com Promotion suspicious, especially given its lack of connection to the company. As P16 said, "The link looks weird. It doesn't look like it even has the name of the store there." Tabassum et al. similarly observed confusion around shortened URLs [19]. URLs with legitimate top-level domains also sometimes drew suspicion. For example, four participants indicated that the URL within the FedEx Notification was suspicious, despite the domain being FedEx.com. P14 referred to parameters in the URL, stating, "I don't know. I've never seen a message that had that T and the N and the underscore U S... I just get this feeling it's not right." Such comments indicate that URL appearance was a highly salient heuristic, although it was sometimes inconsistently applied, aligning with prior work on phishing URL judgments [58].

Linguistic features, such as grammatical errors or perceived awkward phrasing, were another frequent focus. For example, 10 participants noted grammatical errors in the Fake Bank of America Alert message when explaining their recommended actions. While viewing under the save-contact notice, P1 stated, "the text message looks careless and suspicious with the ... bad spacing and the bad punctuation." While commonly emphasized in anti-scam education, these

types of cues can be unreliable, making well-crafted scam messages seem real or casting doubt on benign messages with perceived errors. For instance, the Healthcare Reminder message concluded with the statement that recipients could "Reply XRemind to opt out." Several participants (5) pointed to this as unusual or even suspicious. P19 stated, "it looks like a standard UPMC visit reminder, but it's got some typos at the bottom or confusing text, so it could be a scam either way."

Looking beyond specific message features, participants frequently discussed their *personal experience* and resulting expectations about what was normal. This code was the most common cue to legitimacy discussed by participants, and it was often paired with specific message reasoning codes. For example, 12 participants stated that their personal experience with healthcare appointment reminders made the Healthcare Reminder seem legitimate. When summarizing her reasons for recommending that Bob interact with the message, P4 stated, "It looks like other appointment reminders I've received. It has that little label at the beginning... because sometimes services share the short codes... It looks like the right format." This quote was coded as referencing both *linguistic features* and *personal experience*. Even when not referring specifically to *personal experience*, participants also discussed other specific expectations about messages. For many participants, the IRS Arrest Threat message was clearly ridiculous due to the *communication channel* (14 participants). As P6 stated when viewing the message with the contextual warning (long, confident framing), "IRS doesn't send text messages. Period. End of discussion." Participants also pointed to the presence of *absurd offers* in the Employment Scam (13 occurrences) and the Fake T-Mobile Giveaway (11 occurrences).

4.2.3. Effect of Warnings. Participants saw warnings on half the messages they reviewed, including five of the scam messages and one legitimate message (120 observations). In approximately half of these cases, participants mentioned the warning in their reasoning (68 instances of *warning (unprompted)*), and the exact frequency varies by type of notice. Only five participants mentioned the *save contact notice* without being prompted. The *spam warning* was also mentioned by less than half of the participants, with only eight participants describing it as a factor in their reasoning without being prompted. In contrast, the *long/confident contextual warning* and *false positive* warning were both mentioned by 15 participants. These differences in unprompted mentions may suggest that certain warning designs were more salient or influential in shaping participants' reasoning.

In many cases, participants did not clearly attribute an effect to warnings. Nevertheless, we identified four main ways that the SMS warnings helped users' reasoning. First, the presence of the warning was sometimes an important piece of evidence about the legitimacy of a message (38 occurrences of *warning as red flag*). The degree of effect varied. For some participants, the presence of the warning seemed to be the primary deciding factor. For example, when evaluating the Bank of America Scam message with the *short/uncertain contextual warning*, P5 stated, "Well, I'm not even gonna read it... if my phone flags it as a scam, I will just like report and block." P10 also adopted a policy of following the guidance of warnings. In explaining why she would click the report and block button on the *false positive* warning for the FedEx Notification, she said, "What would they have to gain by warning you? Nothing, nothing. There's no ulterior motive. It's for your benefit that they warn you." P19 also demonstrated this type of reasoning strategy across multiple messages, explaining, "Well, I don't like being on my phone.... So one way to reduce the amount of time is to not even bother reading it and just click 'Report and block' immediately... If it really is a serious issue, then they're gonna find some other way to get in touch with me." In these cases, warnings functioned less as a supplement to user reasoning and more as a shortcut, where participants outsourced judgment to the system.

This tendency to defer to the system seems to extend to inaccurate alerts. Participants generally trusted *false positive* warnings, with a little more than half of the participants (11) recommending that Bob report and block the message with the *false positive* warning. Only four participants indicated that they believed the warning they saw was incorrect.

Other participants indicated that they viewed the warning as only one factor among many. When reviewing the Overstock.com Promotion with a *false positive* warning, P18 explained, "I do pay attention to [warnings]. If there's a warning that says this is likely a scam, I'm more likely to pay closer attention to what does the link look like? Any spelling mistakes? It's just more scrutiny." They subsequently explained that they had a *general distrust* of warnings due to bad experiences, "Typically, I have had an email... falsely mislabeled something as a scam... in general it just encourages more scrutiny." While also grouped under the same theme, this kind of reasoning places greater emphasis on users' judgment of the message in conjunction with the warning's content.

In a similar vein, participants sometimes described warnings as providing reassurance or as affirming a conclusion they had already reached (10 occurrences of *affirmed right decision*). For example, P9 explained that even before opening the Fake Bank of America Alert, the grammatical errors present in the inbox preview made it look suspicious: "Even before I opened it I could see the exclamation mark, and I can see the 'we've placed a hold' so immediately it's like, Oh, this is not a real message. [The warning] just sort of confirms what I already knew." In the case of emotionally charged scams, like the IRS Arrest Threat, a warning may be emotionally reassuring. In explaining his decision to report and block, P2 stated, "Without the scam message I might be a little bit worried... I wouldn't actually call them, but it would be a little bit like concerning." These examples illustrate the value of warnings, even when they did not decisively change a participant's behavior.

Warnings also served to make certain actions more apparent or efficient (15 occurrences of *highlighted particular action*). This was especially noted as an effect of the *save contact notice* (6 occurrences). For example, when viewing

the Bank of America Scam Alert, P1 initially stated that she would "probably just delete it" before seeing the notice and stating, "Oh, I guess there is a 'report spam'. I would probably click that. I like it when they make it easy." The effect of options was less frequently noted for other warnings, but it seems apparent based on participants' recommended actions. Participants recommended reporting and blocking a message in 54 of the 80 observations of warnings with a report and block option present.

Finally, some participants discussed how a warning highlighted specific evidence of fraud (9 occurrences *highlighted evidence*). This was noted for both the contextual warnings and the *save contact notice*. The contextual warnings served to highlight specific suspicious cues within the message. For example, P16 stated, "The pay being ridiculously... generous, for like an hour of work, I might have missed that if I didn't have the information on the message." The *save contact notice* served only to highlight the sender's information or that the number was not saved in their contacts. For example, P2 indicated that the notice influenced their reasoning when viewing the Fake T-Mobile Giveaway, sharing, "Often, like T-Mobile or your provider has already sent you messages... they're not an unknown or a new number."

