# Design and Evaluation of a Human-Centered Explainable AI Dashboard for Cybersecurity Education

**Author:** Dr. Steven A. Schilhabel

**Institution:** Information Systems Department, University of Wisconsin – Oshkosh

Address: 800 Algoma Blvd, Oshkosh, WI 54901, United States

**Email:** schilhabels@uwosh.edu

**ORCID**: 0009-0009-2566-7606

**Corresponding Author:** Dr. Steven A. Schilhabel (schilhabels@uwosh.edu)

# Design and Evaluation of a Human-Centered Explainable AI Dashboard for Cybersecurity Education


## Abstract

While artificial intelligence (AI) is applied more often in cybersecurity training and instruction, the black-box nature of AI systems' decision-making processes can detract from their value for education and impair the trust of the users. In this paper, we share how we designed, prototyped, and evaluated a Human-Centered Explainable AI (HC-XAI) dashboard as a tool for supporting and empowering cybersecurity learning with a focus on phishing threat identification. The artifact was developed in response to educational challenges from the field, following the principles of Design Science Research (DSR). It incorporates three complementary explanation modalities, including: (1) rule-based logic; (2) natural language explanations, generated using large language models; and (3) visual heatmap visualizations of token-level attention. The HC-XAI dashboard was rigorously evaluated using a two-phase methodology, which included both an expert heuristic walkthrough and a mixed-methods user study with a sample of 23 cybersecurity students. The findings demonstrate the dashboard's strong usability, positive impact on learner trust, and variation in user preference across the three modalities. The paper's contributions are fourfold: (1) to the literature, the study shares how to operationalize and evaluate key design choices of HC-XAI; (2) to cybersecurity training practice, the work presents actionable directions for educational program designers and teachers on how to use explainable AI for improving IS and cybersecurity students' education and learning; (3) to DSR research, the work makes a unique contribution in the context of user training and shows how design artifacts can support not only technical practice but classroom activities as well; and (4) to the research

community, the study provides a robust HC-XAI dashboard artifact as a proof-of-concept to inspire and support applied, student-centered research.

**Keywords:** Explainable Artificial Intelligence (XAI), Human-Centered AI, Cybersecurity Education, Design Science Research, Usability, Trust in AI

## 1. Introduction

Artificial intelligence (AI) has become a standard and ubiquitous tool in cybersecurity, driving many applications from phishing detection to intrusion monitoring and anomaly detection. While these AI/ML-based systems achieve high accuracy, they often operate as "black boxes" and are notoriously hard to understand or explain, creating adoption and trust issues. For instance, it is well-documented that security analysts regularly ignore and "turn off" model outputs due to this opacity, a phenomenon known as algorithmic aversion (Jin et al., 2021). This is especially problematic for students, who need to be able to explain and interpret what is going on under the hood in order to understand AI-augmented predictions or recommendations (Dietvorst, Simmons, & Massey, 2015; Logg, Minson, & Moore, 2019).

Explainable artificial intelligence (XAI) methods were developed to tackle the "black box" problem by providing more transparency, interpretability, and trust in and of AI systems (Adadi & Berrada, 2018; Gunning et al., 2019). However, existing XAI solutions are typically designed to meet the needs of skilled data scientists and operate at scale in high-risk/high-stakes domains, such as healthcare and finance (Linardatos, Papastefanopoulos & Kotsiantis, 2020). In contrast, the vast majority of cybersecurity education utilizes static text tutorials or, at most, simplified "training wheels" dashboards that do not enable students to see the intermediate reasoning steps behind a black box model's phishing classifications. Therefore, students are not able to deeply understand or manually retrace the logic behind these AI-augmented threat predictions.

Human-Centered XAI (HC-XAI) is a new emerging branch of XAI work that places a special focus on user-centered explanations, adapting the type and level of explanations to the user's needs, cognitive style, and learning context (Ehsan & Riedl, 2020). While HC-XAI has a great deal of potential to be leveraged in an educational context, to the best of our knowledge, little work has been done to understand how such tools can be integrated into training environments to help students understand and make use of AI-augmented decision-making. In particular, only a small number of applied studies have examined how data science students, a population of expert data scientist users (Nguyen et al., 2020), or novice end users (such as students using explainable AI tools for the first time) interact with multimodal explanations. Even less work has been done to understand the impact of design features on the trust, clarity, and cognitive effort of explanations in an educational context.

In this paper, we design, implement, and evaluate a Human-Centered Explainable AI (HC-XAI) dashboard for phishing threat detection that was specifically developed for the context of cybersecurity training and education. Leveraging the principles of Design Science Research (DSR), we create a novel artifact that incorporates three different explanation modalities: rule-based logic, natural language explanations from a large language model, and token-level visual heatmaps. We conduct a two-part evaluation of the XAI dashboard, first performing an expert walkthrough using Nielsen's heuristic evaluation criteria to identify usability and explainability issues, and second, a mixed-methods user study where we measure and compare the impact of different explanation modalities on students' comprehension and cognitive effort.

This paper makes three contributions to the research community. First, we design and evaluate a novel HC-XAI dashboard for cybersecurity education, demonstrating that by providing multimodal explanations, students can utilize the tool better and trust it. Second, we present an empirical study that measures the impact of different explanation modalities on students' understanding of the AI-augmented dashboard and cognitive effort. Third, we provide pedagogical recommendations and practical advice for instructors and teachers to include the use of explainable AI dashboards in their Information Systems and cybersecurity training curricula to prepare students for their future workplace better. Unlike most existing XAI dashboards developed for professional analysts, our artifact is explicitly designed for educational settings, bridging a gap in the literature by adapting multimodal explanations to novice learners."

