Can Large Language Models Improve Phishing Defense? A Large-Scale Controlled Experiment on Warning Dialogue Explanations

Federico Maria Cau[a], Giuseppe Desolda[b,*], Francesco Greco[b], Lucio Davide Spano[a], Luca Viganò[c]

[a] *University of Cagliari, Via Ospedale 72, Cagliari, 09124, Italy*
[b] *Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, Bari, 70125, Italy*
[c] *King's College London, Bush House, Strand Campus, 30, Aldwych, London, WC2b 4BG, UK*

Abstract

Phishing has become a prominent risk in modern cybersecurity, often used to bypass technological defences by exploiting predictable human behaviour. Warning dialogues are a standard mitigation measure, but the lack of explanatory clarity and static content limits their effectiveness. In this paper, we report on our research to assess the capacity of Large Language Models (LLMs) to generate clear, concise, and scalable explanations for phishing warnings. We carried out a large-scale between-subjects user study (N = 750) to compare the influence of warning dialogues supplemented with manually generated explanations against those generated by two LLMs, Claude 3.5 Sonnet and Llama 3.3 70B. We investigated two explanatory styles (feature-based and counterfactual) for their effects on behavioural met-

*Corresponding author
Email addresses: federicom.cau@unica.it (Federico Maria Cau), giuseppe.desolda@uniba.it (Giuseppe Desolda), francesco.greco@uniba.it (Francesco Greco), davide.spano@unica.it (Lucio Davide Spano), luca.vigano@kcl.ac.uk (Luca Viganò)
URL: https://orcid.org/0000-0002-8261-3200 (Federico Maria Cau), https://orcid.org/0000-0001-9894-2116 (Giuseppe Desolda), https://orcid.org/0000-0003-2730-7697 (Francesco Greco), https://orcid.org/0000-0001-7106-0463 (Lucio Davide Spano), https://orcid.org/0000-0001-9916-271X (Luca Viganò)

rics (click-through rate) and perceptual outcomes (e.g., trust, risk, clarity). The results indicate that well-constructed LLM-generated explanations can equal or surpass manually crafted explanations in reducing susceptibility to phishing; Claude-generated warnings exhibited particularly robust performance. Feature-based explanations were more effective for genuine phishing attempts, whereas counterfactual explanations diminished false-positive rates. Other variables such as workload, gender, and prior familiarity with warning dialogues significantly moderated warning effectiveness. These results indicate that LLMs can be used to automatically build explanations for warning users against phishing, and that such solutions are scalable, adaptive, and consistent with human-centred values.

Keywords: Phishing, Warnings, Human factors, Alert, LLMs, Explanations

1. Introduction

Phishing attacks are among the most widespread and harmful cyber threats, leveraging human behaviour instead of technical weaknesses (Khonji et al., 2013; Prakash et al., 2010). There is a vast number of automatic filtering techniques (e.g., blacklists, machine learning-based detection). Still, these automatic defence systems are pretty far from perfect (Basit et al., 2021), and there is the strong possibility of false positives and, even more importantly, false negatives, which entail that users might receive harmful material.

To overcome this issue, most defensive systems engage the human in the decision-making process by presenting users with *warning dialogues*, allowing them to make the final decision and possibly prevent a phishing attack (Desolda et al., 2019). Although warning dialogues are commonly used, they work well only when describing the risk and influencing the decision to act upon (IBM, 2024). Most dialogues used today are passive and mostly static, with no clear explanations as to why a particular message is harmful (Egelman et al., 2008; Wu et al., 2006).

In the last few years, researchers working in *Human-Centred Security* and *Explainable AI (XAI)* have investigated whether feature-based explanations might be effective in mitigating phishing attacks (Guidotti et al., 2019b; Gunning and Aha, 2019). The clear advantage of such explanations is the reduction of susceptibility to phishing, but they are hand-made, expensive to maintain, and might not easily adjust to the changing circumstances of

the threats (Desolda et al., 2023; Buono et al., 2023; Greco et al., 2025). Other ways of generating explanations have also been considered in general, such as the use of counterfactuals (Guidotti et al., 2019a) or the use of *Large Language Models (LLMs)*, but they have not been investigated extensively in the specific field of cybersecurity (Xu et al., 2024).

To bridge these gaps, we have carried out a large-scale, between-subjects user study ($N = 750$). We investigated how well phishing warning dialogues with explanations can be generated with LLMs. We assessed the result of using a commercial closed-weight model (Claude 3.5 Sonnet) and an open-weight alternative (Llama 3.3 70B). In addition, we considered two styles of explanation, i.e., feature-based and counterfactual explanations. A baseline with feature-based, manually-written explanations has also been considered (Desolda et al., 2023; Greco et al., 2025). The effectiveness of the the different experimental conditions have been evaluated by considering both measures of behaviour (e.g., click-through rate on phishing links) and user perceptions (e.g., clarity, trust, perceived risk), and modelled the degree to which individual factors, including cognitive style and knowledge of cybersecurity issues, influence results.

We found that explanations generated by LLMs can be as effective as those produced manually; specifically, explanations generated by Claude were better than those crafted manually. Concerning the type of explanation, counterfactual ones have an interesting power in ensuring false positive cases, and feature-based ones have certain benefits in true positive cases. Moreover, user-related factors such as gender, familiarity, and mental workload significantly influence the effectiveness of the warning.

The remainder of this paper is structured as follows. Section 2 reviews related work on phishing defences, warning dialogue design, and explainable AI in cybersecurity. Section 3 presents the experimental design and the explanation generation process. This section also reports the formalisation of the four research questions this study aims to answer. Section 4 details the study procedure and methodology. Section 5 reports the quantitative and qualitative results addressing our four research questions. Section 6 discusses threats to validity. Section 7 synthesises the findings as a set of lessons learned. In conclusion, Section 8 summarises the paper and emphasises potential areas for future research.

2. Related work

2.1. Warning dialogues for Phishing Defence

The complete elimination of phishing through automated filtering methods such as blocklists (Gupta and Kumaraguru, 2014) and AI-based methods Basit et al. (2021) is currently unfeasible. No existing tools can detect phishing content, such as websites or emails, with perfect accuracy (El Aassal et al., 2020; Mehdi Gholampour and Verma, 2023). Furthermore, attempts at reducing the number of false negatives (i.e., phishing emails that go undetected) tend to increase the number of false positives (i.e., legitimate emails misclassified as phishing) in classification tasks (Saxena, 2018). This eventually leads to filtering out genuine, relevant emails, potentially disrupting user productivity.

The solution currently adopted in most of the systems is not to automatically filter suspicious content classified as phishing with lower confidence; instead, these emails are presented to users alongside a warning dialogue that alerts them about possible threats while leaving the final decision to the users (Kumaraguru et al., 2010). Despite warning dialogues, many users fail to grasp their significance or meaning, so phishing attacks remain highly effective (IBM, 2024).

One key problem of everyday warning dialogues is their passive behaviour. Popular email clients such as Google Gmail implement static toolbars above the email content. Warnings like these are largely ineffective, as users often overlook or ignore them (Egelman et al., 2008; Wu et al., 2006). On the other hand, warnings that implement an active behaviour, i.e., interrupt user interactions and demand attention, have been proven to be significantly more effective (Wogalter et al., 2002; Petelka et al., 2019; Buono et al., 2023; Egelman et al., 2008; Greco et al., 2025). However, such warnings are currently implemented only in web browsers (Desolda et al., 2019), despite the benefits of active behaviour having also been demonstrated for warnings shown in email clients (Greco et al., 2025; Buono et al., 2023).

Another limitation of warnings is the phenomenon known as *habituation*. This behaviour occurs when users are repeatedly exposed to the same warning stimulus, even under varying risk conditions, leading to decreased attentional responses and increased likelihood of bypassing the warning (Akhawe and Felt, 2013; Kim and Wogalter, 2009). To mitigate this problem, *polymorphic* warnings were demonstrated to be significantly more resistant to habituation than static warnings (Anderson et al., 2015). Such warnings are called

polymorphic since they dynamically change their appearance or content to provide varied visual stimuli. Nonetheless, they are still not implemented in browsers or email clients despite their advantages (Desolda et al., 2019). Only two studies have explored using polymorphic warnings paired with tailored explanations to enhance their effectiveness (Desolda et al., 2023; Buono et al., 2023). Desolda et al. found that including manually designed explanations of malicious features in warning dialogues in web browsers led to warnings that were more familiar and understandable than those in the most common browsers. Buono et al. studied warnings with explanations in the context of email clients and found that active warnings with explanations based on phishing features largely outperformed contextual warnings without explanations. This study confirmed the importance of interrupting user interaction flow to focus entirely on the message, similar to active warnings shown in browsers. Additionally, the study demonstrated the importance of explaining to users why content is malicious, helping them make better-informed decisions.

A further issue that compromises the warnings' effectiveness is the absence of detailed explanations about the phishing risk (Bravo-Lillo et al., 2011). Ideally, warnings should communicate clearly why specific content is dangerous and outline the potential consequences of ignoring it. However, many current warnings provide vague messages that fail to identify the malicious elements within an email or website. This places the burden of risk assessment on the users, who often lack cybersecurity expertise or additional information to make informed decisions (Desolda et al., 2021). By incorporating explanations into warnings, the users' understanding of the threat can be improved, addressing knowledge gaps, and potentially increasing compliance and trust in the system (Bravo-Lillo et al., 2011; Vilone and Longo, 2021).

The use of explanatory warnings in web browsers has been explored in (Desolda et al., 2023). The authors identified seven website features that can be intuitively explained to users and produced two explanation messages for each. These messages were refined through an iterative process to optimise text readability and sentiment. Moreover, the explanation messages followed a well-defined structure grounded in warning theory (Bauer et al., 2013), which led to the identification of a template for generating explanation messages: "Feature description + Hazard Explanation + Consequences of not complying with the warning". The authors empirically validated with 150 users the adoption of warnings with explanations, compared with warn-

ings without explanations. Results showed that the proposed warnings were clearer and more effective than those implemented in modern browsers, which lack explanations. Greco et al. (2025) explored a similar approach in the context of email clients, combining active warning dialogues with embedded explanations. The findings indicated that this combination of active behaviour and explanation significantly outperformed the state-of-the-art email client warnings proposed by Petelka et al. (2019).

As security software engineers, we must understand the underlying decision-making algorithm to generate explanations. In the case of AI-based phishing filtering methods, explanations can be produced by employing *eXplainable AI (XAI)* techniques (Guidotti et al., 2019b; Gunning and Aha, 2019).[1] These allow one to explain the behaviour and decisions of AI models, which are often black-box, and thus not directly interpretable even by AI engineers. With XAI techniques, for example, the system can provide a rationale for the reasons that led to the automatic classification of an email as phishing. Explanations in warning dialogues can contain such a rationale to improve the transparency of the AI system. Let us explore these aspects in the following subsection.

2.2. Human-Centred Explainable AI for Phishing Attacks

Integrating XAI solutions into cybersecurity systems is essential to maintain user trust and confidence. This underscores the importance of user-centred experimental evaluations that balance security, usability, and resilience to adversarial attacks (Charmet et al., 2022; Rjoub et al., 2023; Sarker et al., 2024; AL and Sagiroglu, 2025). In this regard, the broader *Human-Centred XAI (HCXAI)* literature offers valuable insights into multiple explanation types and their effects on users. The most prevalent explanations in AI-assisted decision systems are the feature-based ones (Lai et al., 2023), clarifying *why* an AI system produced a specific decision by highlighting the most important features driving the decision. Feature-based explanations might enhance user understanding, uncertainty awareness, trust calibration, and decision accuracy, though they can sometimes foster over-reliance on AI (Zhang et al., 2020; Wang and Yin, 2021; Chen et al., 2023; Ma et al., 2023; Fok and Weld, 2024; Cau and Spano, 2025). Counterfactual

[1]See (Viganò and Magazzeni, 2020) for a discussion on the relationship between explainable cybersecurity and explainable AI.

explanations constitute another, less explored style: they provide not only *why* an AI system reached a specific decision, but also *how* it is possible for users to alter input features to achieve a potential desired outcome by presenting contrastive *"what-if"* statements. While they are mainly used to offer actionable recourse when negative decisions occur—e.g., loan denial—by indicating necessary feature changes for a positive result—e.g., increasing income for approval—(Koh et al., 2024; Verma et al., 2024; VanNostrand et al., 2024; Upadhyay et al., 2025), they have also demonstrated improvements in helpfulness, decision alignment, and accuracy, sometimes matching or exceeding feature-based methods, in AI-assisted decision-making tasks (Scharowski et al., 2023; Celar and Byrne, 2023; Cau et al., 2023; Teso et al., 2023; Lee and Chew, 2023; Gentile et al., 2025).

Regarding XAI in cybersecurity and phishing in particular, only a small number of proof-of-concept studies have assessed explanation effectiveness with real users (Desolda et al., 2023; Greco et al., 2025). For example, Desolda et al. (2023) found that adding a feature-based explanation to a phishing warning dialogue made the warning more understandable and familiar, which may deter users from visiting malicious sites. In a similar study, Greco et al. (2025) compared warnings with and without feature-based explanations and the effect of showing the warning before versus after opening an email. They showed that adding explanations improved users' comprehension of the phishing threat and, when the warning appeared after an email opening, led to a measurable reduction in click-through rate (CTR) despite the additional effort required by participants. However, these investigations have focused exclusively on manually crafted feature-based explanations, leaving the generation of feature-based or alternative explanation types via large language models largely unexplored (Sarker et al., 2024).

This work addresses these gaps through a user study in a phishing email scenario. We compare LLM-generated, feature-based, and counterfactual explanations to manually crafted ones within warning dialogues, and evaluate their impact on users' CTR on potentially harmful links and subjective perceptions of the warning dialogue.

3. Approach

In this study, we focus on the impact of two factors on user decision-making in phishing-warning dialogues: (i) the style of the explanation messages in the warning dialogue and (ii) the LLM employed to generate the

explanation message.