While one might expect that incorrect evidence in a *false positive* warning would prompt users to question the warning's accuracy, some participants instead rationalized it. For example, P19 credited the evidence within the *false positive* warning as decisive, explaining "It says that... the link is pretending to be from FedEx. So, despite it saying FedEx.com, it would take you to a different location." This may reflect a misunderstanding of the differences between email and SMS phishing. In email, a hyperlink's visible text can differ from its actual destination, but SMS messages display the full URL directly, meaning the link shown is the true destination. This highlights a broader risk of explanatory warnings: when false positives include incorrect or irrelevant evidence, users may treat the added detail as reinforcing the warning's credibility rather than questioning it. In such cases, explanations can make inaccurate warnings more persuasive, rather than helping users recognize inconsistencies.

Taken together, these findings suggest that warnings influenced participants in multiple, sometimes overlapping ways: as decisive signals that replaced deliberation, as one factor within broader reasoning, as affirmations of suspicions or reassurance in stressful scenarios, and as cues that highlighted actions or evidence. While not always the primary driver of decision-making, warnings consistently influenced how participants interpreted and responded to messages, underscoring their role as credible, context-sensitive supports in scam detection.

4.3. Effect of Warning Design Parameters

In this section, we discuss how participants reacted to the different warning design patterns, with a focus on the effect of the contextual explanations. Participants' reactions revealed several recurring design tensions that cut across warning types. They valued context-specific guidance that clarified what action to take and why, yet too much explanatory text risked overwhelming readers. While confident, cautionary framing encouraged warning adherence, AI disclaimers sometimes undermined trust by introducing uncertainty. Replicating prior results, we observed the importance of making it easy to take action (e.g., the one-tap "Report and Block" option) and found that strong visual cues, like the color red and hazard icons, enhanced salience and urgency.

4.3.1. Warning explanations. Participants' reactions to warning text reflected a central design tradeoff between direction and evidence: they wanted clear, actionable guidance on what to do, but differed in how much explanation they needed to justify that advice. Most appreciated the contextual warnings for offering concise, situation-specific direction (17 participants coded as saying *likes level of direction* in response to one of the contextual warnings). For example, P4 reacted to the *short/uncertain contextual warning* by saying, "I like that it gives a warning, and that it says to not click links and to contact your bank directly. That is very good advice." This type of feedback was sometimes given with the proviso that the speaker did not need the advice, but it would be *helpful for others* (9 occurrences). P1 commented on the *short/uncertain contextual warning* for the Fake Amazon Account Alert: "This one does tell you not to click the link and to verify via the official app or website, which I think is a good reminder for people who might not be familiar with... text... like an older adult... I'm thinking about my dad."

The *long/confident contextual warning* provided detailed evidence of fraud alongside directions. Participants' reactions to this explanatory content were mixed, with nine participants stating that they *like the level of evidence* on the *long/confident contextual warning*. P15 saw the warning last, on the Fake Bank of America Account Alert, and felt that it was the best warning, explaining, "It talks about the... specific URL that's not affiliated with a Bank of America. It talks about the grammatical error I saw, and it also specifically says not to click on the link and to contact the bank through official channels... this warning message has all the bases covered." An additional five participants stated that they liked the level of evidence presented in the false positive warning; however, this evidence was often misleading (e.g., indicating that an Amazon.com URL was suspicious).

Others felt that the longer text risked overwhelming users, preferring briefer explanations that conveyed the main point without extra reading. Seven participants indicated that they *dislike the level of detail* on the *long/confident contextual warning* or false positive warning. For example, when asked for their thoughts on the warning, P19 stated, "It's a lot of words... I don't like phone reading in general.... I feel like that could just be in the 'Learn more' section. All I need to know is the recommended action." Other participants expressed dissatisfaction with the amount of text in the long

warning, but were less clear about what should be excluded. P7 stated that they liked the text for the Employment Scam, but felt like it may be too long: "It's a lot of text, but it is informative. My experience is that people often don't read messages that are too wordy."

For some who disliked the longer text, the *short/uncertain contextual warning* was more effective. However, by prioritizing brevity, the short warnings sometimes introduced confusion. P14 noted that the short contextual warning's advice not to click links seemed inconsistent with the presence of a "learn more" link: "It says, 'Don't click links.'... [but] where there's learn more that's clicking a link." This sort of confusion was also experienced by P10 and P7. The longer warnings more clearly referenced the link within the text message, which prevented this sort of issue.

The generic scam warning, which omitted explanatory text, similarly illustrated a tradeoff between clarity and brevity. Some participants appreciated its simplicity (5 occurrences of *likes level of detail*). P19 liked this warning best, explaining, "Fewer words is better." P20 initially expressed a preference for more text, but changed her mind upon seeing the warning with no text: "This is actually... a very effective message... what they're trying to convey to you is this is likely a scam. So pay attention..." More common was a desire for some explanation to be added (*dislikes level of evidence* or *dislikes level of direction*). For example, P8 reacted negatively to the warning, saying, "I mean, it's pretty unhelpful like this is like saying, this is likely a scam, but it doesn't tell me why. So I'm like... it could be, but... how do you know that?" She went on to also recommend that the warning include "like a suggested action of 'Don't click the link' would be good."

Finally, opinions on the explanation from the *spam warning* were divided. Some participants felt the general explanation ("Similar messages you received were identified as spam") was a sufficient level of evidence to justify the warning (7 occurrences of *likes level of evidence*), while others desired a more specific explanation of why a message looked like spam (4 occurrences of *dislikes level of evidence*). Some participants also felt the warning lacked clear instructions for the user (3 occurrences of *dislikes level of direction*). Interestingly, several participants commented on the second sentence of the warning: "Google keeps your personal information private, safe, and secure." This boosted confidence in some, like P17, who directly contrasted it with the contextual warning AI disclaimer: "I think it's nice that it says that... Google keeps your personal information private, safe and secure.... It suggests to me that I'm keeping some form of privacy... in a way that I think marking it as AI in the other spam message didn't." Others viewed the message as useless or deceptive, with P18 calling it "false" and P20 referring to it as "a little ad for Google." These tensions highlight how even brief, auxiliary text can influence users' perceptions of a warning.

4.3.2. Warning framing and confidence. Participants emphasized that the tone and confidence of a warning were also important in shaping their trust and intended response. Participants most often commented on warning framing and confidence in three contexts: the *save contact notice*, the *spam warning*, and the *short/uncertain contextual warning*. The *save contact notice* was criticized by several participants for lacking an explicit warning about the associated scam message (6 occurrences of *dislikes tone of warning*). P12 felt that the notice may encourage some people to "trust this number." P7 similarly felt the warning was "trying to get you to add [the number] rather than reporting it as spam." A few participants also argued that the Google Messages *spam warning* suffered from a similar issue (4 instances of *dislikes tone of warning*). In recommending that the headline be changed, P16 explained "this is more like, 'Why, this could be a spam'.... It sounds a little more doubtful, so people might not pay that much attention to it." These reactions indicate that when warnings convey uncertainty or a weak cautionary tone, users may either ignore the message or interpret it as implicitly endorsing the sender.