The remainder of the paper is organized as follows. Section 2 provides background on related work on XAI and HC-XAI, their application to cybersecurity education, and the development of related artifacts and tools. Section 3 describes the Design Science Research methodology used. Section 4 details the artifact's design and implementation. Section 5 presents the evaluation, including both the heuristic walkthrough and user study, as well as the resulting data analysis and findings. Section 6 discusses the research and educational implications of the findings. Finally, Section 7 concludes with directions for future work.

## 2. Literature Review

As AI use in cybersecurity education becomes more prominent, it is increasingly important to situate this work within the broader context of other research on explainability, trust, and human-centered design. A literature review of other areas enables more clarity on the relationship between how explainable AI (XAI) has developed in response to issues with black-box models, how human-centered extensions of XAI have been built on XAI to produce personalized explanations to different users, and the unique challenges faced by cybersecurity and cybersecurity education as a context. The following subsections cover these bases, which are necessary to provide context for the current study in terms of AI transparency, human–AI interaction, and educational applications.

**Explainable AI (XAI) and Trust**

Artificial intelligence (AI) is rapidly expanding into various areas of human activity, including cybersecurity, finance, and healthcare. There have been mounting concerns related to the explainability of machine learning algorithms. End users have typically not been able to see or interpret how AI-based systems, especially those based on deep learning approaches, reach their conclusions. State-of-the-art machine learning models are described as “black boxes,” making them hard to adopt in high-stakes environments in which both accountability and the need to justify an outcome have been demanded (Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017).

In recent years, the field of explainable artificial intelligence (XAI) has been developed, which aims to provide better transparency and explainability for machine learning algorithms, without sacrificing prediction performance. In the U.S., DARPA (Defense Advanced Research Projects Agency) has initiated the XAI program to spur research in the field, with an explicit focus on human-machine teaming and user trust (Gunning & Aha, 2019).

Trust has been identified as a key factor influencing end-user adoption of AI. In organizational behavior, for example, studies have indicated that humans are likely to reject an algorithm’s output after one or two small mistakes. This phenomenon is referred to as “algorithm aversion” (Dietvorst, Simmons, & Massey, 2015). In another experiment, when presented with a choice between algorithms and human advice, participants were willing to follow the algorithms’ recommendations with an imperfect record, as long as they were given some explanations about the process (Logg, Minson, & Moore, 2019). Trust in AI systems is a more complex construct than accuracy alone.

Cybersecurity is one of the areas where the lack of explainability in AI can be damaging. Security analysts are required to make quick judgments in highly uncertain conditions. In the event of an attack, for instance, a time delay can be critical. A system that raises an alert without explanation can be left unused or relied upon too much. In both cases, the system is less effective than it could be if end users could better calibrate their trust (Carvalho, Pereira, & Cardoso, 2019). XAI can help in cybersecurity with both transparency and adoption, and as a human-in-the-loop support, enabling the human end user to question a particular AI classification decision.

**Human-Centered Explainable AI (HC-XAI)**

Existing XAI research has predominantly treated explanations as an endpoint for opening the "black box" through technical interpretability techniques. However, a parallel line of work, in Human-Centered Explainable AI (HC-XAI), has argued that explanations must not only be technically sound but also human-centric by addressing human cognition and task context (Ehsan & Riedl, 2020; Liao & Varshney, 2022). In HC-XAI, explanation quality depends on the correctness of a model's internal working, as well as factors related to the design of explanations, which take into account their presentation, user experience, and the understandability of the target model (Ehsan et al., 2021).

Reviews of recent HC-XAI research in social sciences journals have identified three common focuses across this work: (1) building interactive explanation systems to support user exploration and information needs, (2) designing explanations for users of different levels of expertise, and (3) adapting explanation approaches based on cognitive theory to evaluate mental workload and measure calibrated trust (Rong et al., 2023; Kim, Maathuis, & Sent, 2024). This finding aligns with the results in the broader human–computer interaction literature, which suggests that explanations are more effective when perceived as dynamic and interactive for human reasoning, rather than fixed text or image outputs (Ehsan et al., 2021).

HC-XAI studies also commonly integrate multimodal explanations combining complementary modalities to provide more complete and transparent reasoning. For example, researchers have combined natural language rationales with rule-based or adversarial examples with visual output (Wang et al., 2020). The provision of multiple modalities for explanation supports human-AI teaming by allowing a user to cross-validate AI decisions through different information sources, as well as to select the explanation modality preferred by the user or most relevant to the task context (Wang et al., 2020). Designing an approach that integrates multiple modalities and meets the HC-XAI design criteria, however, requires care not to overload users or to lead to misinterpretations (Alqaraawi, Schuessler, Weiß, & Kulesza, 2020).

Overall, by centering on usability, design, and user experience, HC-XAI work is directly applicable to the development of a cyber defense tool that can support student learning and explainable outcomes. In particular, the HCI-informed criteria of interactivity, adaptation to users, and calibrated trust provide a lens for understanding the value of integrating multiple information modalities for learners with varying experience and domain knowledge.

**XAI in Cybersecurity Contexts**

Cybersecurity is one of the fields in which the need for XAI has been described as particularly dire. State-of-the-art AI systems have been applied to a variety of cybersecurity tasks, including phishing detection (Alqaraawi, Schuessler, Weiß, & Kulesza, 2020), intrusion detection (Mohale & Obagbuwa, 2025), malware classification (Mohseni, Zarei, & Ragan, 2021), and anomaly detection (Stamm, Liebelt, Mayr, & Morik, 2019). However, without clear explanations of their rationale, security analysts cannot be expected to understand the model reasoning, much less trust and accept it, leaving such systems with little to no value in practice (Molnar, 2019; Mohseni, Zarei, & Ragan, 2021).