Regarding the explanation style, the first approach we considered is based on *feature-based explanations*, a widely adopted approach in HCI systems for communicating the most influential features contributing to an AI decision (Lai et al., 2023; Fok and Weld, 2024). In our context, these explanations highlight specific elements of an email, such as a masked link or a suspicious sender address, that led the AI to classify it as phishing. Complementing this standard, we also include *counterfactual explanations*, which offer an alternative, theoretically grounded approach to supporting user understanding. Rather than simply highlighting influential features, counterfactuals offer contrastive explanations answering implicit *"why not?"* questions by describing how small, meaningful changes would alter the AI's prediction (Lee and Chew, 2023; Koh et al., 2024; Gentile et al., 2025). This explanation technique clarifies decision boundaries and supports actionable recourse, particularly relevant in high-stakes settings such as fraud detection or security. In our phishing warning dialogues, we embed these explanations to show how the email could have avoided being flagged, for example, stating that the email would be considered safe if the link label matched the actual destination URL (e.g., "protect your account" linking to facebook.com/protect-account).

The second factor concerns the LLM model used to generate the explanation message. While past studies often relied on manually written explanation messages demonstrating their effectiveness (Desolda et al., 2023; Greco et al., 2025; Buono et al., 2023), the growing availability of LLMs might enable automatic generation of tailored content in both explanation styles. If explanations produced by LLMs prove as effective as manually crafted messages, which are both time-consuming to create and must be updated whenever new features are introduced, then LLMs could fully automate their generation. This study considered explanations generated by an open-source model (Llama 3.3) and a closed-source commercial model (Claude 3.5 Sonnet). These two LLMs were considered to evaluate whether a commercially available LLM (Claude) could provide better explanations than an open-weight solution such as LLama 3.3. This aspect is critical, as an organization could have strict privacy requirements that prohibit the sharing of sensitive data (i.e., emails) with third parties like Claude (Shashidhar et al., 2023; Panwar et al., 2024; Zhang et al., 2024; Adams et al., 2024; López Espejel et al., 2025).

Building on these premises, we aim to address the following research questions:

- **RQ1** Can LLMs generate explanations as effective as manually-generated ones in protecting users in warning dialogues?
- **RQ2** Are feature-based explanations as effective as counterfactual explanations in warning dialogues?
- **RQ3** How do users perceive LLM-generated warnings?
- **RQ4**: How do user-related factors affect click trough rate (CRT) of phishing links?

In the following subsections, we detail i) how we designed the warning dialogues, ii) how the two explanation styles have been integrated into the warning dialogues, and iii) the selection process of the two LLMs considered in this study.

3.1. Warning dialogues design

The phishing warning dialogues explored in this study have the same structure to isolate the effects of the explanation message, depicted in Figure 1. A dialogue opens with an attention-grabbing header flagging the link as potentially dangerous, followed by a summary that attributes this judgment to the AI system. The main part of the interface is the explanation message, presented as a short paragraph embedded within the dialogue box. The bottom part contains buttons to get back to the email visualization or, with a less prominent presentation, a link to continue browsing the destination. All messages have the same typographic style and layout to control for visual biases. This design ensured that any differences in user response could be attributed to the nature of the explanations rather than variations in formatting.

3.2. Explanation Styles

The explanation message included in the main part differs according to the considered styles. Feature-based explanations emphasize observable cues extracted from the email, such as sender identity mismatches or suspicious links. The left part of Figure 4 shows sample explanations for this style, realised through declarative statements pointing to concrete phishing indicators. All messages share a common structure consisting of three sentences, inspired by prior work Desolda et al. (2023), which builds on

Figure 1: Anatomy of the warning dialogue used in the study. The top part (A) includes a brief warning working as the header of the box, the main part (B) presents the explanation message, while the footer (C) shows controls to get back to the email visualization or to proceed with browsing.

the Communication-Human Information Processing (C-HIP) model (Wogalter, 2018) and previous studies on effective warning communication (Bauer et al., 2013). The same structure was explicitly adopted when prompting the LLMs to generate the warnings, ensuring consistency across human- and AI-authored messages.

In contrast, counterfactual explanations present a hypothetical scenario in which the email would not have been flagged. These explanations reframe the AI's decision as contingent on a small change, inviting users to consider alternatives. The examples adopt a conditional format, underscoring a key difference in framing: while feature-based explanations justify the current decision, counterfactuals highlight what could have made the outcome different. The right part of Figure 4 shows sample warnings using a counterfactual style.

3.3. Generating the explanations

The other difference in the experimental conditions is the LLM used to generate the warning text. We selected the two LLMs to create explanations for this study through a rigorous evaluation process. Notably, 13 different LLMs (six commercially available, seven open-source) were included in our evaluation process as possible candidates. As grounds for this evaluation, we considered the three emails (two true positives and one false positive) whose

links, when clicked, generated a warning dialogue during the experiment. To evaluate each LLM, we followed this process:

- Following the approach and the prompts proposed in Desolda et al. (2025), for each of the three emails, the LLM was prompted to generate an explanation message to be included in the warning, given the email body and one of the features to explain. We considered the following three email features, which were found to be helpful to users when making decisions regarding phishing content (Greco et al., 2025; Desolda et al., 2025), i.e., (i) the Top-Level Domain of the URL in the email is misplaced (i.e., "www.amazon.com.cz"); (ii) the URL in the email is an IP address; (iii) the link shown in the email and its actual destination mismatch (e.g., "click here" is shown instead of "www.facebook.com/...") — this feature was found in the false positive email.

- For each of the three emails, we generated a counterfactual explanation message through a modified version of the prompt proposed in Desolda et al. (2025).

- Two researchers qualitatively assessed the six explanation messages to highlight whether an explanation was not generated correctly (e.g., by reaching the token limit early) or contained hallucinations. In this case, the previous step was repeated, and the produced explanation was analysed again for a maximum of three times. If every attempt did not lead to a coherent explanation, the LLM was marked as inadequate for this task. This process led to the discarding of Llama 3.2 1B-Instruct.

- The explanations underwent an automatic measurement of text readability with the FK (Flesch-Kincaid) Grade Level (Kincaid, 1975) and of text understandability with the SMOG (Simple Measure of Gobbledygook) Index (Laughlin, 1969).

- Explanations with low SMOG Index and FK Grade values were preferred as they indicate more readable and understandable text. Feature-based explanations' readability metrics are shown in Figure 2, while counterfactual-related ones are shown in Figure 3.

To select the best commercial and open-weight LLMs, we initially selected a set of LLMs with the lowest, thus best, SMOG Index and FK Grade for

both feature-based and counterfactual explanations. From this analysis, we selected Claude 3.5 Sonnet and Gemini 1.5 Flash as the best candidates for commercial LLMs, while all the Llama LLMs are candidates for the open-weight solutions. Then, two experts in HCI and cybersecurity performed a manual inspection of the generated explanations. This analysis removed Gemini 1.5 Flash because it generated explanations that were either too generic or contained hallucinations, especially for the counterfactual style (e.g., for the Amazon phishing email it generated an explanation reporting "[..] *A safe email address wouldn't have strange parts in its name.* [..]"). Moreover, Llama 3.2 90B (as well as Gemini 1.5 Pro) was excluded, as it could not be forced to produce an explanation for the false-positive email. Llama 3.1 8B was a good candidate for feature-based explanations, although readability and understandability metrics were mediocre in the context of counterfactual explanations. Moreover, the generated feature-based explanation for the Amazon email contained a hallucination ("[..] *It looks like an Amazon link, but the '.com' name is typically found in other types of companies.* [..]"). Therefore, considering the qualitative assessment and the average metrics for both explanation styles, we selected Llama 3.3 70B as it represents a good trade-off between the resources needed to deploy this LLM and the quality of the explanations. The explanations produced by Claude 3.5 Sonnet and Llama 3.3 70B were then featured in the user study as described in Section 4.1.

3.4. Claude vs Llama

While selecting the two LLMs to be used in the study, we identified some general characteristics of the generated explanation text.

Claude tends to produce detailed and elaborate explanations. Its feature-based explanations frequently embed causal logic and downstream outcomes: for instance, a warning is issued that supplying data on a spoofed website risks account takeover. Within counterfactual scenarios, Claude employs conditional phrasing ("The email would have been safe if. . . ") while offering idealized reference links illustrating what a genuine message might look like. This is a continually organized discussion that incorporates technical pointers and a clear emphasis on the user experience that may occur.

Instead, Llama delivers explanations that are somewhat more concise and technical in tone. Its feature-based outputs highlight specific anomalies (e.g., numeric IP addresses, suspicious domain extensions) with minimal elaboration. The counterfactuals follow a similar structure to Claude's but are gener-

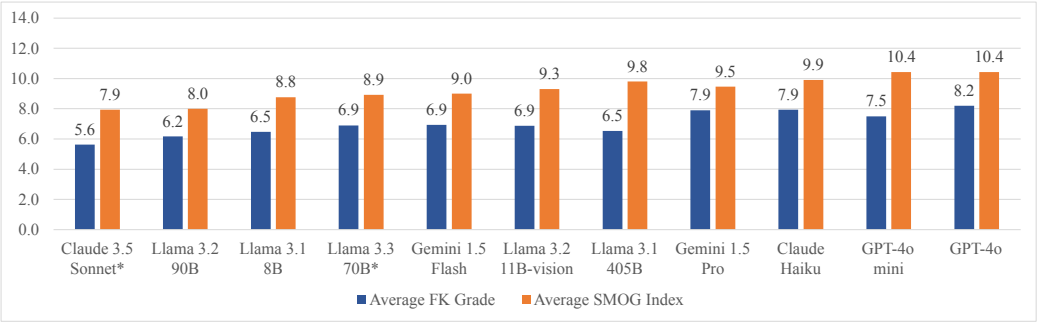

Figure 2: Average understandability and readability metrics for the feature-based explanations generated by the different LLMs - the lower, the better. The asterisks indicate the LLMs that were finally chosen.

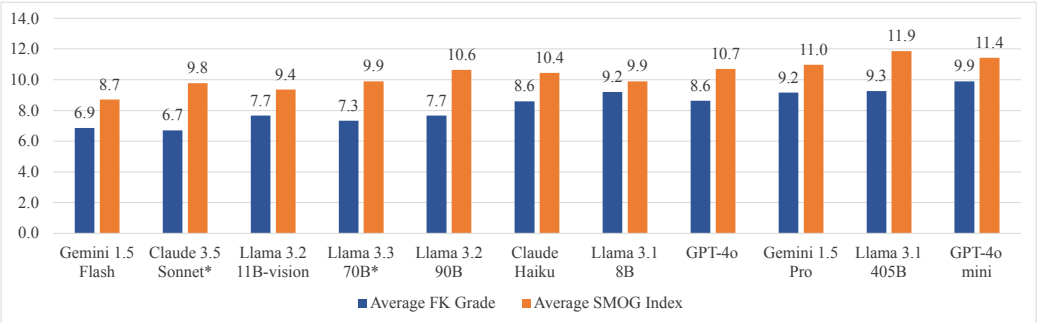

Figure 3: Average understandability and readability metrics for the counterfactual explanations generated by the different LLMs - the lower, the better. The asterisks indicate the LLMs that were finally chosen.

ally more direct and less descriptive, focusing on the change required to alter the classification without extending into consequences. Llama also tends to use simpler language and more neutral phrasing, which may be more mechanical but still effective. These stylistic differences are illustrated in Figure 4. The top row includes examples generated with Claude 3.5 Sonnet, while the bottom row includes examples generated with Llama 3.0 70B.

Figure 4: Four versions of the warning dialogue for the same deceptive link. Rows correspond to the text generated by the two LLMs employed in this study (Claude 3.5 Sonnet and Llama 3.0 70B), while columns refer to explanation styles (feature-based and counterfactual).

4. Methods

In this section, we present a controlled experiment that compares the effectiveness of various explanations displayed in warning dialogues to protect users from phishing emails. The study specifically focused on evaluating explanations based on two factors we identified and discussed in the previous section: the *explanation style* (either feature-based or counterfactual) and the *LLM model* used to generate the explanation message (LLama 3.3 or Claude 3.5 Sonnet). We also considered a baseline warning including manually-generated messages that was empirically demonstrated to outperform all the existing warnings for phishing attacks (Greco et al., 2025). The study procedure and materials (i.e., warning design and emails included) were identical to the study in Greco et al. (2025) to make the comparison with the baseline as fair and rigorous as possible.

4.1. Study Design

This study adopted a between-subjects design with the explanation style and LLM used to generate the explanations as the independent variables. Five total between-subject levels were considered, four experimental warnings plus a baseline:

1. **Llama Feature-based**: warnings containing feature-based explanations generated by the open-weight LLM Llama 3.3 70B.
2. **Llama Counterfactual**: warnings containing counterfactual explanations generated by the open-weight LLM Llama 3.3 70B.
3. **Claude Feature-based**: warnings containing feature-based explanations generated by the commercial LLM Claude 3.5 Sonnet.
4. **Claude Counterfactual**: warnings containing counterfactual explanations generated by the commercial LLM Claude 3.5 Sonnet.
5. **Baseline**: warnings that included feature-based explanations, which were manually written by experts (existing dataset - from Greco et al. (2025))

To improve the external validity of the study, for each experimental condition, we generated three explanations related to three different features. The complete list of explanation messages contained in each warning is reported in Appendix A.

4.2. Participants

A total of 750 participants were involved, with equal allocation (n = 150) to each condition. The manually generated feature-based explanation condition included existing data (collected separately, n = 150), while the remaining 600 participants were recruited specifically for this study through the Prolific online platform[2]. The inclusion criteria for participants were as follows: i) fluency in English, ii) an even split between men and women (50-50), and iii) 18 or more years of age. A total of 61 participants either failed attention checks or were identified as inattentive and were thus excluded. To balance participation across the experimental conditions, we recruited 61 more participants to reach a final sample of exactly 750 participants. The final sample had an average age of 32.98 years (sd = 11.01) and reported

[2] `https://www.prolific.co`

spending 7.94 hours per day on the Internet (sd = 3.95). Participants were 375 males and 375 females. The participation in the study lasted approximately 20 minutes, and participants were rewarded with £3.00, in line with Prolific's recommended participation fee of £9.00/hour.