The presence of an AI disclaimer in the *short/uncertain contextual warning* sparked polarized reactions, revealing a tension between transparency and authority. Half of the participants felt it was unnecessary or even undermined the warning (10 occurrences of *warning shouldn't include ai disclaimer*). P8 explained that seeing the note about AI generation made them feel they needed to double-check the message, "I don't think it's useful to be told 'this was generated by AI. It may not be a hundred percent accurate' because I mean it's already saying 'this may be a scam'... it's adding more layers of maybes." Only a few participants thought the AI disclaimer was necessary (3 occurrences of *warning should include ai disclaimer*). For P18, who discussed their negative experiences with email spam warnings, the AI disclaimer increased her trust in the warning by directly acknowledging potential inaccuracy: "I might click the report and block on this one just because it says... 'this was generated by AI and may not be 100% accurate,' which is my experience." For P6, the text was useful since it alerted him to a *privacy concern*. He shared, "They tell you the AI read the text message... That always bothers me... that somebody or something is reading my text messages. I think that's a privacy problem." These divergent responses underscore the challenge of communicating AI uncertainty in the warning context: while disclaimers may increase trust for some users by acknowledging potential inaccuracy, for others, they introduce additional ambiguity that can reduce the perceived authority of the warning.

4.3.3. Warning options. The options within the warnings were less often discussed than the textual features, but participants generally seemed to value ease of access to defensive actions. Replicating the findings of Tabassum et al. [19], several users commented negatively on the "Not spam" option present on the Google Messages spam warning (6 occurrences of *dislikes options*). These participants generally wanted the ability to confirm that a message was spam. In a typical example, P2 explained, "I wish it had a button that lets me immediately say that it is spam...It makes

it a little harder to actually report this as spam and easier just to report it as not spam." In contrast, seven participants commented positively on the options available on the *save contact notice* or the four hypothetical warning conditions. When P15 was asked to provide feedback on the *generic scam warning*, he referred positively to the options, "there's a... report and block button right there to click on... easy peasy." In general, this suggests that users value the ability to take defensive actions from the message screen.

4.3.4. Visual design. Participants occasionally commented on the visual elements of the warnings, indicating the use of red color and the stop sign icon as influential. Some participants indicated they liked the red icon and color in the four hypothetical warning designs (5 participants coded as saying *likes iconography* or *likes color* for the hypothetical warnings). In discussing the *contextual warning (short, uncertain framing)*, P5 said, "It stands out because it's red and it's like the first thing you see when you open up the message." The red octagon icon, which appeared in the inbox view for all but the *save contact notice*, reportedly influenced some users' perceptions before they opened the message. When discussing the IRS Arrest Threat, P9 explained, "Even before you open it, there are exclamation marks next to it. That sort of biases your viewpoint, unless you know the number." More decisively, in discussing the icon associated with the warning on the Fake Bank of America Alert, P8 stated, "I probably just wouldn't have opened this in the first place to be honest."

In contrast, the blue or purple coloring of the Google Messages notices was generally viewed negatively by participants who commented on it (5 participants coded as saying *dislikes color* or *dislikes iconography* for *save contact notice* or *spam warning*). Discussing the *spam warning* on the Fake Bank of America Account Alert, P1 stated, "It's also not in red. So it's not as like glaring to people... if someone's not really paying attention... they might just click that link." Similarly, P5 indicated that the blue color of the warning failed to capture her attention, saying, "the warning message... doesn't really... stand out... so I naturally gravitated towards the actual content of the message below." In sum, participants viewed the red visual elements as an effective cue that reinforced the warning's message and, in some cases, supported snap judgments about suspicious messages without the need for careful reading.

5. Discussion

In this section, we discuss our results, focusing on our novel findings regarding warnings that include context-specific explanations.

Our findings affirmed several patterns in how participants reasoned about unsolicited SMS messages. As a group, participants responded decisively to clearly scam-like messages (often recommending that they be reported or blocked), but adopted more conditional and uncertain strategies for legitimate messages. Participants commonly reasoned based on message features, their personal experience, and other expectations about what "normally" occurs. These results confirm prior work on SMS message processing [17], [19]. Our findings extend the existing literature by demonstrating how warnings actively support users' reasoning. They sometimes serve as a decisive signal that short-circuits further evaluation. In other cases, they acted as one factor among many or as a source of reassurance that confirmed existing suspicions. Warnings also highlighted particular actions or pieces of evidence, shaping both perception and behavior.

Our exploration of design features revealed some patterns that align with prior warning research. Participants favored visually salient designs that used strong color contrast and clear iconography to convey danger, particularly red elements that stood out against the blue and purple hues of the Google Messages interface. While the preference for red may be culturally specific [59], prior work supports the value of contrasting colors and distinctive iconography for capturing users' attention in warnings [38], [60]. Participants also preferred warnings that offered immediate defensive actions, such as "Report and Block," over options that appeared to downplay risk (e.g., "Not spam"), mirroring the results of Tabassum et al. [19] and reflecting best practices [2].

Beyond visual and action-oriented features, participants also responded positively to warnings that included context-specific text. They often described these as clearer or more informative, with a particular emphasis on the ability to provide specific guidance and contextual evidence.

The designs we tested revealed a key trade-off between **detail and readability**. Longer explanations conveyed clearer evidence but were visually dense, whereas shorter explanations were easier to scan but sometimes introduced confusion or contradictory cues (e.g., "don't click links" next to a "learn more" link). This tension may be especially consequential in mobile contexts, where users often interact under time pressure or partial attention. Although participants in our study were highly attentive by design, their comments about the dense text suggest that it could exceed attention limits in everyday use.

Improved formatting (e.g., increased white space or structured bullet points) could help mitigate this problem and preserve the benefits of contextual detail [21]. This may be challenging to control due to the stochastic nature of GAI output. Alternatively, contextual warnings could also be made configurable, allowing users to select a warning format based on their preferences. Designers could also offer expandable explanations, such as through "accordion" interfaces that reveal longer reasoning on demand. Prior work on notice design suggests that few users voluntarily access secondary text [32], [61]. Still, such layered designs could preserve attention and readability while allowing motivated users to access the richer, context-specific evidence that supports comprehension.

A second tension arose around **confidence and transparency**. Participants generally preferred warnings that appeared certain and directive. The AI disclaimer, especially the phrase noting that the warning "may not be 100% accurate," undermined perceived danger for many. At the

same time, a few participants viewed the disclaimer positively. For them, acknowledging AI involvement increased transparency and trust, either by aligning their experiences with those of spam detection systems or by highlighting potential privacy implications. Balancing these perspectives is difficult: transparency is ethically important, but excessive qualification can reduce authority. A more neutral phrasing, such as "This warning was generated by AI," may preserve transparency without signaling excessive doubt.