Empirical research on the interpretability needs of cybersecurity analysts is starting to emerge. Mohale and Obagbuwa (2025) point to interpretability as a key enabler for effective use of intrusion detection systems, since "applications of deep learning … which are mostly opaque, are likely to have a limited adoption by analysts, and are not easily amenable to the high level of accountability" (p. 1). Speith (2022) similarly observes that the vast majority of extant cybersecurity dashboards incorporate static, rule-based explanations, or post-hoc saliency maps (i.e., token-based attribution). She cautions that such approaches are ill-suited to the needs of security analysts and expert-level end users, for whom decision-making is often a real-time task with high stakes, requiring both accuracy and high confidence.

In the phishing detection task, the presence of XAI has been shown to impact user understanding and behavior. Alqaraawi et al. (2020) conducted a user study on visual saliency-based explanations, in which the rationale for a model's prediction was shown via a heatmap of suspicious terms/tokens. Although they found that XAI improves user understanding of model predictions, the authors caution that "care is needed in deciding how information is visualized and how much information is presented to the user" (p. 9) to avoid cognitive overload. Other work has presented rule-based explanations that, for example, specifically highlight "trigger words" (Verma et al., 2017) or heuristics like malicious URLs (Gunasekaran et al., 2022). Textual, natural language rationales generated using large language models have also been recently proposed (Kulapov & Kulapova, 2023; Weiß et al., 2023) as a means to present phishing explanations in human-readable language; however, they have been empirically evaluated to a much lesser extent.

Note, however, that a vast majority of XAI research in cybersecurity and other AI application areas focuses on professional cybersecurity analysts and end users in operational, as opposed to educational, environments. In this work, we ask how students, as a proxy for future cybersecurity experts, engage with multimodal, human-centered phishing explanations.

**Educational Applications of XAI and Cybersecurity Training Tools**

Research in explainable artificial intelligence (XAI) has primarily been applied to professional and expert user settings such as health and financial services. However, in education, students face the dual task of learning how to use AI systems and how to trust AI-based predictions. The lack of explainability mechanisms leaves students unable to interpret what they are being shown, e.g., correct or incorrect classification labels (Holstein et al., 2019).

In several IS and cybersecurity education streams, research has shown the usefulness of tool interaction, often with the use of dashboards, for providing guidance to students and facilitating learning. Dashboards have been used to provide students with a visualization of real-time data and the behavior of related systems for the development of analytical reasoning and in cybersecurity training, helping students recognize cyber attacks (Arora & Rahman, 2017; Davis, Dehlinger, & Hilburn, 2020; Kraemer, Carayon, & Clem, 2009). Interactive tools have been shown to facilitate student learning but have not, to date, been developed with embedded explainability components.

In XAI education research, Ribeiro, Singh, and Guestrin (2016) model-agnostic interpretability (LIME) framework has been used in classroom settings to "extract human-understandable rationales for how AI systems classify new data points" (Holstein et al., 2019, p.

318). Moreover, recent IS research in the area of visualization-based learning and student cognition has focused on artifacts, such as interactive dashboards, as a mechanism to link technical training and usability with student knowledge (Whitman, 2019). This is in line with prior work in the design science tradition for developing and using educational tools in higher education.

Few have explored human-centered, multimodal XAI in cybersecurity training. In prior research, much of the work has been applied to professionals and expert users in real-world settings. Few studies have investigated how users, in this case, students, interact with and react to XAI explanations, including the potential challenges that they encounter in terms of calibration of trust in the AI and the cognitive load that may be associated with it.

**Research Gap and Objectives**

The literature has illuminated the value of explainability to engender trust in AI (Holzinger et al., 2021; Liao & Varshney, 2022; Rieger et al., 2023), the potential of human-centered AI that personalizes explanations to users' needs (Ehsan & Riedl, 2020), and the utility of dashboards and interactive tools for IS and cybersecurity education (Whitman, 2019; Davis, Dehlinger, & Hilburn, 2020). However, several opportunities for further research have been identified. First, to date, the bulk of XAI scholarship for cybersecurity has been applied or prescriptive, targeting practicing analysts or operational systems (Mohale & Obagbuwa, 2025; Speith, 2022), with little attention to how novice learners might benefit from explainable AI systems in an educational context. Second, while HC-XAI research has called for adaptivity, multimodal explanations, and user-centered design (Ehsan & Riedl, 2020; Liao & Varshney, 2022), few applied studies have tested or validated how users, in this case students, engage with various types of AI explanations or how explanation design impacts comprehension, trust, or cognitive load. Third, while IS education scholarship has established the utility of dashboards and interactive learning artifacts (Whitman, 2019; Davis, Dehlinger, & Hilburn, 2020), many existing cybersecurity training tools do not yet include explainable AI elements that elucidate the rationale for an automated classification.

A body of research is needed to develop, test, and validate the integration of HC-XAI design principles into practical tools with applied outcomes for educational learners. The current study takes a step in this direction by both conceptualizing and validating a Human-Centered Explainable AI dashboard for phishing detection, designed to be applied in a cybersecurity education context. Informed by Design Science Research (DSR), this study has four primary goals:

1. Design an HC-XAI dashboard artifact that leverages multimodal explanations (rule-based explanations, natural language rationales, and visual heatmaps) to facilitate cybersecurity learning for students.
2. Assess the artifact's usability and trustworthiness using expert heuristic walkthroughs and mixed-methods user studies with cybersecurity students.
3. Investigate the impact of different explanation modalities on user comprehension, trust calibration, and cognitive load.
4. Offer prescriptive recommendations to IS and cybersecurity educators on the effective incorporation of HC-XAI tools into educational curricula to better prepare students for AI-augmented cybersecurity work environments.

Through these goals, the present research aims to contribute both to HC-XAI scholarship, through extending its application into education, and to cybersecurity education, through the design and validation of a pedagogically-oriented HC-XAI artifact.