4.3. Materials

We developed a web platform with Laravel 9[3] to conduct the study online. A total of 14 emails were included in the study: 2 of them were (harmless) phishing emails, while the remaining 12 were genuine ones designed starting from legitimate services/companies.

A warning dialogue, such as that shown in Figure 1, was shown to users who clicked on a phishing link. To simulate the case in which the system mistakenly classifies a genuine email as phishing, i.e., a false positive situation, we forced the platform to show a warning also for one of the genuine emails. After the study, each user may have seen a maximum of three warning dialogues.

A full suite of user interactions was recorded in this study: hovering over phishing links, clicking on them, visiting associated URLs, and responding to warnings, which allowed calculating the click-through rate (CTR) for each participant. The CTR was the proportion of accessed suspicious links to the number of displayed warnings, and it was an objective measure of the effectiveness of the warning.

The survey instrument employed by Desolda et al. (2023) was administered to comprehend users' perceptions of the warnings. All the questions and their formats of responses are given in Table 1.

To test for the effects of users' knowledge about basic cybersecurity concepts, a 10-item test questionnaire was used (Olmstead and Smith, 2017).

Finally, a short, 6-item version of the Need for Cognition Scale (NCS-6) (Coelho et al., 2020) was used to measure users' *Need for Cognition*, which is the tendency to engage in and enjoy effortful cognitive activities (Cacioppo and Petty, 1982). Previous work highlighted that individual differences like NFC might significantly impact the effectiveness of AI support and explanations in terms of performance, overreliance on AI, and learning (Cacioppo et al., 1996; Buçinca et al., 2021; Gajos and Mamykina, 2022; Buçinca et al., 2024)

[3] `https://github.com/IVU-Laboratory/llm_warnings_explantions`

Table 1: The questionnaire used to evaluate the warnings (*warning questionnaire*)

#	**Questionnaire item**	**Possible Answers**
1	Did you read the entire text of the warning dialogue?	[yes; partially; no]
2	When you saw the warning dialogue, what was your first reaction?	[free text]
3	I understood the warning dialogue.	[5-point Likert scale, from "Strongly disagree" to "Strongly agree"]
4	I am familiar with this warning dialogue.	[5-point Likert scale, from "Strongly disagree" to "Strongly agree"]
5	I am not interested in this warning dialogue.	[5-point Likert scale, from "Strongly disagree" to "Strongly agree"]
6	Which word(s) did you find confusing or too technical?	[free text]
7	Please rate the extent of risk you feel you were warned about.	[very low risk; low risk; no risk; risky; very high risk]
8	What action, if any, did the warning dialogue want you to take?	[to continue to the website; to be careful while continuing to the website; to not continue to the website; I did not feel anything]
9	What do you think this warning dialogue means?	[free text]
10	Please rate your level of trust in this warning dialogue.	[not at all confident; not very confident; neutral; confident; very confident]
11	What is the first word in this warning dialogue?	[free text]

4.4. Procedure

The study procedure followed the same one in Greco et al. (2025) and is depicted in Figure 7. To prevent priming effects regarding the experiment's true objective, participants were initially deceived about the study's real purpose (Sotirakopoulos et al., 2011; Distler et al., 2021). Participants were told the aim was to evaluate the user experience of a new email client.

After the Prolific task was accepted, the participants were redirected to a web platform where the study was hosted. The (deceptive) purpose of the study, the data usage policies, and the withdrawal right were described on the landing page. According to the ethical guidelines, the participants were asked for their informed digital consent. All the data on participants and interactions were completely anonymized and safely kept on the university servers. Any concerns of the participants could be addressed after the study, and the design was such that it did not cause excessive psychological stress, e.g., phishing scenarios were realistic yet not too alarming.

Once consent was given, all the participants were randomly assigned to four experimental conditions with balanced participation across conditions. Then, participants were requested to follow a case of a fictional user named Alice, who was testing a new email client she had adopted in her workplace. The respondents were asked to interact with the email client by reading messages in the inbox and trying the functionality of embedded links.

The scenario was presented in a banner above the simulated email client interface, which displayed randomised email previews, including the sender's name, title, and subject (Figure 5). Clicking an email revealed its content; phishing links triggered warnings (Figure 6). After each interaction with a phishing message (via link click or by navigating back), participants answered two attention-check questions: "Who was the sender of the email?" and "Were there links in the email?". These checks help identify inattentive respondents, which are especially typical in remote settings (Matsuura et al., 2021).

Upon completing their interactions, participants were debriefed and informed of the actual goal of the study: assessing user responses to warnings triggered by clicks on suspicious links. The debriefing step is a critical ethical practice in studies involving deception (Sotirakopoulos et al., 2011). After the debriefing, participants were asked to reconfirm their consent. Those choosing not to continue would have had their data deleted, though all participants opted to proceed. They then completed a series of questionnaires: first, the warning perception questionnaire from Desolda et al. (2023), followed by the Raw NASA TLX workload assessment (Hart, 1986), a cyberse-

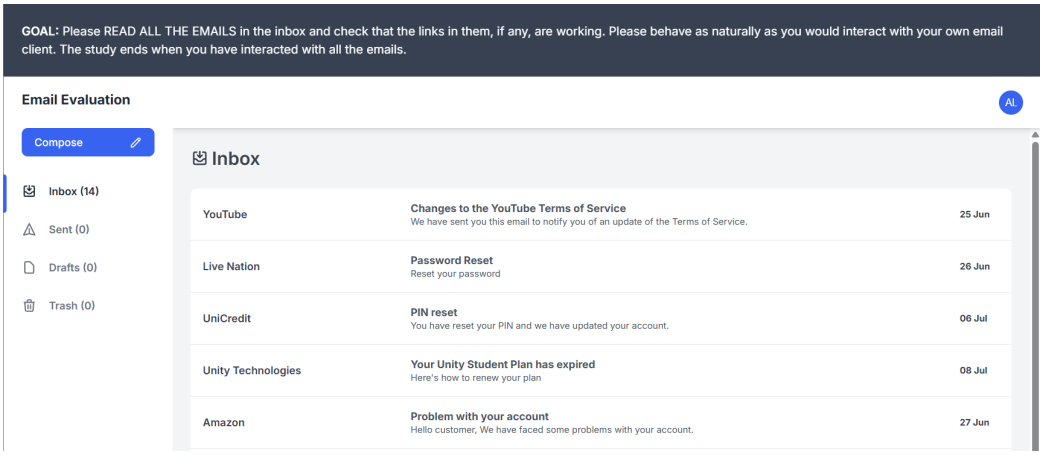

Figure 5: Web platform used for the study.

curity proficiency questionnaire (Olmstead and Smith, 2017), and the Need for Cognition Scale (Coelho et al., 2020). Finally, participants completed a demographic survey stating their age, gender, and average daily internet use. Participants were thanked for their involvement and redirected to Prolific to claim compensation.

The study procedure was approved by the Research Ethics Committee of King's College London, reference number HR-23/24-41353.

4.5. Data Analysis

To answer RQ1 ("Can LLMs generate explanations that are as effective as manually-generated ones in protecting users in warning dialogues?") and RQ2 ("Are feature-based explanations as effective as counterfactual explanations in warning dialogues?"), we applied the Chi-Square test of independence. When differences emerged, Bonferroni adjustment for pairwise comparisons was used as a post-hoc test to analyse differences among individual conditions.

For answering RQ3 ("How do users perceive LLM-generated warnings?"), we employed the Kruskal-Wallis test to analyse differences across the experimental conditions for items 3, 4, 5, 7, and 10 of the warning questionnaire (regarding *understandability*, *familiarity*, *interest*, *felt risk*, and *trust*, respectively) and items of the NASA-TLX questionnaire. In the case of statistical significance, Dunn's test was applied to analyse differences between individual conditions. The Chi-Square test of independence was used to analyse

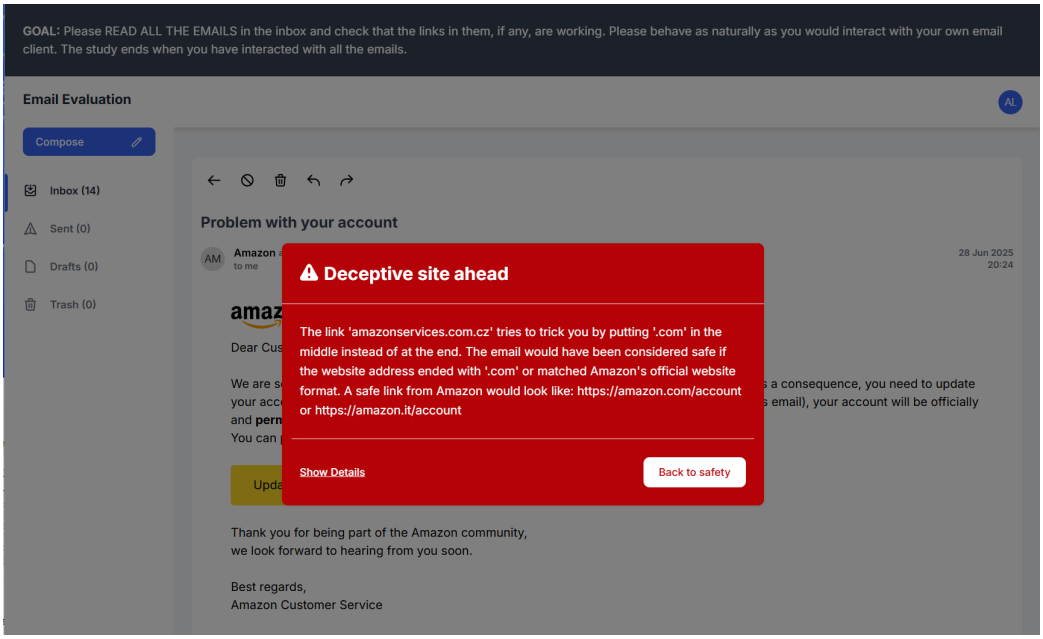

Figure 6: Warning shown during the study for a fake Amazon email that includes a mimicking URL with a suspicious domain (“.com.cz”).

item 8 of the warning questionnaire (relative to the intended action to take).

Moreover, to answer RQ3, a deductive thematic analysis (TA) was also conducted to examine qualitative responses from open-ended questions of the warning questionnaire (see items 2, 6, and 9 in Table 1) following Braun and Clarke’s six-step framework (Braun and Clarke, 2006).

Finally, for RQ4 (“How user-related factors affect CTR?”), we employed a binomial logistic regression to analyse the effect of user-related factors, such as demographics, need for cognition, and warning perception, on CTR.

In each analysis of the RQs, three splits of the dataset were considered: firstly, the whole dataset to analyse the general CTR; secondly, only the data relative to the True Positive emails to examine the warning’s level of protection against actual phishing emails; and thirdly, only the data relative to the False Positive email, to gain insights about the interaction with a warning dialogue that is wrong. Lower CTRs in the case of true positive emails indicate a higher level of protection, as the user is effectively avoiding actual phishing links. On the other hand, a lower CTR in the case of the false positive indicates that the warning does not support the user in recognising

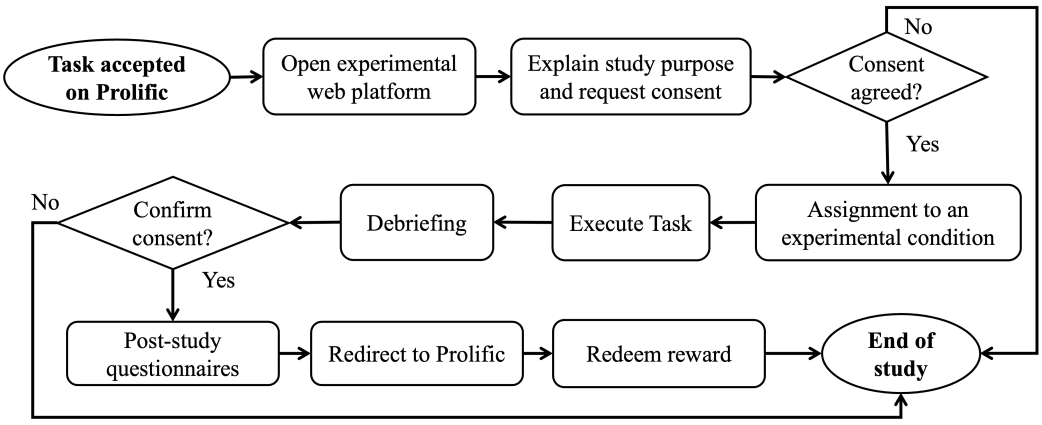

Figure 7: Study Procedure - adapted from Greco et al. (2025).

when the system is incorrect; thus, a higher CTR is preferable for the false positive condition only. All the quantitative data were analysed using R 4.2.2. An alpha level of .05 was used for all statistical tests.

5. Results and Discussion

In this section, we report the results of the controlled experiment to answer each research question. The results of the statistical tests are reported only when significant differences emerged; the full results of the statistical analysis are reported as additional resources in our GitHub repository.

5.1. RQ1 - LLM vs Manual Explanations

To answer the first research question, we compared the results of experimental conditions with the baseline from Greco et al. (2025), which employed warnings with manually generated explanations. To obtain a fair comparison, we excluded from this analysis the conditions that employed counterfactual explanations, as the baseline included a feature-based explanation (n = 300 for the aggregated experimental conditions, n = 150 for the baseline sample, n = 450 in total). The CTRs for the baseline and LLM feature-based conditions (both individual and aggregated) are reported in Figure 8. As it can be observed, LLM-generated warnings generally led to a higher CTR when aggregated, compared to the baseline. A more in-depth analysis revealed that the Claude FB condition had the lowest CTR (9.39%), especially for true positive emails (7.48%), while the baseline led to the lowest CTR in the case

of the false positive (9.78%). On the contrary, Llama FB led to the highest CTR in all cases (14.65% in total, 14.62% for true positives, and 14.43% for false positives).