These dynamics were also reflected in participants' responses to false-positive warnings. Even when a legitimate message was incorrectly flagged, most participants accepted the warning's authority and recommended reporting or blocking the sender. This behavior could reflect trust in the warnings developed over the course of the study, leading participants to reflexively follow them. This result could also be confounded by the study's design. For example, participants may have complied with false-positive warnings presented later in the study to reduce cognitive burden, rather than out of trust in the system. In many cases, deference to the system may be fail-safe: it is arguably better for users to heed an occasional false alarm than to overlook a genuine threat. However, this pattern also underscores the need for calibrated trust. Users should view warnings as informative signals rather than unquestionable truths, attending to the evidence they provide while recognizing that even contextual, AI-supported explanations can be wrong or even intentionally manipulated through prompt injection. These issues are common to all GAI systems [62], but are especially challenging to address in the context of warnings.

These results are most directly applicable to the Android instant messaging context; however, some may generalize more broadly. For example, while iMessage (the default SMS app for iOS) does not presently include scam or spam warnings, it does include a notice on all messages from unknown senders, which highlights the option to "Report Spam" [63]. This UI element may help highlight defensive actions in the same way as the *save contact notice* on Android, though the iOS implementation is more subtle, appearing as text embedded in the background rather than a distinct pop-up box.

Our results also must be interpreted within the limits of the study context. Participants viewed messages in a controlled, task-focused setting where attention to potential deception was unusually close. In everyday use, where users are multitasking and cognitively loaded, warnings may be overlooked or interpreted differently, meaning the reasoning patterns observed here likely represent a best-case scenario for user engagement with warning text. Future work should test context-specific warnings under more ecologically valid conditions. While real-world or longitudinal studies would provide the strongest evidence of how users respond to warnings over time, controlled laboratory settings can also approximate real-world constraints through methods such as introducing time pressure or divided-attention tasks. For example, Sherman et al. enforced a time limit in their evaluation of spam call indicators [64].

6. Conclusion

We conducted an interview study with 20 participants to evaluate how SMS scam warnings impact users' reasoning and how different warning design choices—especially context-specific explanations generated with GAI—affect users' decision making. Confirming prior work, we found that participants relied on message features (e.g., URLs, linguistic features, etc.), personal experience, and expectations of "normal" communication to identify scams. Warnings supported decision making in diverse ways, serving as a decisive signal, acting as one piece of evidence among many, spotlighting particular message features, or highlighting available actions. Most participants valued the context-specific direction and evidence in GAI explanations, but opinions were split on the preferred level of detail and whether an AI disclaimer was necessary. The perceived level of danger, as well as visual and interactive elements such as red iconography and one-tap "Report & Block" options, further influenced perceptions. These results provide guidance for designers of next-generation, AI-powered SMS warnings.

7. Ethics Considerations

The study protocol and materials were approved by our Institutional Review Board. All participants completed an informed consent form that explained the risks and potential benefits of their participation in the study. Participants were compensated $20 in the form of an Amazon gift card. While we presented real scam stimuli in our experiment, our mockups were designed such that it was not possible to click any links or copy text within messages. Moreover, all scam URLs were dead links that would no longer function if visited.

8. LLM Usage Considerations

LLMs were used for editorial purposes in this manuscript (i.e., improving phrasing, laying out figures, and formatting tables), and all outputs were inspected by the authors to ensure accuracy and originality. For the study stimuli, ChatGPT was used to generate contextual warning text. Because the model is proprietary, exact reproduction of the warning text may be limited. To mitigate this issue, we share the prompts used to generate the output and provide the verbatim text that appeared in the warnings. The exact carbon footprint is unknown; however, warnings were generated using a consumer ChatGPT account with single queries for each warning (18 in total). An unknown number of additional queries were made during the development of the prompt. In future work, smaller, on-device models could reduce environmental impact.

Acknowledgments

This material is based upon work supported by the Innovators Network Foundation, a Carnegie Mellon University

CyLab Presidential Fellowship, and a Google Academic Research Award.

References

[1] Federal Trade Commission, "New ftc data show top text message scams of 2024; overall losses to text scams hit $470 million," 2025. [Online]. Available: https://web.archive.org/web/20250421172343/https://www.ftc.gov/news-events/news/press-releases/2025/04/new-ftc-data-show-top-text-message-scams-2024-overall-losses-text-scams-hit-470-million

[2] L. Bauer, C. Bravo-Lillo, L. Cranor, and E. Fragkaki, "Warning design guidelines," Cylab, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CyLab-13-002, Feb. 2013. [Online]. Available: https://kilthub.cmu.edu/articles/journal_contribution/Warning_Design_Guidelines_CMU-CyLab-13-002_/6468131?file=11896667

[3] R. Reeder, E. Cram Kowalczyk, and A. Shostack, "Poster: Helping engineers design neat security warnings," in *Proceedings of the Symposium On Usable Privacy and Security (SOUPS '11)*. Pittsburgh, PA: Association for Computing Machinery, 2011.

[4] W. J. Von Eschenbach, "Transparency and the black box problem: Why we do not trust ai," *Philosophy & technology*, vol. 34, no. 4, pp. 1607–1622, 2021.

[5] J. Bawa, "How we're using ai to combat the latest scams," retrieved from https://blog.google/innovation-and-ai/technology/safety-security/how-were-using-ai-to-combat-the-latest-scams/ on April 14, 2026.

[6] Y. Wang, H. Zhai, C. Wang, Q. Hao, N. A. Cohen, R. Foulger, J. A. Handler, and G. Wang, "Can you walk me through it? explainable SMS phishing detection using LLM-based agents," in *Proceedings of the Twenty-First Symposium on Usable Privacy and Security (SOUPS '25)*. Seattle, WA, USA: USENIX Association, 2025, pp. 37–56.

[7] G. Desolda, F. Greco, and L. Viganò, "APOLLO: A gpt-based tool to detect phishing emails and generate explanations that warn users," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 4, pp. 1–33, 2025.

[8] F. M. Cau, G. Desolda, F. Greco, L. D. Spano, and L. Viganò, "Can large language models improve phishing defense? a large-scale controlled experiment on warning dialogue explanations," *arXiv preprint arXiv:2507.07916*, 2025.

[9] B. Lim, R. Huerta, A. Sotelo, A. Quintela, and P. Kumar, "Explicate: Enhancing phishing detection through explainable ai and llm-powered interpretability," *arXiv preprint arXiv:2503.20796*, 2025.

[10] A. Nahapetyan, S. Prasad, K. Childs, A. Oest, Y. Ladwig, A. Kapravelos, and B. Reaves, "On SMS phishing tactics and infrastructure," in *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP '24)*. San Francisco, CA, USA: IEEE, 2024, pp. 1–16.