# 3. Design and Development

This section presents the design and development process of the Human-Centered Explainable AI (HC-XAI) dashboard. Following the principles of Design Science Research, the artifact was developed as an iterative process that considered technical and user-centered aspects at each step. Each subsection addresses a different aspect of the development process, from the choice of explanation modalities to the implementation details, the refinements based on expert feedback, and its alignment with educational goals.

## Design Rationale

Guided by the key design principles of usability, cognitive accessibility, and educational value, the artifact was developed to meet the diverse needs of cybersecurity students. The HC-XAI dashboard makes use of three multimodal explanation approaches in the artifact, namely: (i) rule-based logic, (ii) natural language rationales, and (iii) token-level heatmaps. A multimodal approach was chosen due to the following considerations. As summarized in Table 1, each explanation modality was selected to align with different cognitive preferences and educational needs, while also recognizing the potential challenges of each approach.

Capturing Diverse Cognitive Preferences: Students have diverse ways of processing information. Some prefer the structure of rule-based logic, others may resonate more with natural language explanations, and some lean toward visual cues such as those provided by heatmaps. Offering a spectrum of explanation modalities ensures that students can find an approach that resonates with their unique style of learning and understanding.

Facilitation of Cross-Verification: Studies on human–AI teaming have found that layered explanations can allow users to calibrate their trust in the system by comparing results across modalities. In the context of the HC-XAI artifact, if a user is uncertain about the system’s highlighting of a suspicious URL in the rule-based approach, they could easily verify it by referencing the corresponding natural language explanation and visual prominence in the heatmap. This cross-modal validation not only enhances clarity but also boosts confidence.

Meeting Educational Needs: Students are inherently in a learning phase and, unlike domain analysts, often require detailed insights to grasp the logic behind classifications. The HC-XAI artifact's use of easy-to-display rules and succinct AI-driven narratives makes it especially fitting for educational environments, where interpretability should take precedence over other factors like processing speed or automation.

*Table 1 - Rationale for Explanation Modalities*

| Explanation Modality | Primary Purpose | Benefits for Students | Potential Challenges |
|---|---|---|---|
| **Rule-based Logic** | Show predefined phishing indicators (e.g., suspicious URLs, urgency keywords) | Provides transparency and reproducibility; aligns with textbook heuristics | May appear rigid or incomplete if indicators are not triggered |
| **Natural Language Rationale (LLM)** | Translate classification into a concise, human-like explanation | Enhances clarity; mirrors how instructors might explain a phishing case | Risk of over-simplification or variability in AI output |
| **Visual Heatmap** | Highlight tokens most influential in classification | Appeals to visual learners; supports interactive exploration | Can appear dense or confusing without additional scaffolding |

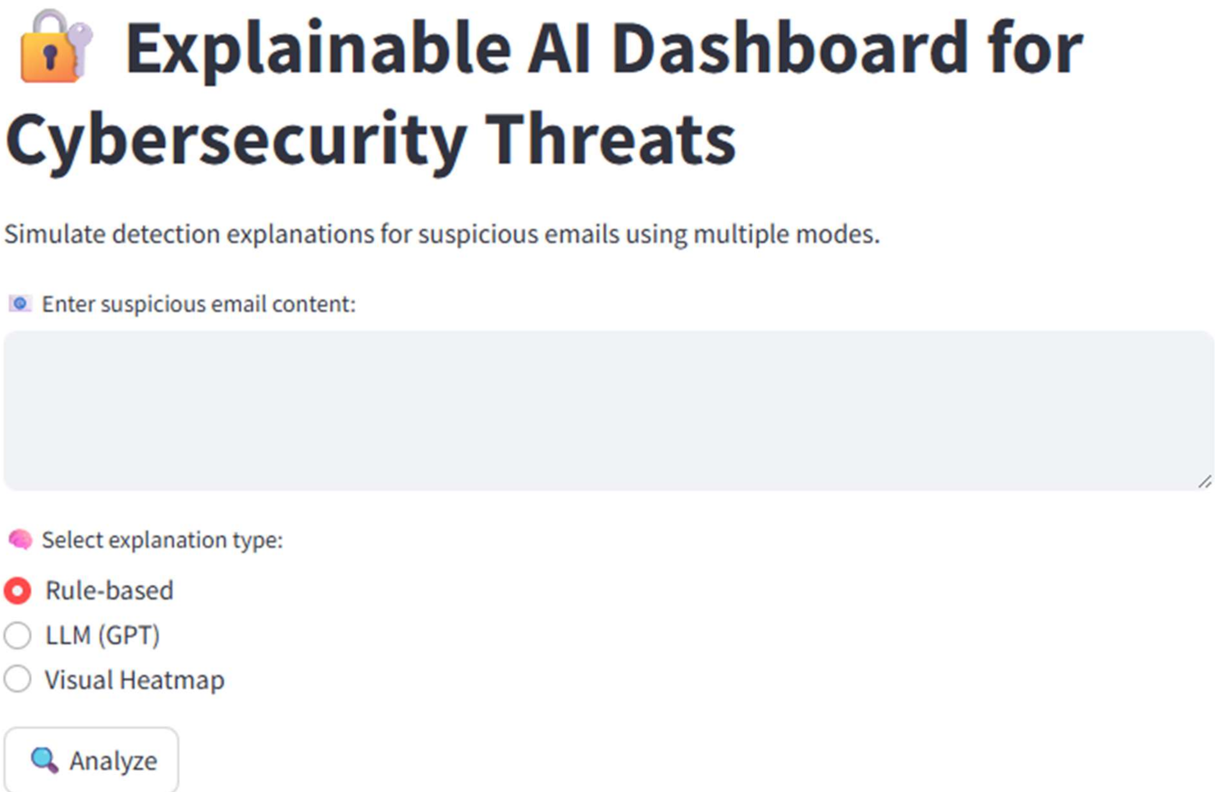


*Figure 1 – Interface of the Explainable AI Dashboard*

**Artifact Implementation**

HC-XAI Dashboard Artifacts Design and Implementation. The HC-XAI dashboard was implemented as a web-based application, with the dual goals of facilitating ease-of-access (important for in-classroom applications) and enabling rapid iteration and testing. Implementation was guided by DSR principles, with a focus on producing a usable and stable artifact that could be readily demonstrated as a proof of concept for multimodal explanations in an educational environment.

