Enabling Augmented Reasoning in Email Phishing Defence

Lorin Schöni
ETH Zurich

Adrian Brechtken
ETH Zurich

Abstract

Phishing warnings often fail because opaque technical signals force users to rely on habit rather than evidence. We introduce a human-centred interface that shifts from static binary alerts to augmented reasoning. In our on-going research, we evaluate underlying system features to determine how best to provide adaptive visual explanations. A cognitive walkthrough reveals that users strongly prefer a hybrid approach: a low-friction Banner for rapid triage paired with in-situ Highlights for granular evidence. By spatially coupling risk indicators directly with textual evidence, our design aims to manage trust miscalibration and reduce warning fatigue.

1 Introduction

Phishing persists by exploiting the gap between opaque technical detection and limited human attention. Automated systems typically exclude users from decision-making [14], yet relying entirely on manual vigilance causes warning fatigue and habituation [1, 19]. Generic binary warnings are insufficient to disrupt habitual behaviours, particularly in edge cases requiring human judgment, such as ambiguous emails.

We propose moving from static alerts to augmented reasoning. High technical accuracy is insufficient if a system remains an opaque "black box". Following human-computer collaboration principles [2, 17], users must understand the underlying reasoning to calibrate their trust accurately. We evaluate Large Language Model (LLM) setups to translate technical signals into interpretable evidence. By aligning visual explanations directly with the user's locus of attention, we aim to reduce the cognitive load needed for verification. This work details our design rationale and presents findings from a cognitive walkthrough to guide the development of transparent, user-assisting security tools.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

2 Related Work

Due to the increasing integration of LLMs into cybersecurity, detection shifts from static rules to dynamic reasoning. With this shift, techniques like semantic analysis and multi-agent debates are increasingly devised for user-facing security tasks [12, 13, 20]. However, these models typically operate as opaque "black boxes". In phishing contexts where uncertainty triggers habitual behaviours [14], accuracy without clarity quickly degrades trust [9]. Research shows that contextual explanations prove far more effective at building decision confidence than binary labels or static dashboards [7, 21]. Furthermore, user engagement drops sharply when security interventions fail to provide clear, actionable feedback [5] or ignore individual variations in user proficiency [15].

To address these limitations, we ground our design in warning science [22] to disrupt automatic processing [8] as modelled by the SCAM framework [19]. By combining question-driven XAI [11] with visual information-seeking principles [16], we attempt to introduce just the amount of cognitive friction needed. Such an approach is ideal at fostering appropriate reliance [4] and encourages deeper user engagement through targeted explanations [6].

3 System Design & Development

We developed a functional prototype as a Chrome Extension for Outlook Web, supported by a Python backend. To balance generalisability with complex reasoning demands, we selected an interactive, text-only, remote architecture (see table 1 in section 5). Early evaluations of a single-agent approach revealed consistency issues that risked undermining user trust. To resolve this, we transitioned to a multi-agent

debate architecture where specialised agents collaborate to extract context and synthesise a threat classification. While this introduces a minor latency trade-off, dividing the workload significantly improves output reliability. We manage these interactions through a custom YAML-driven framework, allowing dynamic runtime adjustments to the debate structure and enabling organisations to tailor the system's strictness to their own threat models.

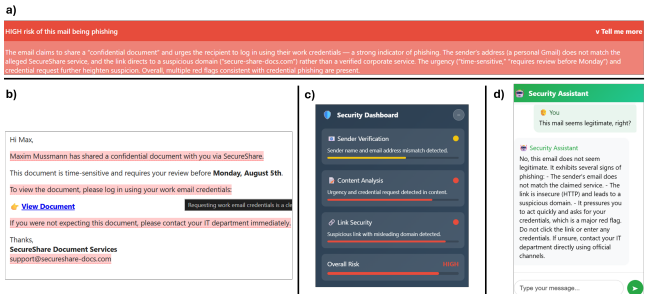

Figure 1: The interface variants implemented in our prototype. Users can toggle between: (a) a concise Banner, (b) in-text Highlights, (c) a detailed Dashboard, and (d) a conversational Chatbot. The example shows responses to a high-suspicion email across all four views.