[11] D. Timko and M. L. Rahman, "Smishing dataset I: Phishing SMS dataset from smishtank.com," in *Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy (CODASPY '24)*. Porto, PRT: Association for Computing Machinery, 2024, p. 289–294.

[12] S. Agarwal, A. Papasavva, G. Suarez-Tangil, and M. Vasek, "Fishing for smishing: Understanding sms phishing infrastructure and strategies by mining public user reports," in *Proceedings of the 2025 ACM on Internet Measurement Conference (IMC '25)*. Madison, WI, USA: Association for Computing Machinery, 2025.

[13] L. Razaq, T. Ahmad, S. Ibtasam, U. Ramzan, and S. Mare, "" we even borrowed money from our neighbor" understanding mobile-based frauds through victims' experiences," *Proceedings of the ACM on human-computer interaction*, vol. 5, no. CSCW1, pp. 1–30, 2021.

[14] C. Balim and E. S. Gunal, "Automatic detection of smishing attacks by machine learning methods," in *Proceedings of the 1st International Informatics and Software Engineering Conference (UBMYK)*. Ankara, TUR: IEEE, 2019, pp. 1–3.

[15] A. K. Jain, S. K. Yadav, and N. Choudhary, "A novel approach to detect spam and smishing sms using machine learning techniques," *International Journal of E-Services and Mobile Applications (IJESMA)*, vol. 12, no. 1, pp. 21–38, 2020.

[16] G. Sonowal and K. Kuppusamy, "Smidca: an anti-smishing model with machine learning approach," *The Computer Journal*, vol. 61, no. 8, pp. 1143–1157, 2018.

[17] D. Timko, D. H. Castillo, and M. L. Rahman, "Understanding influences on sms phishing detection: User behavior, demographics, and message attributes," in *Proceedings of the Symposium on Usable Security and Privacy (USEC) 2025*. San Diego, CA, USA: The Internet Society, Feb. 2025.

[18] E. A. Katsarakes, M. Edwards, and J. D. Still, "Where do users look when deciding if a text message is safe or malicious?" in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 68, no. 1. Phoenix, AZ, USA: SAGE Publications, 2024, pp. 221–225.

[19] S. Tabassum, C. Faklaris, and H. R. Lipford, "What drives SMiShing susceptibility? a US. interview study of how and why mobile phone users judge text messages to be real or fake," in *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS '24)*. Philadelphia, PA, USA: USENIX Association, 2024, pp. 393–411.

[20] H. Siadati, T. Nguyen, P. Gupta, M. Jakobsson, and N. Memon, "Mind your smses: Mitigating social engineering in second factor authentication," *Computers & Security*, vol. 65, pp. 14–28, 2017.

[21] M. S. Wogalter, C. B. Mayhorn, and K. R. Laughery Sr, "Warnings and hazard communications," *Handbook of human factors and ergonomics*, pp. 644–667, 2021.

[22] L. F. Cranor, "A framework for reasoning about the human in the loop," in *Proceedings of the 1st Conference on Usability, Psychology, and Security (UPSEC '08)*. San Francisco, CA, USA: USENIX Association, 2008.

[23] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer, "The emperor's new security indicators," in *Proceedings of the 2007 IEEE Symposium on Security and Privacy (SP '07)*. Oakland, CA, USA: IEEE, 2007, pp. 51–65.

[24] 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 (CHI '08)*. Florence, ITA: Association for Computing Machinery, 2008, pp. 1065–1074.

[25] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Montréal, QC, CAN: Association for Computing Machinery, 2006, pp. 581–590.

[26] M. Wu, R. C. Miller, and S. L. Garfinkel, "Do security toolbars actually prevent phishing attacks?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Montréal, QC, CAN: Association for Computing Machinery, 2006, pp. 601–610.

[27] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter, "Your attention please: designing security-decision uis to make genuine risks harder to ignore," in *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. Newcastle, UK: Association for Computing Machinery, 2013.

[28] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor, "Crying wolf: An empirical study of SSL warning effectiveness." in *Proceedings of the 18th USENIX Security Symposium (USENIX Security '09)*. Montréal, QC, CAN, 2009, pp. 399–416.

[29] A. P. Felt, R. W. Reeder, H. Almuhimedi, and S. Consolvo, "Experimenting at scale with google chrome's SSL warning," in *Proceedings of the SIGCHI conference on human factors in computing systems (CHI '14)*. Toronto, ON, CAN: Association for Computing Machinery, 2014, pp. 2667–2670.

[30] A. P. Felt, A. Ainslie, R. W. Reeder, S. Consolvo, S. Thyagaraja, A. Bettes, H. Harris, and J. Grimes, "Improving SSL warnings: Comprehension and adherence," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems (CHI '15)*, Seoul, KOR, 2015, pp. 2893–2902.

[31] R. W. Reeder, A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, and S. Egelman, "An experience sampling study of user reactions to browser warnings in the field," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Montréal, QC, CAN: Association for Computing Machinery, 2018, p. 1–13.

[32] D. Akhawe and A. P. Felt, "Alice in warningland: a large-scale field study of browser security warning effectiveness," in *Proceedings of the 22nd USENIX security symposium (USENIX Security '13)*. Washington, D.C., USA: USENIX Association, 2013, pp. 257–272.

[33] M. Volkamer, K. Renaud, B. Reinheimer, and A. Kunz, "User experiences of torpedo: Tooltip-powered phishing email detection," *Computers & Security*, vol. 71, pp. 100–113, 2017.

[34] J. Petelka, Y. Zou, and F. Schaub, "Put your warning where your link is: Improving and evaluating email phishing warnings," in *Proceedings of the 2019 CHI conference on human factors in computing systems (CHI '19)*. Glasgow, UK: Association for Computing Machinery, 2019, pp. 1–15.

[35] P. Buono, G. Desolda, F. Greco, and A. Piccinno, "Let warnings interrupt the interaction and explain: designing and evaluating phishing email warnings," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Hamburg, DE: Association for Computing Machinery, 2023, pp. 1–6.

[36] J. Nicholson, L. Coventry, and P. Briggs, "Can we fight social engineering attacks by social means? assessing social salience as a means to improve phish detection," in *Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS '17)*. Santa Clara, CA, USA: USENIX Association, 2017, pp. 285–298.

[37] S. Y. Zheng and I. Becker, "Checking, nudging or scoring? evaluating e-mail user security tools," in *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS '23)*. Anaheim, CA, USA: USENIX Association, Aug. 2023, pp. 57–76.

[38] N. Zare, C. Faklaris, S. Tabassum, and H. R. Lipford, "Improving mobile security with visual trust indicators for smishing detection," in *Proceedings of the 2025 IEEE World AI IoT Congress (AIIoT '25)*. Seattle, WA, USA: IEEE, 2025, pp. 0212–0221.

[39] J. Aneke, C. Ardito, and G. Desolda, "Help the user recognize a phishing scam: Design of explanation messages in warning interfaces for phishing attacks," in *HCI for Cybersecurity, Privacy and Trust: Third International Conference (HCI-CPT '21)*. Berlin, DE: Springer-Verlag, 2021, p. 403–416.