Implemented in Python with Streamlit, the artifact was lightweight and required no complex installation processes, running entirely in a browser. The three explanation modalities were unified into a single interface. The technical components and example outputs of each module are presented in Table 2, which shows how rule-based, natural language, and visual outputs work together to provide complementary explanations.:

1. Rule-based Logic Module – This rules engine detects and highlights known phishing indicators (e.g., urgent language, deceptive URLs, spoofed sender information).
2. LLM-based Explanation Module – Powered by OpenAI GPT, this component generates brief, natural language explanations tailored for non-expert users.
3. Visual Heatmap Module – A token-level saliency map visualization highlights suspicious words/phrases within the email text.

Dashboard Workflow: User inputs suspected phishing email into the dashboard. The system processes the input through each explanation module and presents the results side by side for cross-comparison and interaction.

*Table 2 - Overview of Dashboard Components*

| Component | Technology / Approach | Function | Output Example |
|---|---|---|---|
| **Rule-Based Logic** | Python-based rules engine | Detects and flags phishing indicators | "Suspicious URL detected: http://example-login.com" |
| **LLM Explanation** | OpenAI GPT API (temperature = 0.3, truncated output) | Generates concise natural language rationale for classification | "This email shows urgency and uses a deceptive link, which are common phishing tactics." |
| **Visual Heatmap** | Token-level attention weights (saliency mapping) | Highlights words contributing most to classification | Color-coded words (e.g., "urgent," "verify," "secure link") |

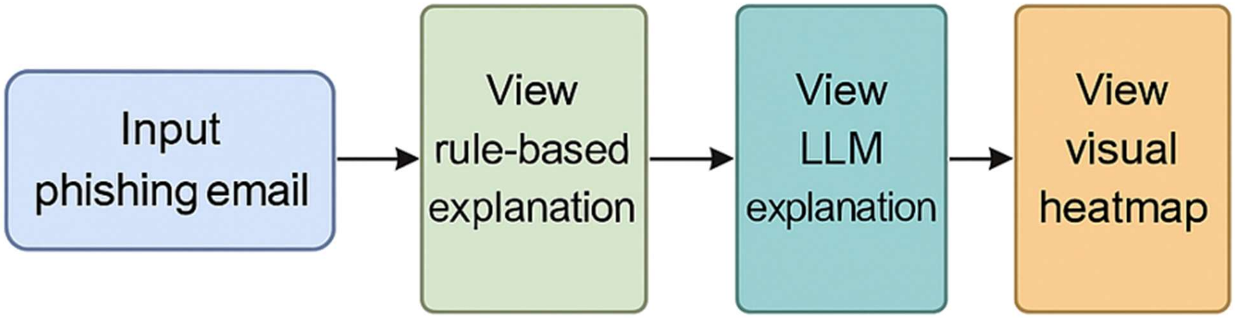


*Figure 2 - Sequence of tasks users perform in the HC-XAI dashboard, including input, analysis, and explanation review.*

The interaction process is further illustrated in Figure 2 (Appendix G), which depicts the sequence of user tasks within the dashboard interface. Ease of use for classroom setting: Streamlit was chosen to minimize technical barriers for classroom use. Determinacy of the outputs: Parameters (temperature 0.3, max length 100) were optimized to return small, deterministic explanations for educational purposes. Visibility: Code was modularized to allow better students to understand the provenance of each rule, prompt for model generation, and the underlying logic of the visualizations.

**LLM Explanation Setup**

Natural Language Explanation Module: At the heart of the HC-XAI dashboard is the natural language explanation module, which leverages a large language model (LLM) to provide brief, student-friendly rationales for phishing classifications. The module was developed adhering to human-centered explainable AI (HC-XAI) principles, focusing on clarity, brevity, and consistency rather than technical specificity or creativity.

Prompt Design: Prompts were designed to elicit non-technical, educationally accessible language. They focused on the four phishing indicators most often covered in cybersecurity training: (1) urgency, (2) deceptive URLs, (3) disguised sender details, and (4) emotionally manipulative language. Responses were limited to under 100 tokens to avoid cognitive overload.

The design parameters for the natural language explanation module are detailed in Table 3, including the rationale for prompt content, output length, and consistency controls.

Output Stability and Control: The model's temperature parameter was set to 0.3 to reduce randomness and maintain consistent experiences across student users, while still ensuring readability. Truncation safeguards were implemented to prevent overly verbose or out-of-context outputs.

Expert Validation: A panel of five expert raters reviewed initial outputs on a 5-point Likert scale (clarity, accuracy, domain alignment). Prompt refinement iterations continued until 90% of explanations received a score of 4 or higher on all three dimensions.

*Table 3 - Prompt Content Rationale for Students*

| Design Element | Implementation Choice | Rationale for Students |
|---|---|---|
| **Prompt Content** | Focus on urgency, URLs, sender, and emotion | Aligns with common phishing heuristics in coursework |
| **Response Length** | ≤ 100 tokens | Prevents cognitive overload, ensures clarity |
| **Temperature Setting** | 0.3 (low variability) | Provides consistent, reproducible outputs for learning |
| **Validation Process** | Expert Likert review, iterative prompt tuning | Ensures explanations are accurate and comprehensible |

Figure 4 (Appendix F) provides a sample output from the LLM module, showing how a phishing email is translated into a concise natural language rationale for students.