4 Preliminary Evaluation & Results

Data collection. We conducted a cognitive walkthrough to assess usability and decision-support capabilities. All participants provided informed consent and no personal information was collected, adhering to institutional and established ethical standards [3]. Nine participants navigated a structured scenario involving benign and high-risk emails, interacting with the core interface modes (see fig. 1). We evaluated ease of use, clarity of feedback, user trust, decision-making support, and satisfaction, achieving thematic saturation after eight walkthroughs. The detailed scenario is provided in section 5.

Results. Users demonstrated a clear preference towards low-friction feedback during initial triage. The Banner's traffic-light metaphor was praised for providing immediate visibility without demanding high cognitive effort. When classification required verification, the Dashboard offered helpful structure. However, its spatial separation from the email body created a contextual disconnect; users likened it to an intrusive pop-up rather than a native security layer, and reported lower perceived trustworthiness. Conversely, the Highlights view was rated highest for engagement because it grounded risk in in-situ evidence (e.g., urgency keywords) rather than abstract summaries.

Main Usability Patterns. Transparency and visual integration emerged as the primary trust drivers. Participants consistently demanded specific risk indicators traceably linked to the LLM's reasoning; when this linkage was weak, trust dropped. Notably, users expressed a strong preference for a hybrid interface: using a Banner for immediate alerting, combined with Highlights for on-demand evidence verification. However, interaction friction heavily constrained adoption of conversational agents: Users rejected the Chatbot for rapid email triage, instead preferring proactive, structured explanations over querying they perceived as time-consuming.

5 Discussion

Our findings show that while technical accuracy is necessary, user trust depends on interaction granularity. Participants required varying evidence levels depending on their doubt. This mirrors question-driven explanation needs [11] and Shneiderman's visual mantra [16]. The Banner provided an initial overview but failed to support the deeper inquiry needed for trust calibration [9]. The Dashboard offered depth but suffered from a lack of natural integration into the email interface.

The preference for the Highlights view shows effective security tools should align with the user's locus of attention [23] and be spatially coupled with the threat. Such in-context highlighting fosters augmented reasoning [21] and shifts users from passive compliance to active evaluation by showing precisely where phishing cues are located and what makes them concerning.

This immediacy carries risks. Incorrect visual explanations can trigger a misinformation effect, where users internalise flawed AI reasoning [18]. While aiming to increase human agency [10] and empower users [17], designs should prevent over-reliance. Cognitive forcing functions offer one solution [4]. Yet, participants firmly rejected forced interaction through a Chatbot, preferring proactive surfacing of evidence over reactive querying.

Future Work. We plan to expand the UI to support dynamic autonomy levels. Moving beyond passive explanations, the updated system should allow users to configure responses ranging from visual guidance to the active remediation of malicious elements. We also intend to test the system using a inbox triage task focusing on ambiguous edge cases where AI certainty is imperfect. We aim to determine whether the hybrid design supports sustained reasoning over time, and whether in-situ evidence mitigates the misinformation effect [18] during false positives.

Acknowledgements

This work is graciously supported by armasuisse Science and Technology. This work was supported by a scholarship of the German Academic Exchange Service (DAAD).

References

[1] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Commun. ACM*, 42(12):40–46, December 1999.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–13, New York, NY, USA, May 2019. Association for Computing Machinery.

[3] APA. Ethical principles of psychologists and code of conduct, 2017.

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, 5:188:1–188:21, April 2021.

[5] Xiaowei Chen, Sophie Doublet, Anastasia Sergeeva, Gabriele Lenzini, Vincent Koenig, and Verena Distler. What motivates and discourages employees in phishing interventions: An exploration of expectancy-value theory. In *Proceedings of the Twentieth USENIX Conference on Usable Privacy and Security*, SOUPS '24, pages 487–506, USA, August 2024. USENIX Association.

[6] Md Montaser Hamid, Jonathan Dodge, Andrew Anderson, and Margaret Burnett. "Loss in Value": What it revealed about WHO an explanation serves well and WHEN. *CEUR Workshop Proceedings*, 3957:9–15, 2025.

[7] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, pages 907–924, New York, NY, USA, March 2025. Association for Computing Machinery.

[8] Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

[9] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–14, New York, NY, USA, May 2019. Association for Computing Machinery.

[10] Vivian Lai and Chenhao Tan. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 29–38, New York, NY, USA, January 2019. Association for Computing Machinery.