[40] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. Siemens, "Drebin: Effective and explainable detection of android malware in your pocket." in *Proceedings of the Network and Distributed System Security Symposium 2014 (NDSS '14)*, vol. 14, no. 1. San Diego, CA, USA: The Internet Society, 2014, pp. 23–26.

[41] K. Althobaiti, N. Meng, and K. Vaniea, "I don't need an expert! making url phishing features human comprehensible," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, 2021, pp. 1–17.

[42] F. Rashid, N. Ranaweera, B. Doyle, and S. Seneviratne, "LLMs are one-shot url classifiers and explainers," *Computer Networks*, vol. 258, pp. 1–15, 2025.

[43] S. S. Roy, C. Torres, and S. Nilizadeh, "'Explain, Don't Just Warn!'–a real-time framework for generating phishing warnings with contextual cues," *arXiv preprint arXiv:2505.06836*, 2025.

[44] M. A. Uddin, M. N. Islam, L. Maglaras, H. Janicke, and I. H. Sarker, "Explainabledetector: Exploring transformer-based language modeling approach for sms spam detection with explainability analysis," *Digital Communications and Networks*, vol. 11, no. 5, pp. 1504–1518, 2025.

[45] Y. Lee and D. Han, "KorSmishing explainer: A korean-centric llm-based framework for smishing detection and explanation generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track (EMNLP '24)*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 642–656.

[46] E. Bouma-Sims, "Extended appendix to from 'be careful' to 'here's why': Investigating user reasoning with context-specific sms scam warnings," Apr 2026. [Online]. Available: osf.io/de89b

[47] M. Wei, J. Mink, Y. Eiger, T. Kohno, E. M. Redmiles, and F. Roesner, "Sok (or solk?): On the quantitative study of sociodemographic factors and computer security behaviors," in *Proceedings of the 33rd USENIX Security Symposium (USENIX Security '24)*. Philadelphia, PA, USA: USENIX Association, Aug. 2024.

[48] S. Zhuo, R. Biddle, Y. S. Koh, D. Lottridge, and G. Russello, "Sok: Human-centered phishing susceptibility," *ACM Trans. Priv. Secur.*, vol. 26, no. 3, apr 2023.

[49] M. Schreier, *The SAGE handbook of qualitative data collection*. SAGE Publications Ltd, 2018, ch. Sampling and generalization.

[50] C. Wang, Z. Jia, H. Benkraouda, C. Zevnik, N. Heuermann, R. Foulger, J. A. Handler, and G. Wang, "VeriSMS: A message verification system for inclusive patient outreach against phishing attacks," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Honolulu, HI, USA: Association for Computing Machinery, 2024.

[51] M. S. Wogalter, V. C. Conzola, and T. L. Smith-Jackson, "Research-based guidelines for warning design and evaluation," *Applied Ergonomics*, vol. 33, no. 3, p. 219–230, May 2002.

[52] N. McDonald, S. Schoenebeck, and A. Forte, "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Nov. 2019.

[53] B. B. Anderson, A. Vance, C. B. Kirwan, J. L. Jenkins, and D. Eargle, "From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it," *Journal of Management Information Systems*, vol. 33, no. 3, pp. 713–743, 2016.

[54] A. Vance, B. Kirwan, D. Bjornn, J. Jenkins, and B. B. Anderson, "What do we really know about how habituation to warnings occurs over time? a longitudinal fmri study of habituation and polymorphic warnings," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Denver, CO, USA: Association for Computing Machinery, 2017, p. 2215–2227.

[55] J. M. Pedroza, "Making noncitizens' rights real: Evidence from immigration scam complaints," *Law & Policy*, vol. 44, no. 1, pp. 44–69, 2022.

[56] A. Sotirakopoulos, K. Hawkey, and K. Beznosov, "On the challenges in usable security lab studies: lessons learned from replicating a study on SSL warnings," in *Proceedings of the Seventh Symposium on Usable Privacy and Security (SOUPS '11)*. Pittsburgh, PA, USA: Association for Computing Machinery, 2011.

[57] S. Fahl, M. Harbach, Y. Acar, and M. Smith, "On the ecological validity of a password study," in *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS '13)*. Newcastle, UK: Association for Computing Machinery, 2013.

[58] S. Albakry, K. Vaniea, and M. K. Wolters, "What is this url's destination? empirical evaluation of users' url reading," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Honolulu, HI, USA: Association for Computing Machinery, 2020, p. 1–12.

[59] C. Kawai, Y. Zhang, G. Lukács, W. Chu, C. Zheng, C. Gao, D. Gozli, Y. Wang, and U. Ansorge, "The good, the bad, and the red: implicit color-valence associations across cultures," *Psychological Research*, vol. 87, no. 3, pp. 704–724, 2023.

[60] M. J. Kalsher, M. S. Wogalter, and B. M. Racicot, "Pharmaceutical container labels: enhancing preference perceptions with alternative designs and pictorials," *International Journal of Industrial Ergonomics*, vol. 18, no. 1, pp. 83–90, 1996.

[61] E. Bouma-Sims, M. Li, Y. Lin, A. Sakura-Lemessy, A. Nisenoff, E. Young, E. Birrell, L. F. Cranor, and H. Habib, "A us-uk usability evaluation of consent management platform cookie consent interface design on desktop and mobile," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Hamburg, DE: Association for Computing Machinery, 2023.

[62] M. Nahar, H. Seo, E.-J. Lee, A. Xiong, and D. Lee, "Fakes of varying shades: How warning affects human perception and engagement regarding llm hallucinations," *arXiv preprint arXiv:2404.03745*, 2024.

[63] Apple, "Report spam and block senders in messages on iphone," retrieved from https://web.archive.org/web/20260302022128/https://support.apple.com/guide/iphone/report-spam-and-block-senders-iph3f94d910d/ios on March 2nd, 2026.

[64] I. Sherman, J. Bowers, K. McNamara, J. Gilbert, J. Ruiz, and P. Traynor, "Are you going to answer that? measuring user responses to anti-robocall application indicators," in *Proceedings of the 2020 Network and Distributed Systems Security Symposium (NDSS '20)*. San Diego, CA, USA: The Internet Society, Feb. 2020.

Appendix A. Interview Guide

The semi-structured interview format allowed the interviewer to ask follow-up questions as needed. Text in italics or square brackets was not read verbatim. Suggested follow-ups are included, but were not always asked. As the task progressed, participants would often begin answering questions without being explicitly prompted, so previously answered questions were not re-asked. The recording of the interview began after the introduction section.