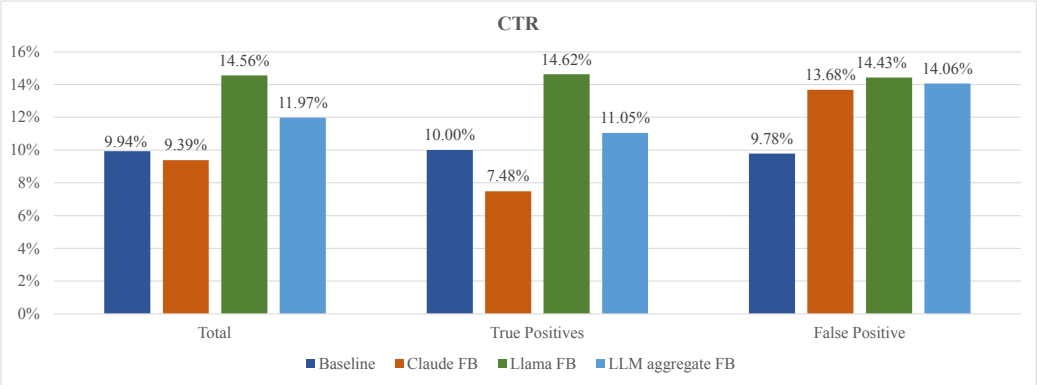

Figure 8: CTRs in the baseline condition (manually-generated, feature-based explanation) and for the LLM-generated feature-based explanations, presented both individually (Claude FB, Llama FB) and aggregated.

The Chi-square test of independence revealed no significant statistical difference between the baseline and the LLM feature-based conditions, both individually and when aggregated ($p > 0.05$). This indicates that LLM-generated warnings are at least as effective as manually-generated ones in protecting users. However, comparing the baseline, Claude FB, and Llama FB led to a significant difference ($\chi^2 = 5.86$, df $= 2$, $p = 0.05$) for the True Positive emails. Although pairwise comparisons did not reveal differences, a trend emerged between Claude-generated feature-based warnings (8.90%) and Llama-generated ones (14.29%), as the post-hoc test indicates a marginal difference between the two ($p = 0.084$).

5.2. RQ2 - Feature-based vs Counterfactual Explanations

The second research question regarded the effectiveness of explanations in warning dialogues based on their type (feature-based or counterfactual). Therefore, we aggregated the conditions according to the explanation type they included and made a direct comparison (n=300 for both samples, n=600 in total). To have a balanced analysis, we excluded the baseline from this comparison. The CTRs for the conditions included in this analysis are reported in Figure 9.

Compared to the feature-based explanations, the warnings with counterfactual explanations led to a lower CTR in general (11.41% vs 11.97%) and for the false positive email (10.00% vs 14.06%), while slightly increased the CTR for the true positives (12.04% vs 11.03%).

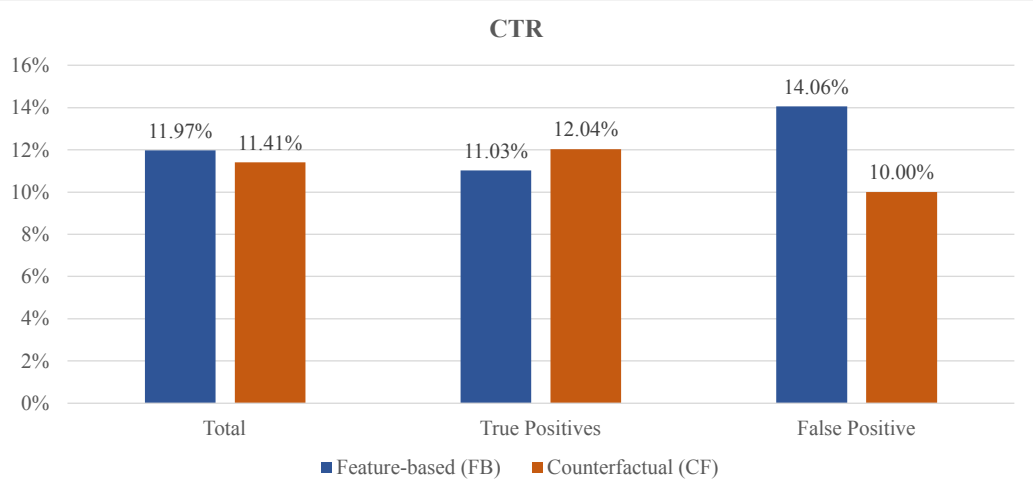

Figure 9: CTRs for the aggregated conditions with feature-based and counterfactual explanations. The baseline is not included.

The Chi-Square test did not highlight any statistically significant difference between the two types of explanation, using the same LLM. This is true when either considering all the emails ($\chi^2 = 0.05$, df $= 1$, $p = 0.827$), only the true positives ($\chi^2 = 0.12$, df $= 1$, $p = 0.724$), or only the false positives ($\chi^2 = 1.13$, df $= 1$, $p = 0.288$).

To deepen the analysis of RQ2, we investigated the CTR for the experimental conditions taken individually (n=150 for each sample, n=600 in total). Figure 10 reports the results. It is noticeable how feature-based explanations from Claude led to the highest level of protection (CTR = 9.39% in general and CTR = 7.48% for true positives). In contrast, feature-based explanations from Llama led to the highest CTR (14.56% in general, 14.62% for true positives, and 14.43% for the false positives). In the case of the false positive email, the baseline led to the lowest CTR (9.78%).

The Chi-Square test highlighted no significant differences. However, the true positive emails led to a slight statistical difference ($\chi^2 = 6.97$, df $= 3$, $p = 0.073$). This result might suggest a difference in CTR between the Claude FB (7.48%) and Llama FB (14.62%) conditions.

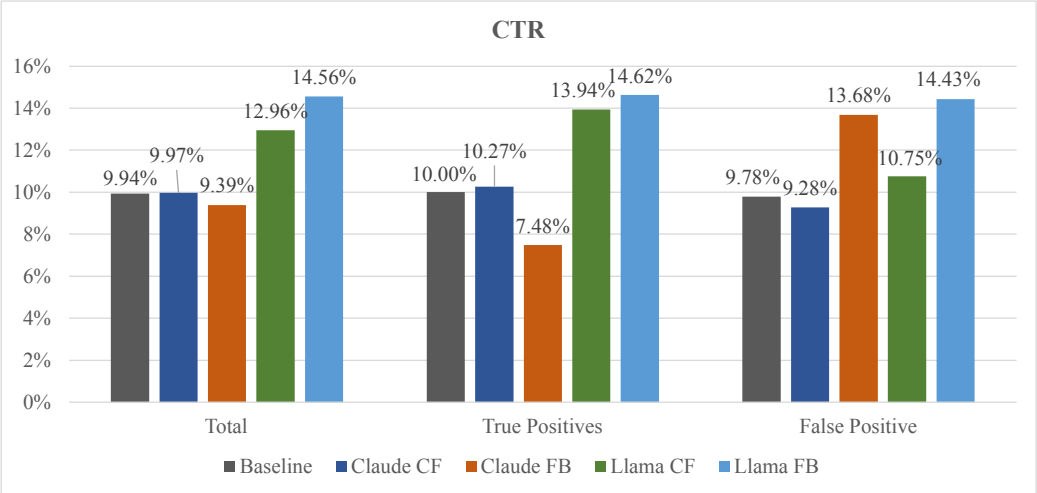

Figure 10: CTRs for the individual experimental conditions and the baseline.

5.3. RQ3 - User Perception of Warnings

The third research question explored how users perceive warnings with LLM-generated explanations. To do this, we examined the results of the warning questionnaire (see Table 1) and the raw NASA-TLX. Responses to the NASA-TLX were considered to measure the workload caused by the experimental condition. Moreover, to get more insights, we also computed the six sub-dimensions of the NASA-TLX, i.e., mental demand, physical demand, temporal demand, perceived performance, perceived effort, and frustration level.

The results of the extracted dimensions for the baseline, the four experimental conditions, and the aggregation based on explanation type and LLM are reported in Figures 11 and 12.

From the Kruskal-Wallis test between the four experimental conditions and the baseline (n=150 each, n=750 in total), a statistically significant difference emerged for warnings' understandability ($H = 15.76$, df $= 4$, $p = 0.003$). Dunn's post-hoc test revealed that warnings generated with Llama were perceived as significantly more understandable than the baseline, both in the case of feature-based ($Z = 3.53$, $p = 0.002$) and counterfactual explanations ($Z = 3.29$, $p = 0.005$). No other significant differences emerged.

Item 8 of the warning questionnaire ("What action, if any, did the warning dialogue want you to take?") required a separate analysis, as the data type was not numeric but categorical. The Chi-square test did not reveal any

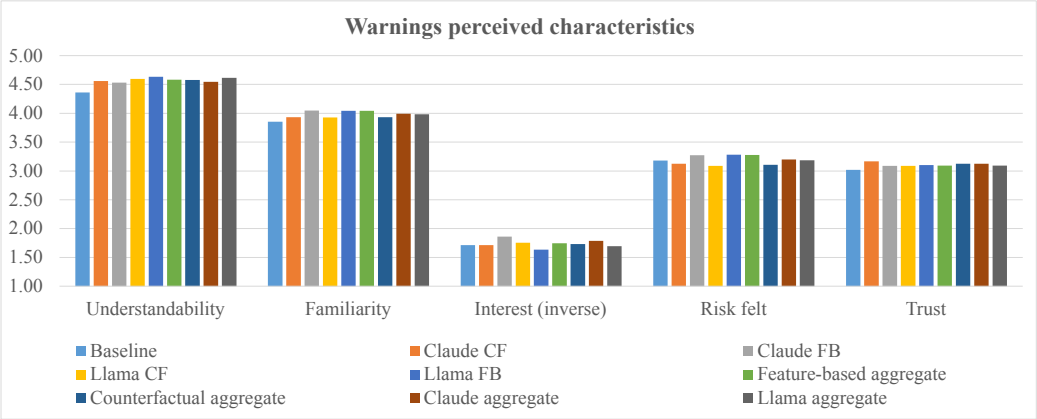

Figure 11: Average values of perceived characteristics emerging from the results of the warning questionnaire.

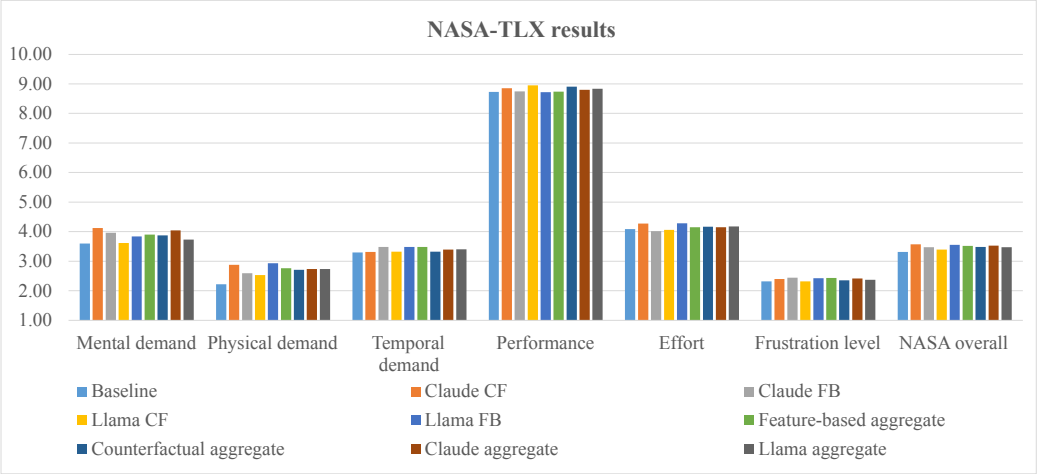

Figure 12: Average values of each dimension of the Raw NASA-TLX and overall scores for each condition and aggregation.

significant difference ($\chi^2 = 9.68$, df $= 5$, $p = 0.644$).

5.4. RQ4 - How User-Related Factors Affect CTR

To investigate how user-related factors affect CTR, we built a series of binomial logistic regression models including demographic measures (age, gender, expertise, average hours online per day, Need-for-Cognition) and warning-perception variables (understandability, familiarity, trust, perceived

Table 2: Answers to Item 8 of the warning questionnaire by warning condition. Percentages represent the proportion of participants selecting each response option.

Condition	**"Don't continue"**	**"Be careful"**	**"Continue"**	**"None"**
Baseline	74.0%	24.0%	0.0%	2.0%
Claude CF	71.3%	25.3%	2.7%	0.7%
Claude FB	78.0%	18.7%	2.7%	0.7%
Llama CF	74.0%	21.3%	3.3%	1.3%
Llama FB	74.0%	24.0%	1.3%	0.7%
FB aggregate	76.0%	21.3%	2.0%	0.7%
CF aggregate	72.7%	23.3%	3.0%	1.0%
Claude aggregate	74.7%	22.0%	2.7%	0.7%
Llama aggregate	74.0%	22.7%	2.3%	1.0%

suggested action, mental workload). Different models were trained for three outcomes: CTR across all emails (ALL), true positives (TP), and false positives (FP), across all conditions and LLM variants.

Moreover, odds ratios (OR) were computed by exponentiating the estimated coefficients from the logistic regression models, by applying the formula $\text{OR} = e^{\beta}$. This transformation allows interpreting each β coefficient as the multiplicative change in the odds of clicking associated with a one-unit increase in the predictor. An OR greater than 1 indicates a higher likelihood of clicking, whereas an OR below 1 indicates a lower probability. To further simplify the interpretation of the OR, we mapped the ORs into *effect sizes*, categorised as small (OR 1.2–1.5 / 0.67–0.83), moderate (1.51–2 / 0.5–0.66), and large (>2 / <0.5), following the guidelines suggested by Chen et al. Chen et al. (2010). The majority of significant effects observed in our data fell either into the "small" (OR 1.2–1.3) or "large" (OR>2 or <0.5) categories, with relatively few predictors yielding odds ratios in the moderate range. Odds ratios of 0.84 to 1.19 were considered as negligible because they show a difference in odds of less than 20% in clicking (which will practically be insignificant even in the case of a significant difference). This strategy is guided by the prescriptions of the applied behavioural research to differentiate between the statistical and practical significance of the requirements (Kirk, 1996; Fritz et al., 2012).