[11] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–15, New York, NY, USA, April 2020. Association for Computing Machinery.

[12] Ruofan Liu, Yun Lin, Xiwen Teoh, Gongshen Liu, Zhiyong Huang, and Jin Song Dong. Less defined knowledge and more true alarms: reference-based phishing detection without a pre-defined reference list. In *Proceedings of the 33rd USENIX Conference on Security Symposium*, SEC '24, pages 523–540, USA, August 2024. USENIX Association.

[13] Hyunsol Mun, Jeeeun Park, Yeonhee Kim, Boeun Kim, and Jongkil Kim. PhiShield: An AI-Based Personalized Anti-Spam Solution with Third-Party Integration. *Electronics*, 14(8):1581, January 2025. Publisher: Multidisciplinary Digital Publishing Institute.

[14] Tarini Saka, Kalliopi Vakali, Adam D G Jenkins, Nadin Kokciyan, and Kami Vaniea. Judging Phishing Under Uncertainty: How Do Users Handle Inaccurate Automated Advice? In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–18, New York, NY, USA, April 2025. Association for Computing Machinery.

[15] Lorin Schöni, Neele Roch, Hannah Sievers, Martin Strohmeier, Peter Mayer, and Verena Zimmermann. It's a Match - Enhancing the Fit between Users and Phishing Training through Personalisation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–25, New York, NY, USA, April 2025. Association for Computing Machinery.

[16] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343, September 1996. ISSN: 1049-2615.

[17] Ben Shneiderman and Ben Shneiderman. *Human-Centered AI*. Oxford University Press, Oxford, New York, January 2022.

[18] Philipp Spitzer, Joshua Holstein, Katelyn Morrison, Kenneth Holstein, Gerhard Satzger, and Niklas Kühl.

Don't Be Fooled: The Misinformation Effect of Explanations in Human–AI Collaboration. *International Journal of Human–Computer Interaction*, 0(0):1–29, November 2025. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2025.2574511.

[19] Arun Vishwanath, Brynne Harrison, and Yu Jie Ng. Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. *Communication Research*, 45(8):1146–1166, December 2018.

[20] Ngoc Tuong Vy Nguyen, Felix D Childress, and Yunting Yin. Debate-Driven Multi-Agent LLMs for Phishing Email Detection. In *2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5, April 2025. ISSN: 2768-1831.

[21] Yizhu Wang, Haoyu Zhai, Chenkai Wang, Qingying Hao, Nick A. Cohen, Roopa Foulger, Jonathan A. Handler, and Gang Wang. Can you walk me through it? explainable SMS phishing detection using LLM-based agents. In *Proceedings of the Twenty-First USENIX Conference on Usable Privacy and Security*, SOUPS '25, pages 37–56, USA, August 2025. USENIX Association.

[22] Michael S Wogalter, Vincent C Conzola, and Tonya L Smith-Jackson. Research-based guidelines for warning design and evaluation. *Applied Ergonomics*, 33(3):219–230, May 2002.

[23] Min Wu, Robert C. Miller, and Simson L. Garfinkel. Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 601–610. Association for Computing Machinery, 2006.

Walkthrough Scenario & Guideline

We conducted cognitive walkthroughs, which lasted for 30 to 45 minutes. The goal was to gather feedback on the interface and identify frictions and preferences among different users. Each walkthrough was conducted by two researchers, with one person participating. After providing informed consent, one researcher accompanied the user with navigating the tool, whereas a second researcher noted all interactions and exchanges. On a rolling basis, we conducted a light thematic analysis to count frequencies, commonalities, and differences between each participant's preferences. After 8 walkthroughs, the researchers noted thematic saturation, with no novel insights provided by the eight walkthrough. In the following, we detail the walkthrough procedure.