## Iterative Refinement

HC-XAI Dashboard Usability Refinements The HC-XAI dashboard was refined through an iterative development cycle that incorporated feedback from both expert reviewers and early student testers. This process ensured that the artifact aligned with usability, interpretability, and educational accessibility goals. Table 4 summarizes the iterative development process, documenting the issues identified at each phase and the corresponding refinements implemented.

Phase 1: Expert Heuristic Walkthrough An initial prototype was evaluated by a panel of domain experts using a customized HC-XAI heuristic checklist. Feedback emphasized the need for clearer terminology in the rule-based module, improved accessibility of heatmaps, and onboarding support for first-time users.

Phase 2: Prototype Adjustments. In response to expert feedback, several targeted modifications were made:

- Rule-based logic: Simplified technical terms into plain language (e.g., “malicious URL” → “suspicious link”).
- Heatmaps: Adjusted color gradients for readability, especially for color-blind users.
- LLM prompts: Reworded for brevity and consistency across explanations.
- User onboarding: Added a short tutorial to guide first-time users.

Phase 3: Pre-Study Testing. A revised prototype was shared with a small pilot group of students prior to the formal evaluation study. Feedback confirmed improved clarity but suggested further improvements to visual explanations, such as tooltips for highlighted terms.

*Table 4 - Iterative Refinement Summary*

| **Development Phase** | **Feedback Source** | **Key Issues Identified** | **Refinements Implemented** |
|---|---|---|---|
| **Expert Walkthrough** | 5 domain experts | Dense terminology, color contrast issues, and a lack of onboarding | Simplified language, heatmap gradient adjustment, and added a tutorial |

| | | | |
|---|---|---|---|
| **Prototype Adjustments** | Internal dev team | Inconsistent LLM phrasing | Reworded prompts, capped token length |
| **Pre-Study Testing** | Pilot student group | Visual overload in heatmaps | Added tooltips and simplified visual highlighting |

## Educational Orientation

The HC-XAI dashboard was designed with educational use cases, information systems, and cybersecurity programs in mind. While there are various explainable AI systems designed for professional analysts working in high-stakes environments, this artifact was created for teaching and learning purposes. Table 5 highlights the educational features intentionally built into the dashboard, connecting technical design choices with classroom applications.

In the classroom: The HC-XAI dashboard was designed to be used as an ancillary learning platform alongside other cybersecurity curricula. As such, the intention was to allow students to:

- Explore phishing examples with multimodal explanations in an interactive environment
- Map theoretical concepts to system outputs (e.g., map common phishing heuristics to the LLM-generated explanation)
- Critique and analyze the outputs of different explanation modalities

Pedagogical Design Considerations: A few design decisions in the dashboard were intended for students and instructors:

- Use of plain language in the rule-based and LLM modules to make the explanations more accessible for novices.
- Inclusion of visual scaffolding cues (heatmaps, tooltips) to help students with multimodal preferences.

- Low technical threshold (web-deployed) to allow for use both in a classroom setting or remotely without any technical barriers to set up.
- Utility for instructors, where the explanations could be displayed onscreen in class or assigned as a part of lab activities.

*Table 5 - Educational Features*

| Feature | Educational Purpose | Classroom Benefit |
|---|---|---|
| **Rule-based logic in plain terms** | Links technical detection rules to beginner-friendly language | Helps students bridge textbook heuristics and system reasoning |
| **Natural language explanations** | Provides narrative-style rationale | Mirrors instructor explanations; aids comprehension |
| **Heatmaps with tooltips** | Highlights key suspicious terms visually | Supports visual learners and interactive exploration |
| **Web-based deployment** | Simple browser access | Reduces IT setup; easy for labs and online teaching |
| **Instructor demonstration mode** | Allows use in lectures and labs | Supports active, discussion-based learning |

# 4. Evaluation

After following the design and development process as explained in the previous section, the HC-XAI dashboard needed to be evaluated to assess its usability in an educational setting. The evaluation of the system, based on the Design Science Research guidelines presented in the first chapter, followed both the expert view and student view with the evaluation and assessment of usability, interpretability, and trust. This section explains the two-step evaluation process and presents the quantitative and qualitative results.

## Evaluation Overview

The HC-XAI dashboard artifact was evaluated through a two-phased approach to understand expert and student feedback and to test the system more effectively. In line with DSR, the goal of the evaluation was to evaluate the artifact's usability and trust implications as well as the quality and clarity of its explanations.

Phase 1: Expert heuristic walkthrough using human-centered XAI evaluation checklist. This phase of the evaluation identified usability issues with the system at the outset and set the course for artifact improvements to be made in advance of student testing.

Phase 2: Student usability and trust study with 23 cybersecurity students. The participants of the study interacted with the HC-XAI dashboard using all three modalities (rule-based logic, natural language explanations, and heatmaps). They provided both quantitative ratings using Likert scales and qualitative feedback using open-ended questions.

This method provided a combination of both quantitative and qualitative analysis, allowing both the measurement of factors of interest as well as the ability to gain deeper insights about student users and the system as a whole. Examples of the rule-based, LLM, and heatmap explanation modalities used by the students during the study are shown in Figures 3–5 in the Appendix.

*Table 6 - Evaluation Phases and Objectives*

| Phase | Participants | Method | Focus | Outcomes |
|---|---|---|---|---|
| **Phase 1: Expert Walkthrough** | 5 domain experts | Heuristic evaluation with HC-XAI checklist | Usability, clarity, accessibility | Identified terminology, color, and onboarding issues |

| | | | | |
|---|---|---|---|---|
| **Phase 2: Student Study** | 23 cybersecurity students | Mixed-methods: Likert survey + open-ended responses | Trust, comprehension, cognitive effort | Quantitative ratings and qualitative insights on each modality |

**Phase 1: Expert Heuristic Walkthrough**

The initial iteration of artifact evaluation was an expert heuristic walkthrough of the HC-XAI dashboard. Five cybersecurity and information systems domain experts interacted with the artifact, then completed a usability evaluation of the artifact using a custom HC-XAI checklist. The purpose of this phase was to identify usability concerns, assess explanation quality, and uncover potential barriers before beginning the study with students.