Familiarity with warnings showed consistent large effects: higher famil-

iarity was associated with markedly higher CTR in multiple conditions. ORs ranged from 1.8 to 3.24, with significant effects in Counterfactual aggregate (ALL, TP, FP), Claude Counterfactual (ALL, TP, FP), Claude aggregate TP, and Llama Counterfactual (ALL, TP).

Interpreting the warning as suggesting "don't continue" led to large decreases in CTR across different conditions: ORs between 0.10 and 0.41, significant in Counterfactual aggregate (ALL, TP), NoLLM (ALL, TP), Llama aggregate (ALL, TP), and Claude aggregate (ALL, TP).

Gender showed large and asymmetric effects. Female participants consistently had much lower CTR: ORs as low as 0.0003 (NoLLM FP) and generally below 0.05 in several conditions (Counterfactual, NoLLM, Llama aggregate, Claude aggregate, Claude Counterfactual, Llama Counterfactual). Male participants had higher CTR in some conditions: OR=2.62 (Counterfactual aggregate TP), OR=2.13 (Llama aggregate TP), while in one case, male gender was associated with lower CTR (NoLLM FP, OR=0.13).

NASA-TLX (mental workload) was associated with moderate to significant reductions in CTR: ORs between 0.45 and 0.74, significant in Feature-based FP, Llama aggregate (ALL, TP), Counterfactual aggregate (ALL, TP), Claude aggregate (ALL, FP), Claude Feature-based FP, and Llama Feature-based FP.

Understandability also showed protective effects: higher scores correlated with lower CTR, with ORs between 0.48 and 0.51 across Counterfactual aggregate (ALL, TP) and Claude aggregate FP.

Trust in warning (Llama Feature-based FP) also showed a significant reduction in CTR (OR=0.44).

Average hours online per day had small positive effects: OR=1.24 (NoLLM ALL) and OR=1.22 (NoLLM TP).

Interactions between being **male** and **need-for-cognition** produced small to large effects: OR=1.44 in Counterfactual aggregate for false positives, OR=2.15 in Claude Counterfactual for false positives, OR=1.33 in Claude Counterfactual for all emails, OR=1.32 in Claude Counterfactual for true positives, a smaller effect of OR=1.21 in Counterfactual aggregate for all emails.

Other smaller effects were observed for the interaction between **expertise trust in warning** (OR=0.44).

Full results, including ORs, p-values, and effect size categorisations, are detailed in Table A.4 reported in Appendix A.6.

5.5. Qualitative results

This subsection describes the themes identified through deductive thematic analysis of Items 2, 6, and 9, as presented in Table 1. This analysis strictly followed Braun and Clarke's six-step framework (Braun and Clarke, 2006): *data familiarisation*, *code generation*, *theme identification*, *theme review*, *theme definition and naming*, and *report production*. The deductive approach was selected to align the coding framework with the prior study that inspired our experimental design and provided the baseline warning condition (Greco et al., 2025). Two independent researchers with a background in Human-Computer Interaction and qualitative research methods carried out the process. They were blind to the participants' experimental conditions during the first phases to avoid bias. In particular, the researchers first familiarized themselves with the data by repeatedly reading transcripts and noting initial impressions. Then, they independently coded relevant excerpts. Inter-rater reliability was evaluated using Cohen's kappa ($\varkappa$) values: Item 2 = .81, Item 6 = .84, and Item 9 = .79. These values indicate good agreement (Landis and Koch, 1977). Discrepancies were resolved through discussion until a consensus was reached. Related codes were grouped into preliminary themes (80% initial agreement), which were collaboratively refined to ensure an accurate representation of the data. Each theme was then clearly defined and named, and the final report included interpretive commentary and supporting data excerpts. For each of the three items, the themes captured the essence and meaning of the users' answers, namely, the users' initial reactions to the warnings (item 2), terms perceived as confusing (item 6), and interpretations of warning meanings (item 9).

In particular, Figures 13 and 14 show the frequencies for Item 2 and Item 9 across our experimental conditions, while Table 3 reports the themes, counts, and participants' example responses emerged from the thematic analysis considering confusing words for Item 2.

5.5.1. First reaction to warning (Item 2)

In Item 2, participants answered the question: "When you saw the warning, what was your first reaction?". Five recurring themes were identified across the four experimental conditions: *Perceived threat*, *Take safe action*, *Further checking*, *Warning not trusted*, and *Positive reaction to the warning*. These themes are reported in Figure 13, which shows their frequency across the four experimental conditions defined by the LLM (LLaMA or

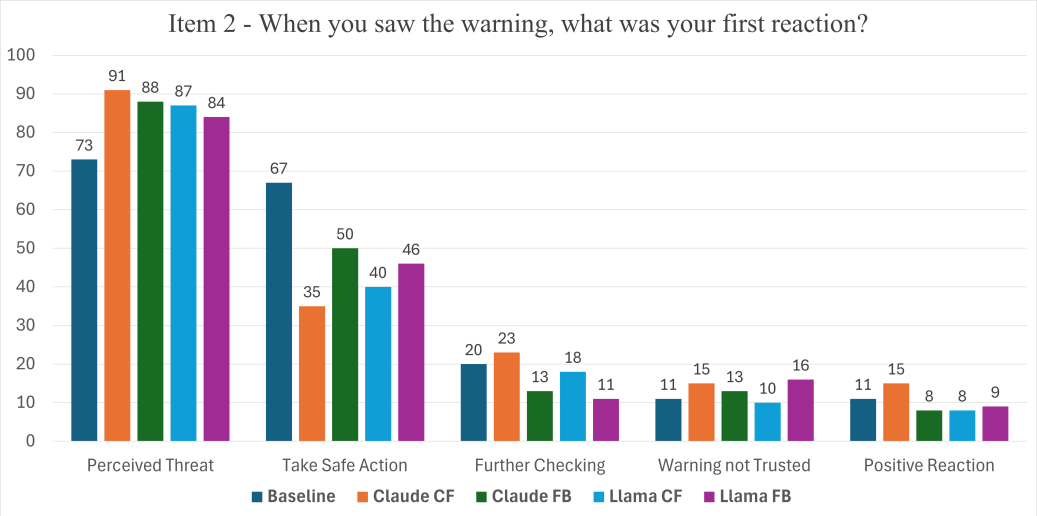

Figure 13: Distribution of first reaction themes (Item 2) across experimental conditions, including the manually crafted feature-based baseline from Greco et al. (2025), and the combination of LLM (LLaMA, Claude), explanation style (feature-based, counterfactual). Each theme's total number of instances is indicated on top of each bar.

Claude), the explanation style (feature-based or counterfactual), and the manual feature-based explanation from Greco et al. (2025).

Perceived threat. This theme captured participants' immediate perception of danger, often accompanied by feelings of uncertainty or concern. For instance, one participant (Claude, feature-based) remarked: *"It was strange, but I thought it had something to do with some particular links not being recognized by the computer"* (P290). Another (LLaMA, counterfactual) noted: *"That the site I am potentially about to visit is malicious and may cause harm to my computer."* (P393). Claude FB tends to elicit this perception more than the other conditions (91 times), whereas the baseline elicits this perception to a lesser extent (73 times).

Take safe action. Many users responded by choosing the safest path, typically avoiding interaction with the email or the suspicious link. A participant (Claude, counterfactual) stated: *"To click on 'Back to safety"'* (P9), while another (LLaMA, feature-based) wrote: *"That the site I am about to visit is at a high risk of being a scam site and I am advised to be careful"* (P512). Here, it is evident that in the baseline condition, this action is elicited very often (67 times). In contrast, for LLM-generated conditions, this idea is much less considered by users (35 to 46 times).

Further checking. Some participants described their intention to verify the warning by inspecting the email or sender. For example, a participant (Claude, feature-based) commented: *"Stop and think about what to do"* (P518), and another (LLaMA, counterfactual) shared: *"I wanted to verify who the sender was and if the link made sense"* (P120). The feature-based conditions elicit this action less frequently (13 times Claude FB, 11 times Lama FB) than the other conditions (from 18 to 23 times).

Warning not trusted. A subset of responses expressed scepticism or confusion about the alert. One user (Claude, counterfactual) admitted: *"I was a bit worried if this is not a real warning, but of course it is not."*, while another (LLaMA, feature-based) stated: *"I didn't trust the warning, it seemed like a false alarm"* (P263). In this case, the results appear more balanced across all the experimental conditions.

Positive reaction to the warning. Some users welcomed the alert or found it informative. A participant (LLaMA, counterfactual) reflected: *"It was helpful to get an alert like that, made me feel safer"* (P370). In this theme, the results are generally more balanced across all experimental conditions, except for the Claude FB, which elicited a positive reaction from participants (15 times).

5.5.2. Confusing words (Item 6)

With this item, we asked participants to indicate which words they found confusing or too technical in the warning dialogue with an open question. Overall, four themes emerged[4], summarized in Table 3. The most common theme was "Warning is clear", highlighting that participants did not find any word confusing or too technical. The second theme, "Unclear general word", includes responses where participants reported confusion about general terms used in the warning text (e.g., "deceptive", "harmful", or "suspicious"). The third theme, "Unclear technical term", refers to participants who reported difficulties with technical terms included in the warning (e.g., "IP address", or "https"). The last theme is "Unclear UI element", referring to participants who found the words and buttons of the warning dialogue misleading (e.g., the function of the button "click here (not safe) to read" was sometimes misinterpreted by participants).

[4]Please note that 179 participants did not respond, as answering was optional.

Table 3: Themes, counts, and example answers for confusing words item: " *Which word/s did you find confusing or too technical?*".

Themes	**Theme count**	**Answer examples**
Warning is clear	282	"I didn't find any part too confusing or technical." (P230)
Unclear general word	109	"The numbers provided in the texts, such as the email points" (P52), "The word *Deceptive*" (P41)
Unclear technical term	5	"The word *https*" (P230)
Unclear UI element	13	"To still have to click the link clearly specified *not to be safe* at the end of the warning" (P11)

5.5.3. Warning meaning (Item 9)

In this item, participants responded to the question "What do you think this warning means?". Thematic analysis revealed eight distinct themes, presented in Fig. 14, along with their frequencies across the four conditions and the manually crafted feature-based explanation (baseline).

Clear threat. Participants indicated the presence of a generic threat in the suspicious email or website, prompting them to avoid clicking on links or accessing the content (e.g., "It means that the website you're trying to enter is a scam and not real. ", P77). Here, the feature-based conditions elicit this perception more frequently (54 times Claude FB and 46 times Lama FB) than the other conditions (from 35 to 38 times).

Possible threat. Participants interpreted the warning as indicating that the content could be dangerous or lead to undesirable consequences, though not with certainty, prompting a need for further investigation (e.g., "It warns me against a possible phishing scam ", P80). It is noticeable that the baseline, with a manually crafted feature-based explanation, instils less perceived threat (45 times) than LLM-generated explanations (from 76 to 99 times).

Malicious website. Participants' interpretations of the warning varied, particularly concerning the specific source of danger. The threat was often associated with the website linked in the email. Additionally, some participants perceived the warning as indicating a particular threat (e.g., " That the link takes you to a deceptive website and not to the one that was supposed to be.", P85) while others saw it as more general (e.g., " It's letting you know that it is fake and that you can potentially be hacked.", P532). Participants

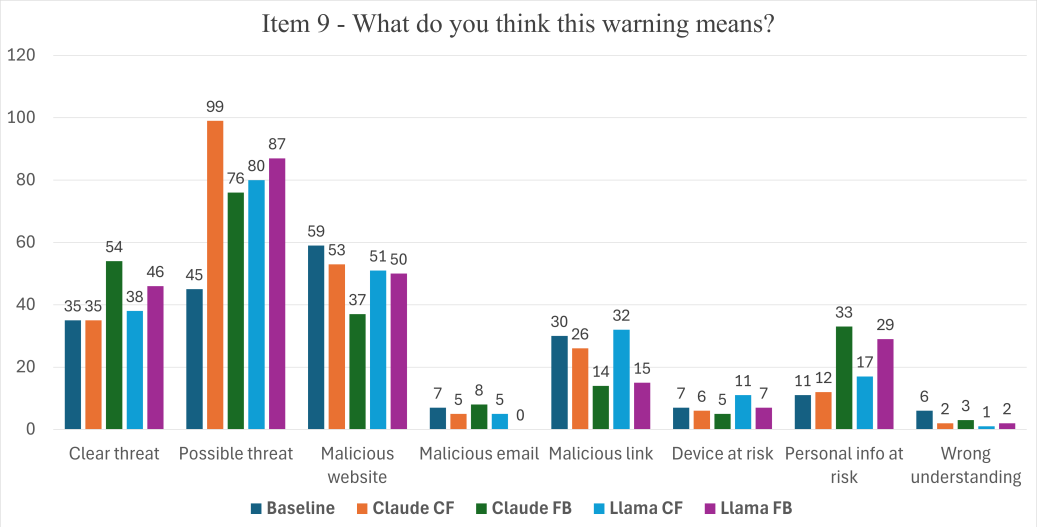

Figure 14: Distribution of warning meaning themes (Item 9) across experimental conditions, including the manually crafted feature-based baseline from Greco et al. (2025), and the combination of LLM (LLaMA, Claude), explanation style (feature-based, counterfactual). Each theme's total number of instances is at the top of each bar.

had a similar impression across all conditions (from 50 to 59 times), except for Claude FB, which resulted in a lesser impression of a malicious website (37 times).

Malicious email. In some cases, participants interpreted the warning as indicating that the threat resided in the content of the email itself or a scammer sender. Interpretations within this theme varied in terms of perceived threat certainty, ranging from a definitive assessment of maliciousness (e.g., "It means that the email has been identified as being sent from a scammer.", P244) to more tentative assessments of potential risk (e.g., "Warns users to be careful of suspicious links/emails that can take their data", P214). This theme was mentioned infrequently across conditions.