Intended User journey

The user starts Outlook. The extension menu pops up, but the extension is disabled by default. **Before** checking any mails, the user activates the extension by clicking the toggle and selects "**Banner**" from the now activated extension dropdown. Then, the user starts reading mails.
First, they click the most recently received mail (which should be a legitimate mail). While the user reads the mail, the banner evaluates to LOW risk.
The user clicks on the second most recent mail (which should be phishing). While the user reads the mail, the banner evaluates to HIGH risk. The user clicks on "Show me more" to understand the classification. After reading the rather long banner explanation, the user decides to switch to the more concise dashboard by selecting "**Dashboard**" in the extension dropdown.
The user clicks on the next mail (which should be phishing). While the user reads the mail, the dashboard evaluates to high risk. The user did not expect this and wants to know which parts seem suspicious. They switch to the highlights feature by clicking "**Highlights**" in the extension dropdown. The user waits for the highlights to be shown and then hovers over the first two, reading their tooltips for further explanation. The user finally switches to the "**Chatbot**" and asks it to evaluate the email.

Procedure and Questions

1. **Activating the Extension and Selecting "Banner"**
    - Was it clear how to activate the extension? *(Ease of use)*
    - Did you understand what the "Banner" option would do before selecting it? *(Clarity of feedback)*
    - How confident were you that the extension was ready to use after activation? *(User trust and confidence)*
2. **Reading a Legitimate Mail (LOW Risk Banner)**
    - Was the LOW risk banner easy to notice and interpret? *(Clarity of feedback)*
    - Did the banner influence your perception of the email's safety? *(User trust and confidence)*
    - Did you feel the banner provided enough information without needing "Show me more"? *(Decision-making support)*
3. **Reading a Phishing Mail (HIGH Risk Banner)**
    - Did the HIGH risk banner catch your attention effectively? *(Clarity of feedback)*
    - What would motivate you to click "Show me more"? *(Decision-making support)*
    - Was the explanation clear and helpful in understanding the risk? *(Clarity of feedback)*

- Did the length or detail of the explanation feel appropriate? *(Preference and satisfaction)*

4. **Switching to "Dashboard" View**

   - What would make you choose switching to the Dashboard view? *(Preference and satisfaction)*
   - Was it easy to find and switch views? *(Ease of use)*
   - Did the Dashboard feel more or less helpful than the Banner? Why? *(Decision-making support)*

5. **Reading Another Phishing Mail (Unexpected HIGH Risk)**

   - Did the HIGH risk evaluation surprise you? *(User trust and confidence)*
   - Did you trust the tool's judgment? *(User trust and confidence)*
   - What did you want to understand better at this point? *(Decision-making support)*

6. **Switching to "Highlights" View**

   - Was it clear how to switch to Highlights view? *(Ease of use)*
   - Did the highlighted elements help you understand what was suspicious? *(Decision-making support)*
   - Were the tooltips informative and easy to understand? *(Clarity of feedback)*
   - Did this view improve your confidence in the tool's assessment? *(User trust and confidence)*

7. **Switching to "Chatbot" View**

   - Do you think the Chatbot is easy to interact with? *(Ease of use)*
   - What would make you use the Chatbot? *(Preference and satisfaction)*

Final Reflection Questions

- Which view (Banner, Dashboard, Highlights, Chatbot) did you find most helpful overall? Why? *(Preference and satisfaction)*
- Did the tool support your decision-making effectively? *(Decision-making support)*
- Was there any point where you felt confused or unsure? *(Ease of use)*
- What would you improve or change about the experience? *(Preference and satisfaction)*

Development Details

Table 1: Design space and rationale. The **left** column indicates the selected configuration for our prototype.

Design Level	Selected Approach	Alternative Considered
Presentation	**Interactive** *Why: Increases user awareness and trust through active feedback loops.*	**Static (Passive/Disruptive)** *Why not: Ignored by users (passive) or causes annoyance/fatigue (disruptive).*
Model Input	**Text-Only** *Why: Ensures generalisability across platforms and lowers latency.*	**Multimodal (HTML/Image)** *Why not: High computational cost (OCR) and structural parsing complexity.*
Hosting	**Online (Remote)** *Why: Enables continuous model updates and wider accessibility.*	**Local (On-Device)** *Why not: Limited scalability and harder to patch security vulnerabilities.*
Architecture	**Multi-Agent** *Why: Specialised agents handle distinct subtasks collaboratively.*	**Single Agent** *Why not: Monolithic structure struggles with complex, multi-step reasoning.*
Tuning	**Prompt Engineering** *Why: Lightweight and allows rapid iteration without retraining.*	**Fine-Tuning (Weights)** *Why not: Resource-intensive and risks catastrophic forgetting.*