Experts interacted with each explanation modality (rule-based logic, natural language rationale, and heatmaps) using sample phishing emails (Figures 3–5 in the Appendix illustrate sample dashboard outputs). Insights from the Expert Walkthrough:

- Rule-Based Logic: The technicality of the terminology (e.g., “malicious URL”) was highlighted as a concern, as students may not easily understand it; experts suggested substituting more jargon-free terms (e.g., “suspicious link”).
- Heatmaps: Experts indicated that the color gradients were not significantly distinct enough to understand the maps, especially for users with color vision deficiencies. They recommended higher contrast palettes and explanatory legends.
- LLM Explanations: While the clarity of LLM-generated explanations was generally high, experts recommended more concise prompts to ensure consistent outputs.
- Onboarding: Experts identified that inexperienced users would benefit from a tutorial or instructions to familiarize themselves with dashboard navigation.

These insights directly informed changes to terminology, visualization design, and onboarding support before Phase 2.

*Table 7 - Expert Heuristic Walkthrough Findings*

| Explanation Modality | Issue Identified | Recommended Refinement |
|---|---|---|
| **Rule-Based Logic** | Terminology too technical | Simplify to plain language (e.g., “suspicious link”) |
| **Heatmaps** | Color contrast and density issues | Adjust gradient palette; add explanatory legend |
| **LLM Explanations** | Inconsistent phrasing across outputs | Shorten prompts; enforce response length limit |
| **Dashboard Navigation** | Lack of onboarding support | Add a tutorial or a quick-start guide for new users |

## Phase 2: Student Usability and Trust Study

Phase 2 of the evaluation was a student study with 23 participants, all of whom were currently enrolled in cybersecurity courses. This phase of the evaluation helped us understand how students might use and experience HC-XAI in an educational setting, with a focus on the interface’s usability, the user’s comprehension of the system’s explanations, and their trust in the system.

Study Procedure: Students interacted with the dashboard by examining example phishing emails and analyzing output from each modality (rule-based logic, natural language rationale, and heatmaps; see Figures 3–5 in the Appendix for examples of each modality). After using the dashboard, participants completed a Likert-scale survey, including questions that assessed four constructs: trust, confidence, cognitive effort, and confusion. Open-ended questions were also included to allow students to give qualitative feedback on their experience. Evaluation Focus: The study evaluation attempted to answer three questions:

1. Usability: Did students find the dashboard easy to use and navigate?
2. Explanation Clarity: Did students understand why the email was classified as phishing based on output from each modality?
3. Trust Calibration: Did the explanations change students' confidence in the results output?

Results

1. Students generally found the dashboard easy to use, and moving between explanation modalities was relatively easy for all participants.
2. Students found the natural language explanations to be the most clear, followed by rule-based output.
3. Heatmaps were well-liked in terms of visual aesthetics, but many described them as "dense" or "hard to interpret."
4. Most students across all modalities indicated that they were more confident in the result if they had access to more than one form of explanation.

These initial results were used to form more specific quantitative and qualitative analyses detailed in Sections 5.4 and 5.5.

**Quantitative Results**

Student perceptions of the dashboard were measured with four constructs (trust, confidence, cognitive effort, and confusion) in which students were asked to rate their agreement on a 5-point Likert scale. Table 8 displays descriptive statistics, and Table 9 shows inferential statistics for these measures.

Descriptive Results: In general, students reported high confidence and clarity in the natural language explanations, moderate trust in the system output, and low levels of confusion. The Heatmaps modalities were associated with lower clarity and higher cognitive effort in comparison to rule-based and natural language modalities. Table 8 shows the mean and standard deviation values for each of the four constructs across the modalities.

*Table 8 - Descriptive Statistics of Student Perceptions*

| Survey Item | Mean | Std. Dev. | Interpretation |
|---|---|---|---|
| The dashboard was easy to use | 3.76 | 1.64 | Generally usable, though the user experience varied |
| It was clear how to select different explanation modes | 3.81 | 1.66 | Interface navigation was clear for most users. |
| The interface allowed quick application of explanations | 4.69 | 0.48 | Users found the interaction fast and efficient. |
| Rule-based explanations were helpful and clear | 3.24 | 1.55 | Mixed feedback on the clarity of rule-based output |
| The logic behind rule-based detection was clear | 2.90 | 1.58 | Several users had difficulty understanding the logic. |
| The AI-generated explanation was easy to follow | 4.38 | 0.81 | Strong clarity in natural language explanations |
| Explanation matched user expectations | 4.38 | 1.02 | Explanations aligned well with user reasoning |
| Heatmap highlighted key terms clearly | 3.81 | 1.11 | Visual highlights were mostly effective. |
| Visual display aided understanding | 3.75 | 1.18 | Display supported understanding for most users |

| | | | |
|---|---|---|---|
| I trust the system's classification | 3.19 | 1.47 | Trust levels varied, with a moderate overall response. |
| Explanations increased confidence in the system | 4.38 | 0.62 | Most users felt explanations improved confidence. |
| Explanations required significant mental effort | 2.00 | 0.89 | Explanations were easy to process cognitively. |
| Dashboard explanations were confusing | 1.69 | 0.60 | Very few users found the explanations confusing. |

Inferential Results: A Friedman test and Wilcoxon signed-rank post-hoc comparisons were conducted to assess for significant differences across the explanation modalities. There were statistically significant differences in participants' scores for all constructs (see Table 9). This was especially true between confidence and effort, as well as between trust and confusion. Table 9 shows where significant differences were located among constructs. Key takeaways include:

- Natural language explanations were the most favored in terms of clarity and confidence.
- Rule-based outputs received moderate ratings, though students were still tripped up by terminology.
- Heatmaps were significantly higher in effort than other modalities, as was noted by experts in Phase 1.