Malicious link. In some instances, participants understood the warning as indicating that the threat was embedded in the link. The perceived level of risk varied, ranging from a clear and immediate danger (e.g., "That the link I clicked leads to a site that doesn't belong to Facebook and they will steal the information I type.", P361) to a more cautious interpretation of potential harm (e.g., "Warns users to be careful of suspicious links/emails that can take their data", P215). This theme was more frequent for counterfactual

(26 times in the case of Claude CF and 32 times for Llama CF) and baseline explanations than for feature-based (14 times in the case of Claude FB and 15 times for Llama FB).

Device at risk. The warning was interpreted as an indication that the user's device could be at risk, for example, due to potential malware contained within the email content or linked website (e.g., "It means the webpage or link you are about to enter may harm your computer of breach data", P513). This theme was generally less frequent than others, and quite balanced across all conditions.

Personal information at risk. Some participants interpreted the warning as suggesting that their data, such as login credentials, were at risk of being stolen by cybercriminals (e.g., "This warning means the website you're attempting to visit is likely a fake site that's designed to steal login details.", P157). This theme was more frequent for feature-based explanations (32 times in the case of Claude FB and 29 times for Llama FB) than for the baseline and counterfactual ones (12 times in the case of Claude CF and 17 times for Llama CF).

Wrong understanding. The final theme encompasses all instances where participants misinterpreted the warning (e.g., "It means the website has a bug", P131; "That I can not continue with whatever I was doing then", P326). Overall, this was the most infrequent theme, where the manually crafted explanations (baseline) achieved the highest counts.

6. Threats to validity

In this section, we analyse some potential issues that might threaten the study's validity.

6.1. Internal Validity

Internal validity can be threatened by hidden factors that can compromise the achieved conclusions.

Confounding Variables. Latent factors such as prior knowledge and experience with phishing may have influenced behaviour and response to warnings. We used a between-subjects design to mitigate this threat, and participants were randomly assigned to conditions.

Deception and priming. At the beginning of the experiment, participants were informed that the study aimed to evaluate a new email client interface, thus hiding the true objective of the study, which was revealed at

the end to avoid biasing participants towards unnatural behaviour. Even if this procedure was ethically approved to reduce demand characteristics, such framing may have influenced attentiveness and perceived task seriousness.

6.2. External Validity

External validity regards the extent to which the results can be generalized to broader contexts, populations, and real-world settings.

Ecological validity. The email client we created mimics real and common features of web clients; however, users knew that they were involved in a study, thus in a safe environment. This reduced-risk setting can lead to unnatural behaviours that users exhibit when facing phishing attacks. This risk was mitigated by asking users to behave as naturally as possible.

Participant recruiting. To make the study population as representative as possible of a generic population to a phishing attack, we used Prolific by setting only gender and English knowledge as inclusion criteria. This guaranteed that participants were recruited worldwide and of different ages. However, considering the typical Prolific users, our sample may not include older adults, cybersecurity professionals, or individuals with low digital literacy.

Device and context effects. Participants attended the study remotely on their devices in uncontrolled environments. Factors such as device, screen size, multitasking, or ambient noise may have affected attention and decision-making.

6.3. Construct Validity

Construct validity examines whether the study accurately measures its intended concepts, namely, the effectiveness and perception of LLM-generated phishing explanations.

Behavioral measures. We adopted the click-through rate (CTR), the most widely adopted metric to measure the phishing vulnerability. However, clicks may occur in a controlled and safe setting for various reasons, including curiosity or misunderstanding. Thus, this behaviour may not always reflect a failure to detect phishing intent.

6.4. Statistical Validity

Statistical validity refers to the appropriateness of the data analysis methods and the robustness of the inferences drawn.

Multiple comparisons. In our study, different statistical tests such as Chi-square analyses, Kruskal-Wallis tests, and logistic regressions were conducted. However, while Bonferroni correction was applied where applicable, the risk of Type I errors was reduced, but it marginally remained.

7. Lessons learned

Triangulating qualitative and quantitative findings, we distilled lessons learned and directly provided guidance to design and implement future decisions. The four research questions are correlated with these lessons, thus determining the practical implications of our studies. They explain how the interaction between various design options, methods of explanation, and user differences affects the effectiveness and the perceived utility of AI-produced phishing warning messages.

Lesson 1 (RQ1): Automating warnings generation with LLMs without losing effectiveness

Feature-based explanations generated by LLM proved to be in line, or sometimes better, than the phishing protection of manually written explanations. To be more precise, the CTR of the Claude-generated explanations was even lower than that of the baseline and the Llama, thus indicating that a quality LLM-generated explanation is as effective as one written by experts in defending users.

These results align with past studies examining the automatic performance of explanations when used in security settings. For example, Chen et al. (2021) found users deemed automated phishing explanations beneficial and credible, but not all were protective. Similarly, Liao et al. (2020) showed that the overall meaning of an explanation, as opposed to its stylistic structure, has more influence on user trust and behaviour. All these results together demonstrate that properly fine-tuned LLM outputs exceed (or are on par with) human efforts in that area when considering causal reasoning.

Takeaway. If applying high-quality LLMs along with focused prompt engineering, explanation messages on phishing warnings can achieve the performance of an expert-created message, thus making it possible to create explanation systems at scale.

Lesson 2 (RQ2): Feature-based explanations may enhance the accuracy in identifying actual phishing emails.

In investigating feature-based explanation styles, the present study reveals that in the case of true positive emails, feature-based resulted in a lower CTR (11.03%) with respect to counterfactual ones (12.04%). This difference was more pronounced in the case of Claude-generated explanations (Claude FB: 7.48% vs. Claude CF: 10.27%), although this should be interpreted as an observed trend rather than a statistically significant effect, given the p-value exceeding conventional thresholds. These insights are consistent with the recent study by Zhao et al. (2024), showing that slight differences in language model tuning may trigger strong behavioural outcomes. The analysis of Item 9 further illustrates that explanations issued by Claude more frequently prompted interpretations invoking data theft or credential loss, thus suggesting a heightened salience of risk.

Takeaway. Feature-based explanations may help reduce clicks on phishing attempts, although further research is needed to deepen the investigation of this trend; in particular, LLMs such as Claude appear to convey a greater sense of threat.

7.1. Lesson 3 (RQ2): Counterfactuals may reduce false alarm overrides

Even if counterfactuals showed no overall CTR advantage, it emerged that they might reduce false positive overrides (10.00% vs. 14.06% for feature-based). Also, qualitative responses (e.g., "The email would have been safe if. . . " [P393]) indicate counterfactuals help users distinguish true threats from benign anomalies. Thus, counterfactuals help users reason about *why* the content may or may not be dangerous.

Takeaway. Adopt counterfactual explanations when the risk of false positives is high, such as in sensitive domains (e.g., internal communication or finance), where false alarms can disrupt workflows.

Lesson 4 (RQ2, RQ3): Explanation style shapes the user's mental model of phishing risk

The analysis of Item 9 revealed that feature-based and counterfactual explanations elicit different sources of threats in users. In the case of feature-based explanations, users tend to identify the outcome of the threat, for example, stolen credentials or leaked data. On the contrary, in the case of counterfactual explanations, users often pinpoint the mechanism of deception, especially deceptive links and their mismatch with the related labels.

This finding aligns with findings in literature Djatsa (2020); Greco et al. (2025), which revealed that the explanations of the threat influence the users' beliefs about where the risk resides.

Takeaway. Counterfactual explanations should be presented to highlight manipulation tactics, while feature-based explanations should be visualized to reinforce the severity of potential consequences.

Lesson 5 (RQ3): Static metrics and user perceptions can signal warning efficacy.

Empirical evidence indicates that better readability scores (according to the SMOG index and Flesch-Kincaid grade level scores) are usually associated with low CTRs. Two contrasting sets of explanatory messages illustrate this phenomenon: Claude's feature-based phrases, which achieved a SMOG score of 7.9 and a Flesch-Kincaid grade level of 5.6, registered a CTR of 9.39%, whereas Llama's analogous constructs, with a SMOG score of 8.9 and a Flesch-Kincaid grade level of 6.9, demonstrated a higher CTR of 14.65%. The above observations provide tentative support for the hypothesis that adopting less complicated wording might enhance the effectiveness of the warning message.

Moreover, the questionnaire results showed that perceived trust and perceived risk were important predictors of CTR. In the false-positive subset, higher values of perceived trust in the warning significantly decreased the odds of clicking (OR=0.44, p=0.038), suggesting an overconfidence effect in the warning, which may lead to discarding genuine emails.

The strongest individual predictor across all logistic regression models was how users interpreted the action prescribed in the warning: those who reported that the message instructed them to "don't continue" were 78% less likely to click (OR = 0.22, $p < .001$). This effect remained robust across all subsets, including true and false positives, and was consistent across models using both Claude and Llama LLMs.

Takeaway. A combination of readability measures and user-based assessment, such as perceived trust, perceived risk, and message clarity, represents a promising approach to enhancing the efficacy of warnings and minimizing the vulnerability to phishing.

Lesson 6 (RQ4): Habituation and high online exposure both increase risk

The study results suggest that familiarity with warnings and the extended use of the internet daily are both highly significant predictors of clicking on

phishing emails. In particular, the odds of clicking were more than doubled by a one-point increase in familiarity (OR = 2.20, p = .001), which indicates that habituation is a direct detriment to the effectiveness of warnings. Similarly, each extra hour spent online per day increased click probability by about 25% (OR = 1.25, p = .024), implying that individuals exposed to a lot of online content have a high chance of adopting automatic or shortcut browsing behaviours. These findings are consistent with prior studies on habituation in warning design literature (Amran et al., 2018, Akhawe and Felt, 2013, Egelman et al., 2008), which proved that familiarity may reduce people's sensitivity to a familiar warning signal.

Takeaway. Warning designs should disrupt habitual browsing by introducing visual or semantic variability, especially for users with high exposure to online content. For example, designers can rotate, refresh, or personalise warning designs to sustain user attention and reduce the risks posed by the habituation effect.

Lesson 7 (RQ4): Mind the (cognitive) gap – workload slows risky clicks

The data from the NASA-TLX questionnaire revealed a significant negative correlation between the perceived workload and total click-through rate ($\beta = -0.029$, $p = 0.002$). In particular, the increased mental effort was associated with a decline in the probability of impulsive clicks. This observation was supported via logistic regression models, which showed that the odds of clicking reduced by 34% (OR=0.66, $p = 0.012$) with each increment of workload. Further, the impact was similar across the explanations and data subsets styles, such as true-positive (OR=0.65, $p = 0.012$) and false-positive conditions (e.g., Claude aggregate FP; OR=0.53, $p = 0.004$).

The findings are congruent with past results by Sheahan et al. (2024) who state that constructive friction promotes reflection before doing something. Our finding is also in line with Kahneman (2011), which illustrates how shifting from fast (impulsive) to slow (deliberative) thinking improves judgment in risk-laden contexts.

Takeaway. Introducing mild cognitive friction, such as a prompt or delay, might help prevent impulsive decisions on phishing links.

Lesson 8 (RQ4): Gendered clicks - demographics matter

Logistic regressions revealed that gender had significant and systematic effects on click behaviour. The likelihood of female participants clicking on phishing emails was significantly lower, and the odds ratios averaged almost

zero in true-positive and false-positive phishing email conditions. On the contrary, male participants with higher values of Need for Cognition (NFC) were over twice as likely to click in the false-positive conditions with counterfactual warnings (OR = 2.16, p = .012). The gender-by-NFC interaction indicates that some cognitive characteristics, in combination with gender, can increase vulnerability, perhaps due to over-trusting of the reasoning ability when subjected to ambiguous cues by the user. These results support the findings obtained by Greitzer et al. (2021), who also observed that male participants were more susceptible to phishing attacks in simulated attacks; they explained the trend with increased overconfidence or lowered caution.

Takeaway. Security messaging may benefit from adaptive targeting that accounts for demographic factors like gender.

8. Conclusions

This study offers empirical evidence that LLMs can automatically produce phishing warning explanations that are as effective as, and in certain circumstances, more effective than, those crafted by humans. Notably, warnings generated by Claude 3.5 Sonnet markedly reduced click-through rates on phishing links compared with Llama 3.3 70B and human-composed messages, especially when a feature-based explanation style was used. Counterfactual explanations showed initial potential in reducing false positives, yet feature-based explanations remained more reliable for identifying genuine threats.

The viability of LLM-generated warnings was supported with user-perception measures: participants found them clear, trustworthy, and actionable, especially those generated by Claude. The study also found that mental workload, gender, and familiarity with warning dialogues significantly affected the warning efficacy. These results stress the role of individualisation and flexibility of security interfaces.

Overall, employing LLMs in phishing-defence systems can improve the scalability, responsiveness, and user interaction without reducing effectiveness. Future work should examine real-time deployment, user-personalisation strategies, and broader applications of LLM-generated explanations in cybersecurity interfaces.

Declaration of generative AI

During the writing of this paper, the author(s) used *Grammarly Pro* to fix grammatical errors and improve text quality. After using this tool/service,

the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Acknowledgments

This work has been supported by the Italian Ministry of University and Research (MUR) and by the European Union-NextGenerationEU, under grant PRIN 2022 PNRR "DAMOCLES: Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES" (Grant P2022FXP5B) CUP: H53D23008140001.

This work is partially supported by the co-funding of the European Union - Next Generation EU: NRRP Initiative, Mission 4, Component 2, Investment 1.3 - Partnerships extended to universities, research centres, companies and research D.D. MUR n. 341 del 5.03.2022 – Next Generation EU (PE0000014 – "Security and Rights In the CyberSpace – SERICS" - CUP: H93C22000620001).

The research of Francesco Greco is funded by a PhD fellowship within the framework of the Italian "D.M. n. 352, April 9, 2022"- under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 3.3 - PhD Project "Investigating XAI techniques to help user defend from phishing attacks", co-supported by "Auriga S.p.A." (CUP H91I22000410007).

CRediT authorship contribution statement

F. M. Cau: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization.

G. Desolda: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

F. Greco: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization.

L. D. Spano: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition.

L. Viganò: Conceptualization. Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Supervision.

References

Adams, L.C., Truhn, D., Busch, F., Dorfner, F., Nawabi, J., Makowski, M.R., Bressem, K.K., 2024. Llama 3 challenges proprietary state-of-the-art large language models in radiology board-style examination questions. Radiology 312, e241191.