A.1. Introduction

Thank you for taking the time to participate in this interview. We will begin by reviewing key areas of consent. This interview will be audio-recorded and transcribed for the purpose of our note-taking. Is it ok to audio-record you during this interview? *All participants said yes to this question*

Participation in the interview is voluntary, and you may choose to withdraw at any time. It will take around 60 minutes. The format will start with some background questions about your usage of mobile phones. We will then review various instant messages and discuss how you perceive each message. This task will require you to scan a QR code and then view images of instant messages on your phone. To help us protect your privacy, please do not reveal any private or personally-identifiable information about yourself or others, such as names, in response to our questions. Do you have any questions before we begin? *[Answer any questions]*
Feel free to ask any questions throughout the process.

A.2. Background Questions

1) What type of phone do you use on a daily basis?
2) When did you begin using this phone (approximately)?
3) Is this the phone you will be using during the interview today?
4) Who is your mobile phone carrier?
5) What types of things do you use instant messaging for? By instant messaging I mean SMS or RCS, but you may also consider things like WhatsApp or Facebook Messenger.
   a) **Follow-up:** What types of instant messaging do you do?
   b) **Follow-up:** Do you ever receive messages from your doctor or healthcare provider? What kinds of messages?
   c) **Follow-up:** Do you ever receive messages from your bank or other financial institutions? What kinds of messages?
   d) **Follow-up:** Do you ever receive codes that are used to login or sign-up to online services? What kinds of messages?
   e) **Follow-up**: Do you ever receive messages from other companies or websites? What kinds of messages?

A.3. Task Introduction

During the next portion of the interview, we will review some instant messages and get your impression on how you would respond to them. We are going to show you some text messages that were received by a mobile phone user named Bob who lives in Pittsburgh and frequently receives text messages from delivery services like USPS or FedEx, his bank, which is Bank of America, and other companies. He also receives spam. You're trying to decide how to respond to messages. Once the task begins, you will scan a QR code with your phone and view a set of messages . For each message, you should explain what you think Bob should do in response to the message and why. I will share these instructions with you in the zoom chat so you may refer to them throughout the task. Do you have any questions before we get started? *[Answer any questions]*

Feel free to ask any other questions you may have during this task

A.4. Task

[Prompt participant to view Nth message, referring to some salient feature like the phone number or start of text.] Please take a moment to read the message. Let me know when you are finished.

6) What do you think Bob should do after receiving this message? ***Potential follow-up prompting****: For example, you may recommend he reply, ignore, report, delete, click, etc.*
7) *If participant doesn't justify answer:* Why do you think that's the best response?
   a) ***Follow-up:*** Were there specific parts of the message that influenced your thinking?

b) ***Follow-up (if applicable)*:**This message includes a warning/notice. Did that affect what you think Bob should do?
c) ***Follow-up (if applicable)***: Have you seen this warning/notice before?
d) ***Follow-up (if applicable)*:** What do you think of the text of the warning/notice?
e) ***Follow-up (if applicable)***: This warning says that it was generated using AI. Does that impact your opinion of the warning or its content?
f) ***Follow-up (if applicable)***: Are there any changes you would make to improve this warning/notice? Why/Why Not?
g) ***Follow-up (if applicable)***: Do you think this warning/notice is better or worse than the others you have seen? Why or why not?

8) How confident are you in your decision/assessment? From 1 to 10.
9) Have you received a message like this before in real life?

Repeat questions for all 12 instant messages.

A.5. Post-task questions

Thanks for walking through those messages. Now I'd like to ask a few final questions to get your overall thoughts.

10) You saw several different warnings/notices while completing this task. How helpful were the warnings in making your decisions?
    a) **Follow-up:** Are there any warnings that stood out as particularly helpful to you? Any that were particularly unhelpful?
    b) **Follow-up:** Was there anything missing from the warnings or messages that would've helped you decide more easily?
    c) **Follow-up:** Do you think warnings like these would be helpful in real life? Why or why not?
11) Are there any features (other than warnings) that you think your phone should have to help deal with instant messages from unknown senders?
12) Is there anything else you'd like to share about this experience or about instant messaging in general?

Appendix B. Stimuli

All formatting within warning text was generated by ChatGPT. Sender identifiers are described for privacy purposes.

B.1. Fake Amazon Account Alert

Sender: Email address that contains the word "Amazon" followed by random text and an unrelated domain name.
Text: Unusual Activity : We've suspend your Amazon account due to unusual activity. Login attempt blocked from : Mozilla Firefox 8.7 [168.254.81.174 - Turkey] Our system has suspend your Amazon account for security reasons. To unlock your Amazon account, please verify with link below: http://www.xprswebsite.com/free/userf111/redirf111/?=5826865 You need to take action within 2 days before account will be suspended. Regards, Amazon Teams
Long Contextual Warning Text: **Poor grammar**, **a suspicious sender email**, and **a non-Amazon URL** suggest this message is a phishing attempt. **Do not click the link**—go directly to the official Amazon website if you're concerned about your account.
Short Contextual Warning Text: Do not click links; verify directly via official app or website.

B.2. Fake Bank of America Alert

Sender: US number with mid-western area code.
Text: BofA Alert: We 've placed a hold on your debit/atm card due to some unusual activity, To recover Visit https://profib04.com and verify your identity
Long Contextual Warning Text: The message includes a **suspicious URL that is not affiliated with Bank of America** and has **grammar errors** (e.g., "We 've"). **Do not click the link**—contact your bank through official channels to verify any issues.
Short Contextual Warning Text: Don't click links; contact your bank directly.

B.3. Fake Delivery Alert

Sender: US number with west-coast area code
Text: We have detected an issue with your recent shipment; invalid shipping information. Access http://uspsdeliveryrd.ga/ to update your shipping address.
Long Contextual Warning Text: The message uses a **fake USPS-related URL** (".ga" domain) and **creates urgency to trick you** into clicking. **Do not open the link**—check shipment status directly on the official USPS website.
Short Contextual Warning Text: Avoid clicking links; verify shipment status on official site.

B.4. Employment Scam

Sender: US number with Pittsburgh, PA area code
Text: Hello, are you looking for a part-time job now? We provide sales growth services for Amazon, and you can get a stable income of $30-200 in just one hour a day. You do not need to pay any deposit or membership fee, regardless of gender, 23-68 years old, and the Commission is paid every day. If you are interested, please add my telegram. t.me/JRE7708 or WhatsApp: [redacted UK phone number]
Long Contextual Warning Text: The message promises **unrealistically high pay for little work**, includes **unsolicited contact links (Telegram, WhatsApp)**, and uses **vague job details**—common signs of job scams. **Do not respond or click any links**; legitimate employers do not recruit this way.

Short Contextual Warning Text: Do not respond or share personal info; offers like this are often scams.
Note: The phone number was arbitrarily selected for this message, as the SmishTank dataset did not include a phone number for this message.

B.5. Fake T-Mobile Giveaway

Sender: US number with west-coast area code
Text: This is a gift to express our gratitude for your exceptional payment record. HAPPY 2025! T Mobile POSC9.XYZ/YgSpZZjZoh
Long Contextual Warning Text: The message contains a **suspicious, shortened-looking URL** and **unsolicited reward claims**, which are classic signs of phishing. **Do not click the link**—check directly with T-Mobile through official channels if you receive unexpected offers.
Short Contextual Warning Text: Ignore unexpected gifts and don't click links.