Akhawe, D., Felt, A.P., 2013. Alice in warningland: a large-scale field study of browser security warning effectiveness, in: Proceedings of the 22nd USENIX Conference on Security, USENIX Association, USA. p. 257–272. URL: https://dl.acm.org/doi/10.5555/2534766.2534789, doi:10.5555/2534766.2534789.

AL, S., Sagiroglu, S., 2025. Explainable artificial intelligence models in intrusion detection systems. Engineering Applications of Artificial Intelligence 144, 110145. URL: https://www.sciencedirect.com/science/article/pii/S0952197625001459, doi:https://doi.org/10.1016/j.engappai.2025.110145.

Amran, A., Zaaba, Z.F., Mahinderjit Singh, M.K., 2018. Habituation effects in computer security warning. Information security journal: A global perspective 27, 192–204.

Anderson, B.B., Kirwan, C.B., Jenkins, J.L., Eargle, D., Howard, S., Vance, A., 2015. How polymorphic warnings reduce habituation in the brain: Insights from an fmri study. URL: https://doi.org/10.1145/2702123.2702322, doi:10.1145/2702123.2702322.

Basit, A., Zafar, M., Liu, X., Javed, A.R., Jalil, Z., Kifayat, K., 2021. A comprehensive survey of ai-enabled phishing attacks detection techniques. Telecommunication Systems 76, 139–154. URL: https://doi.org/10.1007/s11235-020-00733-2, doi:10.1007/s11235-020-00733-2.

Bauer, L., Bravo-Lillo, C., Cranor, L., Fragkaki, E., 2013. Warning Design Guidelines. Technical Report. Carnegie Mellon University. URL: https://www.researchgate.net/publication/258499093_Warning_Design_Guidelines.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3, 77–101. URL: https://www.

tandfonline.com/doi/abs/10.1191/1478088706qp063oa, doi:10.1191/1478088706qp063oa.

Bravo-Lillo, C., Cranor, L.F., Downs, J., Komanduri, S., Sleeper, M., 2011. Improving computer security dialogs, in: Proceedings of the 13th IFIP TC 13 International Conference on Human-Computer Interaction - Volume Part IV, Springer-Verlag, Berlin, Heidelberg. p. 18–35. URL: https://dl.acm.org/doi/10.5555/2042283.2042286, doi:10.5555/2042283.2042286.

Buçinca, Z., Malaya, M.B., Gajos, K.Z., 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. Proceedings of the ACM on Human-computer Interaction 5, 1–21. URL: https://doi.org/10.1145/3449287, doi:10.1145/3449287.

Buono, P., Desolda, G., Greco, F., Piccinno, A., 2023. 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, ACM, New York, NY, USA. pp. 1–6. URL: https://doi.org/10.1145/3544549.3585802, doi:10.1145/3544549.3585802.

Buçinca, Z., Swaroop, S., Paluch, A.E., Murphy, S.A., Gajos, K.Z., 2024. Towards optimizing human-centric objectives in ai-assisted decision-making with offline reinforcement learning. URL: https://arxiv.org/abs/2403.05911, arXiv:2403.05911.

Cacioppo, J., Petty, R., Feinstein, J., Jarvis, B., 1996. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. Psychological Bulletin 119, 197–253. doi:10.1037/0033-2909.119.2.197.

Cacioppo, J.T., Petty, R.E., 1982. The need for cognition. Journal of personality and social psychology 42, 116. URL: https://psycnet.apa.org/doi/10.1037/0022-3514.42.1.116, doi:10.1037/0022-3514.42.1.116.

Cau, F.M., Hauptmann, H., Spano, L.D., Tintarev, N., 2023. Supporting high-uncertainty decisions through ai and logic-style explanations, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, ACM, New York, NY, USA. p. 251–263. URL: https://doi.org/10.1145/3581641.3584080, doi:10.1145/3581641.3584080.

Cau, F.M., Spano, L.D., 2025. The influence of curiosity traits and on-demand explanations in ai-assisted decision-making, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, ACM, New York, NY, USA. p. 1440–1457. URL: https://doi.org/10.1145/3708359.3712165, doi:10.1145/3708359.3712165.

Celar, L., Byrne, R., 2023. How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains. Memory & Cognition 51. doi:10.3758/s13421-023-01407-5.

Charmet, F., Tanuwidjaja, H.C., Ayoubi, S., Gimenez, P.F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., Zhang, Z., 2022. Explainable artificial intelligence for cybersecurity: a literature survey. Annals of Telecommunications 77, 789–812. URL: https://doi.org/10.1007/s12243-022-00926-7, doi:10.1007/s12243-022-00926-7.

Chen, H., Chen, X., Shi, S., Zhang, Y., 2021. Generate natural language explanations for recommendation. arXiv preprint arXiv:2101.03392 .

Chen, H., Cohen, P., Chen, S., 2010. How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. Communications in Statistics—simulation and Computation® 39, 860–864.

Chen, V., Liao, Q.V., Wortman Vaughan, J., Bansal, G., 2023. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. Proc. ACM Hum.-Comput. Interact. 7. URL: https://doi.org/10.1145/3610219, doi:10.1145/3610219.

Coelho, G.L.d.H., Hanel, P.H.P., Wolf, L.J., 2020. The very efficient assessment of need for cognition: Developing a six-item version. Assessment 27, 1870–1885. URL: https://doi.org/10.1177/1073191118793208, doi:10.1177/1073191118793208, arXiv:https://doi.org/10.1177/1073191118793208. pMID: 30095000.

Desolda, G., Aneke, J., Ardito, C., Lanzilotti, R., Costabile, M.F., 2023. Explanations in warning dialogs to help users defend against phishing attacks. International Journal of Human-Computer Studies 176, 103056. URL: https://www.sciencedirect.com/science/article/pii/S1071581923000654, doi:https://doi.org/10.1016/j.ijhcs.2023.103056.

Desolda, G., Di Nocera, F., Ferro, L., Lanzilotti, R., Maggi, P., Marrella, A., 2019. Alerting users about phishing attacks, in: Moallem, A. (Ed.), HCI for Cybersecurity, Privacy and Trust, Springer International Publishing, Cham. pp. 134–148. URL: https://link.springer.com/chapter/10.1007/978-3-030-22351-9_9, doi:10.1007/978-3-030-22351-9_9.

Desolda, G., Ferro, L.S., Marrella, A., Catarci, T., Costabile, M.F., 2021. Human factors in phishing attacks: A systematic literature review. ACM Comput. Surv. 54. URL: https://doi.org/10.1145/3469886, doi:10.1145/3469886.

Desolda, G., Greco, F., Vigano, L., 2025. Apollo: A gpt-based tool to detect phishing emails and generate explanations that warn users. Proceedings of the ACM on Human-Computer Interaction 9, 33. URL: https://doi.org/10.1145/3733049, doi:10.1145/3733049.

Distler, V., Fassl, M., Habib, H., Krombholz, K., Lenzini, G., Lallemand, C., Cranor, L.F., Koenig, V., 2021. A systematic literature review of empirical methods and risk representation in usable privacy and security research. ACM Transactions on Computer-Human Interaction 28, 1–50. URL: https://doi.org/10.1145/3469845, doi:10.1145/3469845.

Djatsa, F., 2020. Threat perceptions, avoidance motivation and security behaviors correlations. Journal of Information Security 11, 19. doi:10.4236/jis.2020.111002.

Egelman, S., Cranor, L.F., Hong, J., 2008. 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, ACM, New York, NY, USA. p. 1065–1074. URL: https://doi.org/10.1145/1357054.1357219, doi:10.1145/1357054.1357219.

El Aassal, A., Baki, S., Das, A., Verma, R.M., 2020. An in-depth benchmarking and evaluation of phishing detection research for security needs. IEEE Access 8, 22170–22192. URL: https://ieeexplore.ieee.org/document/8970564, doi:10.1109/ACCESS.2020.2969780.

Fok, R., Weld, D.S., 2024. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. AI

Mag. 45, 317–332. URL: https://doi.org/10.1002/aaai.12182, doi:10.1002/aaai.12182.

Fritz, C.O., Morris, P.E., Richler, J.J., 2012. Effect size estimates: current use, calculations, and interpretation. Journal of experimental psychology: General 141, 2. doi:10.1037/a0024338.

Gajos, K.Z., Mamykina, L., 2022. Do people engage cognitively with ai? impact of ai assistance on incidental learning, in: Proceedings of the 27th International Conference on Intelligent User Interfaces, ACM, New York, NY, USA. p. 794–806. URL: https://doi.org/10.1145/3490099.3511138, doi:10.1145/3490099.3511138.

Gentile, D., Donmez, B., Jamieson, G.A., 2025. Human performance effects of combining counterfactual explanations with normative and contrastive explanations in supervised machine learning for automated decision assistance. International Journal of Human-Computer Studies 196. URL: https://doi.org/10.1016/j.ijhcs.2024.103434, doi:10.1016/j.ijhcs.2024.103434.

Greco, F., Desolda, G., Buono, P., Piccinno, A., 2025. Enhancing phishing defenses: The impact of timing and explanations in warnings for email clients. Computer Standards & Interfaces 93, 103982. URL: https://www.sciencedirect.com/science/article/pii/S092054892500011X, doi:https://doi.org/10.1016/j.csi.2025.103982.

Greitzer, F.L., Li, W., Laskey, K.B., Lee, J., Purl, J., 2021. Experimental investigation of technical and human factors related to phishing susceptibility. ACM Transactions on Social Computing 4, 1–48. URL: https://doi.org/10.1145/3461672, doi:10.1145/3461672.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., Turini, F., 2019a. Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems 34, 14–23. doi:10.1109/MIS.2019.2957223.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2019b. A Survey of Methods for Explaining Black Box Models. ACM

Computing Surveys 51, 1–42. URL: https://dl.acm.org/doi/10.1145/3236009, doi:10.1145/3236009.

Gunning, D., Aha, D.W., 2019. DARPA's Explainable Artificial Intelligence Program. AI Magazine 40, 44–58. URL: https://onlinelibrary.wiley.com/doi/10.1609/aimag.v40i2.2850, doi:10.1609/aimag.v40i2.2850.

Gupta, S., Kumaraguru, P., 2014. Emerging phishing trends and effectiveness of the anti-phishing landing page, in: 2014 APWG Symposium on Electronic Crime Research (eCrime), pp. 36–47. URL: https://doi.org/10.1109/ECRIME.2014.6963163, doi:10.1109/ECRIME.2014.6963163.

Hart, S.G., 1986. Nasa task load index (nasa-tlx). URL: https://ntrs.nasa.gov/api/citations/20000021488/downloads/20000021488.pdf.

IBM, 2024. Security x-force threat intelligence index. URL: https://www.ibm.com/reports/threat-intelligence.

Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux.

Khonji, M., Iraqi, Y., Jones, A., 2013. Phishing detection: a literature survey. IEEE Communications Surveys & Tutorials 15, 2091–2121. doi:10.1109/SURV.2013.032213.00009.

Kim, S., Wogalter, M.S., 2009. Habituation, dishabituation, and recovery effects in visual warnings. Human Factors and Ergonomics Society Annual Meeting 53, 1612–1616. URL: https://journals.sagepub.com/doi/abs/10.1177/154193120905302015, doi:10.1177/154193120905302015.

Kincaid, J., 1975. Derivation of New Readability Formulas: (automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Research Branch report, Chief of Naval Technical Training, Naval Air Station Memphis. URL: https://books.google.it/books?id=4tjroQEACAAJ.

Kirk, R.E., 1996. Practical significance: A concept whose time has come. Educational and Psychological Measurement 56, 746–759. URL: https://doi.org/10.1177/0013164496056005002, doi:10.1177/0013164496056005002.

Koh, S., Kim, B.H., Jo, S., 2024. Understanding the user perception and experience of interactive algorithmic recourse customization. ACM Transaction Computer-Human Interaction 31. URL: https://doi.org/10.1145/3674503, doi:10.1145/3674503.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F., Hong, J., 2010. Teaching johnny not to fall for phish. ACM Transactions on Internet Technology 10, 1–31. URL: https://doi.org/10.1145/1754393.1754396, doi:10.1145/1754393.1754396.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q.V., Tan, C., 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA. p. 1369–1385. URL: https://doi.org/10.1145/3593013.3594087, doi:10.1145/3593013.3594087.

Landis, J.R., Koch, G.G., 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics , 363–374.

Laughlin, G.H.M., 1969. Smog grading-a new readability formula. Journal of Reading 12, 639–646. URL: http://www.jstor.org/stable/40011226.

Lee, M.H., Chew, C.J., 2023. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. Proc. ACM Hum.-Comput. Interact. 7. URL: https://doi.org/10.1145/3610218, doi:10.1145/3610218.

Liao, Q.V., Gruen, D., Miller, S., 2020. Questioning the ai: Informing design practices for explainable ai user experiences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15. doi:10.1145/3313831.3376590.

López Espejel, J., Yahaya Alassan, M.S., Bouhandi, M., Dahhane, W., Ettifouri, E.H., 2025. Low-cost language models: Survey and performance evaluation on python code generation. Engineering Applications of Artificial Intelligence 140, 109490. URL: https://www.sciencedirect.com/science/article/pii/S0952197624016488, doi:https://doi.org/10.1016/j.engappai.2024.109490.

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., Ma, X., 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. p. 19. URL: https://doi.org/10.1145/3544548.3581058, doi:10.1145/3544548.3581058.

Matsuura, T., Hasegawa, A.A., Akiyama, M., Mori, T., 2021. Careless participants are essential for our phishing study: Understanding the impact of screening methods, in: Proceedings of the 2021 European Symposium on Usable Security, ACM, New York, NY, USA. p. 36–47. URL: https://doi.org/10.1145/3481357.3481515, doi:10.1145/3481357.3481515.

Mehdi Gholampour, P., Verma, R.M., 2023. Adversarial robustness of phishing email detection models, in: Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics, ACM, New York, NY, USA. p. 67–76. URL: https://doi.org/10.1145/3579987.3586567, doi:10.1145/3579987.3586567.

Olmstead, K., Smith, A., 2017. What the public knows about cybersecurity. URL: https://www.pewresearch.org/internet/2017/03/22/what-the-public-knows-about-cybersecurity/.

Panwar, S., Bansal, A., Zareen, F., 2024. Comparative analysis of large language models for question answering from financial documents, in: Sharma, H., Shrivastava, V., Tripathi, A.K., Wang, L. (Eds.), Communication and Intelligent Systems, Springer Nature Singapore, Singapore. pp. 297–308.

Petelka, J., Zou, Y., Schaub, F., 2019. 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, ACM, New York, NY, USA. p. 1–15. URL: https://doi.org/10.1145/3290605.3300748, doi:10.1145/3290605.3300748.

Prakash, P., Kumar, M., Kompella, R.R., Gupta, M., 2010. Phishnet: Predictive blacklisting to detect phishing attacks, in: 2010 Proceedings IEEE INFOCOM, pp. 1–5. doi:10.1109/INFCOM.2010.5462216.

Rjoub, G., Bentahar, J., Abdel Wahab, O., Mizouni, R., Song, A., Cohen, R., Otrok, H., Mourad, A., 2023. A survey on explainable artificial intelligence for cybersecurity. IEEE Transactions on Network and Service Management 20, 5115–5140. doi:10.1109/TNSM.2023.3282740.

Sarker, I.H., Janicke, H., Mohsin, A., Gill, A., Maglaras, L., 2024. Explainable ai for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. ICT Express 10, 935–958. URL: https://www.sciencedirect.com/science/article/pii/S2405959524000572, doi:https://doi.org/10.1016/j.icte.2024.05.007.

Saxena, S., 2018. Precision vs recall. URL: https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488.

Scharowski, N., Perrig, S.A.C., Svab, M., Opwis, K., Brühlmann, F., 2023. Exploring the effects of human-centered ai explanations on trust and reliance. Frontiers in Computer Science 5. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1151150, doi:10.3389/fcomp.2023.1151150.

Shashidhar, S., Chinta, A., Sahai, V., Wang, Z., Ji, H., 2023. Democratizing LLMs: An exploration of cost-performance trade-offs in self-refined open-source models, in: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore. pp. 9070–9084. URL: https://aclanthology.org/2023.findings-emnlp.608/, doi:10.18653/v1/2023.findings-emnlp.608.

Sheahan, J., Chatting, D., Collins, R., Bley, J., Eriksson, A., Taylor, N., Rozendaal, M.C., 2024. Designing with friction: Inverting notions of seamless technology, in: Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction, pp. 1–4.

Sotirakopoulos, A., Hawkey, K., Beznosov, K., 2011. On the challenges in usable security lab studies: Lessons learned from replicating a study on ssl warnings. URL: https://doi.org/10.1145/2078827.2078831, doi:10.1145/2078827.2078831.

Teso, S., Alkan, Ö., Stammer, W., Daly, E., 2023. Leveraging explanations in interactive machine learning: An overview. Frontiers in Artificial Intelligence 6, 1066049. doi:https://doi.org/10.3389/frai.2023.1066049.

Upadhyay, S., Lakkaraju, H., Gajos, K.Z., 2025. Counterfactual explanations may not be the best algorithmic recourse approach, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, ACM, New York, NY, USA. p. 446–462. URL: https://doi.org/10.1145/3708359.3712095, doi:10.1145/3708359.3712095.

VanNostrand, P.M., Hofmann, D.M., Ma, L., Rundensteiner, E.A., 2024. Actionable recourse for automated decisions: Examining the effects of counterfactual explanation type and presentation on lay user understanding, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA. p. 1682–1700. URL: https://doi.org/10.1145/3630106.3658997, doi:10.1145/3630106.3658997.

Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., Shah, C., 2024. Counterfactual explanations and algorithmic recourses for machine learning: A review. ACM Comput. Surv. 56. URL: https://doi.org/10.1145/3677119, doi:10.1145/3677119.

Viganò, L., Magazzeni, D., 2020. Explainable security, in: IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2020, Genoa, Italy, September 7-11, 2020, IEEE. pp. 293–300.

Vilone, G., Longo, L., 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion 76, 89–106. URL: https://doi.org/10.1016/j.inffus.2021.05.009, doi:10.1016/j.inffus.2021.05.009.

Wang, X., Yin, M., 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making, in: 26th International Conference on Intelligent User Interfaces, ACM, New York, NY, USA. p. 318–328. URL: https://doi.org/10.1145/3397481.3450650, doi:10.1145/3397481.3450650.

Wogalter, M.S., 2018. Communication-human information processing (c-hip) model, in: Forensic human factors and ergonomics. CRC Press, pp. 33–49.

Wogalter, M.S., Conzola, V.C., Smith-Jackson, T.L., 2002. Research-based guidelines for warning design and evaluation. Applied Ergonomics 33, 219–230. URL: https://www.sciencedirect.com/science/article/pii/S0003687002000091, doi:https://doi.org/10.1016/S0003-6870(02)00009-1.

Wu, M., Miller, R.C., Garfinkel, S.L., 2006. Do security toolbars actually prevent phishing attacks?, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. p. 601–610. URL: https://doi.org/10.1145/1124772.1124863, doi:10.1145/1124772.1124863.

Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., Wang, H., 2024. Large language models for cyber security: A systematic literature review. arXiv preprint arXiv:2405.04760 .

Zhang, G., Jin, Q., Zhou, Y., Wang, S., Idnay, B., Luo, Y., Park, E., Nestor, J.G., Spotnitz, M.E., Soroush, A., Campion, T.R., Lu, Z., Weng, C., Peng, Y., 2024. Closing the gap between open source and commercial large language models for medical evidence summarization. NPJ Digital Medicine 7, 239. URL: https://doi.org/10.1038/s41746-024-01239-w, doi:10.1038/s41746-024-01239-w.

Zhang, Y., Liao, Q.V., Bellamy, R.K.E., 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA. p. 295–305. URL: https://doi.org/10.1145/3351095.3372852, doi:10.1145/3351095.3372852.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M., 2024. Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology 15. URL: https://doi.org/10.1145/3639372, doi:10.1145/3639372.

Appendix A. Warning Messages

In the following are reported the explanation messages contained in the warnings for each condition. Each condition had three warnings, one per email: *IP address*, *Top-level domain mispositioned*, and *Link mismatch*.

Appendix A.1. Claude feature-based explanations

- *IP address.* The link shows numbers (92.233.24.33) instead of "instagram.com" like real Instagram emails would use. This is a deceptive trick to send you to a fake website. If you enter your account details there, attackers will gain control of your Instagram account

- *Top-level domain mispositioned.* The website amazonservices.com.cz is trying to look like a real Amazon page by using 'amazon' in its address. This is a trick to make you think you're visiting Amazon when you're not. If you enter your login details, scammers could take control of your real Amazon account.

- *Link mismatch.* The link shown as 'protect your account' may take you somewhere different from what you see. Attackers use this trick to make you think you're going to Facebook when you're not. This could lead to your Facebook password being stolen.

Appendix A.2. Llama feature-based explanations

- *IP address.* The link http://92.233.24.33 is a string of numbers instead of a normal website name. This site might be fake and try to trick you. You might be giving away your private information.

- *Top-level domain mispositioned.* The link https://amazonservices.com.cz/account.php has a strange company name extension. This site might be pretending to be something it's not. You might be disclosing private information.

- *Link mismatch.* The link "https://www.facebook.com/hacked/disavow?u=100000125023309&nArdInDS2&lit_IT&ext1548538159" is an imitation of the original Facebook link. This site might be intended to take you to a different place. You might be disclosing private information.

Appendix A.3. Claude counterfactual explanations

- *IP address.* The link in the email points to a string of numbers (92.233.24.33/instagram/login.php) instead of the official Instagram website name. The email would have been safe if the link used Instagram's actual website name, which helps users verify they're going to the real Instagram site. A safe link would have looked like: https://instagram.com/account/reset

- *Top-level domain mispositioned.* The link 'amazonservices.com.cz' tries to trick you by putting '.com' in the middle instead of at the end. The email would have been considered safe if the website address ended with '.com' or matched Amazon's official website format. A safe link from Amazon would look like: https://amazon.com/account or https://amaz on.it/account

- *Link mismatch.* The email shows a safe link but leads to 'facebook.com/hacked'. The email would have been safe if the link you see matched exactly where it takes you, which helps ensure you're going to the real Facebook website. A safe link would show and lead to the same place, like: facebook.com/help/security

Appendix A.4. Llama counterfactual explanations

- *IP address.* The link http://92.233.24.33 is a string of numbers instead of a normal website name. The email would have been considered safe if it used a recognisable website name, like a company's official name, which helps verify the link's authenticity. A safe link might have looked like: https://instagram.com/login

- *Top-level domain mispositioned.* The link https://amazonservices.com. cz/account.php has a suspicious web address with a ".com" in the wrong place. The email could have been considered safe if the ".com" was at the end of the address. A safe link might have looked like: https://amazon.com/account-update

- *Link mismatch.* The link text "protect your account" hides the actual link "https://www.facebook.com/hacked/disavow?u=100000125023309 &nArdInDS2&lit_IT&ext1548538159" The email would have been considered safe if the link text matched the actual link, which helps ensure the link is trustworthy and not trying to trick you. A safe link might have looked like: https://www.facebook.com/account-security.

Appendix A.5. Baseline - Manually-generated feature-based explanations

- *IP address.* Usually, websites use the URL instead of the IP address to make it easier for you to browse the web. However, an IP address was found in the email. Similar emails are harmful and steal private information. There is a potential risk of being cheated if you proceed."

- *Top-level domain mispositioned.* In the URL present in the email ("https://amazonservices.com.cz/ account.php"), the top-level domain (e.g., ".com") is in an abnormal position. This could indicate that the URL leads to a fake website. Such websites might steal your personal information.

- *Link mismatch.* This email reports a link that is different from the actual one "https://www.facebook.com/hacked/disavow? u=1000001250 23309&nArdInDS2&lit_IT&ext1548538159". This site might be intended to take you to a different place. You might be disclosing private information.

Appendix A.6. Logistic Regression Detailed Results

Table A.4: Logistic regression results on CTR. OR>1 increases click likelihood; OR<1 reduces it. Effect sizes: Small (OR 1.2–1.5/0.67–0.83), Moderate (1.51–2/0.5–0.66), Large (>2/<0.5). ORs between 0.84 and 1.19 are reported as Negligible.

Variable	**Condition**	**Emails**	OR	p	Effect
Familiarity	CF agg	ALL	2.20	.001	Large
	CF agg	TP	2.34	.001	Large
	CF agg	FP	2.53	.012	Large
	Claude agg	TP	1.83	.028	Large
	Claude CF	ALL	2.23	.033	Large
	Claude CF	TP	2.69	.022	Large
	Claude CF	FP	3.24	.050	Large
	Llama CF	ALL	2.14	.021	Large
	Llama CF	TP	2.05	.030	Large
Action="don't continue"	CF agg	ALL	0.21	<.001	Large
	CF agg	TP	0.18	<.001	Large
	NoLLM	ALL	0.14	<.001	Large
	NoLLM	TP	0.10	<.001	Large
	Llama agg	ALL	0.26	.002	Large
	Llama agg	TP	0.23	.001	Large

(continued on next page)

(continued from previous page)

Variable	Condition	Emails	OR	p	Effect
	Claude agg	ALL	0.41	.034	Large
	Claude agg	TP	0.34	.023	Large
Gender=female	CF agg	ALL	0.006	.023	Large
	CF agg	FP	0.005	.003	Large
	CF agg	TP	0.041	.012	Large
	NoLLM	FP	0.0003	.001	Large
	NoLLM	TP	0.0038	$<.001$	Large
	Llama agg	FP	0.03	.011	Large
	Claude agg	TP	0.035	.022	Large
	Claude CF	ALL	0.012	.023	Large
	Llama CF	FP	0.010	.041	Large
Gender=male	CF agg	TP	2.62	.020	Large
	Llama agg	TP	2.13	.037	Large
	NoLLM	FP	0.13	.036	Large
NASA-TLX	FB agg	FP	0.52	.004	Moderate
	Llama agg	ALL	0.63	.008	Moderate
	Llama agg	TP	0.64	.013	Moderate
	CF agg	ALL	0.65	.012	Moderate
	CF agg	TP	0.64	.012	Moderate
	Claude agg	ALL	0.74	.050	Moderate
	Claude agg	FP	0.61	.018	Moderate
	Claude FB	FP	0.47	.033	Large
	Llama FB	FP	0.45	.036	Large
Understandability	CF agg	ALL	0.53	.044	Moderate
	CF agg	TP	0.51	.042	Moderate
	Claude agg	FP	0.48	.046	Large
Trust warning	Llama FB	FP	0.44	.038	Large
Avg hours online/day	NoLLM	ALL	1.24	.024	Small
	NoLLM	TP	1.22	.006	Small
Age: Avg hours online/day	FB agg	ALL	1.007	.025	Negligible
	CF agg	ALL	1.007	.025	Negligible

(continued on next page)

(continued from previous page)

Variable	**Condition**	**Emails**	OR	p	Effect
	Llama agg	ALL	1.007	.025	Negligible
	Claude agg	ALL	1.007	.025	Negligible
	NoLLM	TP	1.02	.027	Negligible
	Claude CF	FP	0.93	.014	Negligible
Expertise:Avg hours online/day	Claude FB	ALL	1.07	.042	Negligible
	Llama CF	ALL	1.11	.039	Negligible
Gender=Male: Need-for-Cognition	CF agg	FP	1.44	.001	Moderate
	CF agg	ALL	1.21	.008	Small
	Claude CF	ALL	1.33	.023	Small
	Claude CF	FP	2.15	.012	Large
	Claude CF	TP	1.32	.024	Small
	Llama agg	ALL	1.04	.032	Negligible
Expertise:Need-for-Cognition	Llama agg	ALL	1.03	.032	Negligible
Age:Expertise	CF agg	ALL	0.98	.048	Negligible
	CF agg	FP	0.97	.045	Negligible
Gender=Male:Age	Llama agg	TP	0.88	.02	Negligible