B.6. IRS Arrest Threat

Sender: US Phone number with Washington D.C. area code.
Text: WARNING: (Criminal Investigation Division) I.R.S is filing lawsuit against you, for more information call on [anonymous US number with south-eastern area code] on urgent basis, Otherwise your arrest warrant will be forwarded to your local police department and your property and bank accounts and social benifits will be frozen by government.
Long Contextual Warning Text: This message uses **threatening language, urgent legal claims**, and **spelling errors** (e.g., "benifits")—tactics commonly used in government impersonation scams. **Do not call the number**; the IRS does not contact people this way.
Short Contextual Warning Text: Do not call or respond; official agencies don't threaten by text.

B.7. FedEx Notification

Sender: 5-digit short code
Text: FedEx: A package from Aarke was delivered 05/13 at 12:08P. See photo and rate delivery: fedex.com/t/[tracking number]/en_US Reply HELP for help. STOP to cancel
False Positive Warning Text: This message is likely a scam because it urges you to click a suspicious link pretending to be FedEx; **do not click links or provide personal information** and verify delivery status directly through official FedEx channels.

B.8. Healthcare Reminder

Sender: 5-digit short code
Text: MyUPMC: You have an upcoming appt. at UPMC. Log in to view appt details with MyUPMC. https://myupmc.upmc.com/appointments Reply XRemind to opt out
False Positive Warning Text: This message pretends to be from UPMC but contains a link that could steal your login; **do not click links or enter your credentials**—always access your account directly through official websites.

B.9. Amazon OTP

Sender: 5-digit short code
Text: Amazon: Your code is 134091. Don't share it. If you didn't request it, deny here https://amazon.com/a/c/r/[25 character base 64 string]
False Positive Warning Text: This message may be phishing for your Amazon credentials—**do not click the link or enter any codes**, and go to Amazon's site directly if you didn't request a login.

B.10. Overstock.com Promotion

Sender: 5-digit short code
Text: Overstock Day is here with our best home deals of the year! TODAY ONLY get 20% off + 2x rewards + free shipping!* https://o.bttn.io/0nvkx2i4Cqp
False Positive Warning Text: This message uses urgent sales language and a suspicious link to lure clicks—**avoid clicking and shop only through official retailer websites**.

B.11. Bank of America PIN Alert

Sender: 5-digit short code
Text: BofA Security: Debit card ending in [4 digits] was declined because a PIN wasn't entered. Please try again with your PIN. Reply STOP to end texts.
False Positive Warning Text: This message impersonates Bank of America to create urgency—**do not reply or share your PIN**, and contact your bank directly through verified channels.

B.12. Political Solicitation

Sender: US phone number with area code located in the same state as Pittsburgh, PA
Text: Hey Bob, OnePA here! Mayor Ed Gainey is standing up to Trump & billionaires trying to buy Pittsburgh's election on May 20th! Read more here to see how Mayor Ed Gainey is fighting back! https://onepa.org/pittsburgh/?id=403V Stop to End
False Positive Warning Text: This message uses **emotional political language and a suspicious link** to manipulate you—avoid clicking links and verify campaign information through trusted sources.

Appendix C. Meta-Review

The following meta-review was prepared by the program committee for the 2026 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

C.1. Summary

The paper performs an interview study with 20 participants that seeks to understand the user perception of SMS messages with scam warnings. The participants are presented with different warning types (none, detailed, high-level only), and their reasoning is evaluated. The analysis leads to interesting findings regarding user perception and factors that impact the user's reasoning and decision-making for legitimate as well as scam SMS messages.

C.2. Scientific Contributions

- Identifies an Impactful Vulnerability.
- Provides a Valuable Step Forward in an Established Field.

C.3. Reasons for Acceptance

1) The paper provides an in-depth analysis of user reasoning in the context of SMS-based spam and uncovers valuable insights.
2) The study includes both current warnings from Google Messages and hypothetical GAI-driven designs. As mobile OS vendors (mainly Android-based phones) move toward integrating local LLMs, understanding how users interpret AI-generated security advice is timely and critical.
3) Several of the insights are novel and will inform future work in the field. In particular, that users invent reasons to agree with incorrect AI warnings is a critical vulnerability in the deployment of GenAI for security. It challenges the prevailing assumption that "more explanation is better." The fact that the study population is predominantly educated, tech-literate users means that the findings present an upper bound in reasoning and decision-making abilities of the general population, as the ability to parse complex "contextual explanations" might be significantly lower in a general population sample.

C.4. Noteworthy Concerns

1) The study setup forces users to stare at messages and warnings. In reality, SMS is an interrupt-driven mechanism often consumed in seconds. The finding that users "liked" the detailed evidence might be an artifact of the interview setting, where they felt compelled to read everything.
2) The paper only briefly addresses possible effects of warning fatigue or "habituation" during the study. This could also have been a contributing factor to why people were inclined to believe the false positive was actually a threat.
3) Although the paper aims to study how users reason about messages in the presence of warnings, the design and analysis do not consistently isolate the role of warnings themselves. In particular, it is unclear how the different types of warnings were systematically identified.
4) Several key constructs (e.g., conditional responses) are not clearly defined, making it difficult to interpret the findings. Participants frequently provided multiple or context-dependent actions, but the paper does not explain in depth what these responses are intended to capture or how they should be interpreted analytically. This raises the possibility that the prevalence of conditional or contradictory recommendations reflects aspects of the study design (e.g., message presentation or response options) rather than users' natural decision-making.
5) The studied sample is skewed toward educated/tech-literate users, which may impact the findings.

C.5. Response to the Meta-Review

We thank the reviewers for their constructive feedback. We acknowledge the limitations of our study design, including that the lab setting does not fully reflect real-world use and that explanation text may be ignored in practice; our results thus reflect an upper bound of warning attentiveness (Concern 1). While the repetitive nature of the task may confound some results, we believe it is unlikely that habituation affected participants' assent to false positive warnings. Habituation refers to users ignoring or "tuning out" a warning after long-term, repeated exposure because the brain becomes less reactive to repeated stimuli [53]. Our participants, instead, heeded the warnings (Concern 2). Qualitative interviews provided deep insight into participants' reasoning processes; however, this makes it difficult to isolate the specific effect of each design parameter. We based our warning selection on recent related work [6] and production warnings used by Android to ensure systematic relevance (Concern 3). As discussed in the text and the extended Appendix, conditional responses are those in which participants indicated they might take multiple actions based on context not provided in the message or the interview instructions. We argue that the prevalence of conditional or contradicting responses reflects the situational heuristics users apply in real-world communication rather than being solely an artifact of the study design (Concern 4). Finally, we acknowledge the skew in our sample towards highly educated, tech-literate people, and that users with different backgrounds may experience these warnings differently (Concern 5).