*Table 9 - Inferential Statistics: Wilcoxon Signed-Rank Tests*

| **Comparison** | **p-value** | **Significant (α = 0.0083)** |
|---|---|---|
| Trust vs. Confidence | 0.0047 | Yes |
| Trust vs. Effort | 0.0019 | Yes |
| Trust vs. Confusion | 0.0002 | Yes |

| | | |
|---|---|---|
| Confidence vs. Effort | 0.0009 | Yes |
| Confidence vs. Confusion | 0.0005 | Yes |
| Effort vs. Confusion | 0.0956 | No |

Note: Friedman test indicated significant differences across constructs, $\chi^2(3) = 35.69$, $p < .001$, Kendall's $W = 0.74$.

Overall, the quantitative results show that multimodal explanations may lead to higher confidence levels, while the type of explanation affects the trust and interpretability of a model.

## Qualitative Insights

In addition to the ratings, the students provided open-ended feedback on their experience with the HC-XAI dashboard. We qualitatively analyzed the free-text comments to gain a deeper understanding of the students' rationale behind their ratings, as well as to identify key design aspects and considerations for classroom application.

**Clarity of Explanations:** Many students appreciated the multimodal design but found that the rule-based module's language was sometimes too technical. One student wrote, "The rules were fine, but some language could be clearer for non-experts."

**Visual Understanding:** Heatmaps were reported as engaging but occasionally overwhelming. While some students liked the visual emphasis (e.g., "I liked the red words, but not all of them seemed relevant"), others found the displays visually dense and required additional scaffolding.

**Trust Enhancement:** A few students indicated that the multiple types of explanations increased their trust in the system. For example, "Having multiple types of explanations made me feel more confident."

**Suggestions for Improvement:** Participants requested additional onboarding support to interpret outputs, such as examples or short tutorials: "A short guide would make it easier to adopt."

**Professional Relevance:** A few students also mentioned the potential to use the model professionally, but needed stronger explanations: "Yes, I would use this, but I would need assurance about accuracy and explanation reliability."

These qualitative insights were largely consistent with the quantitative results: the natural language explanations were consistently clear, the rule-based outputs were potentially useful with less jargon, and the heatmaps were engaging. However, they required additional scaffolding to reduce cognitive load.

**Evaluation Summary**

In this work, we developed an explainable XAI-HC artifact, focusing on phishing detection, in the form of an XAI dashboard. We performed two separate testing phases for our artifact to demonstrate its usability as well as its instructional benefits. Our expert walkthrough further verified the technical soundness of our artifact and enabled us to gather initial feedback for refinement. Feedback from the expert walkthrough indicated that the explanation modality can be further improved by using simpler terms. Further, it also showed that the heatmap is not as accessible to all audiences and can be further improved by including helpful onboarding support.

The student testing of our artifact involved using an independent measures design with three distinct explanation modalities. We conducted both quantitative and qualitative analyses to gain insights into students' preferences and experiences with the explanation modalities. Results of the quantitative analyses have shown significant differences in students' responses on multiple items based on the explanation modality received. The qualitative data have also provided some clues as to why certain modalities are rated differently by the participants. The data from this usability testing clearly shows the benefits of using multimodal explanations. However, we do need to scaffold users further to assist in interpreting the explanation modalities to ensure that they are all usable by all audiences.

In terms of the educational implications, the testing data have clearly shown that explainable AI dashboards can be used as instructional tools in cybersecurity courses. Not only did we provide the students with the opportunity to learn about how the model makes phishing detection decisions, but we also asked them to think about trust in AI systems by comparing the explanation types.

# 5. Discussion

The findings summarized in the preceding section now serve as a basis for discussing what the HC-XAI dashboard brings to the table in terms of cybersecurity education, on the one hand, and human-centered, explainable AI research, on the other. In the following, we interpret the key insights gained from the evaluation, discuss their educational implications and potentials for teaching and learning, elaborate on what the artifact brings to the design knowledge body in the area of HC-XAI, and identify limitations and opportunities for future work.

**Interpretation of Findings**

The evaluation demonstrated that the HC-XAI dashboard is a usable and effective learning tool. However, the three explanation modalities yielded mixed results on the measures of clarity, cognitive effort, and trust impact.

Natural language explanations received the highest overall rating. Quantitatively, students rated them highest in terms of clarity and matched expectations. Qualitatively, they noted that the succinct narrative style approximated the way that an instructor might verbally communicate phishing heuristics, which made the output more palatable and less “threatening.” Taken together, these results suggest that plain-language AI explanations can promote both comprehension and confidence for learning.

Rule-based logic explanations were rated moderately well. Students appreciated the transparency that they offered in terms of explicitly identifying suspicious features. However, students also used terms like “bloated” or “code-like” to describe the output. To some extent, this issue could be mitigated by substituting more accessible language (and the evaluation suggests that rule-based explanations are best used as a complementary rather than primary explanation).

Heatmap visualizations were the least intuitive according to the evaluation. Students described them as visually “pleasing” but rated them as the most cognitively demanding. Students found them helpful as a secondary source of information, but they were largely unusable on their own. These results dovetail with the prior argument that visual explanations are only as useful as their explanatory scaffolding (legends, tool tips, instructor guidance, etc. ).

Regardless of the modality, the qualitative data suggest that students felt that multimodal explanations enhanced trust and confidence in their judgments by providing an opportunity for cross-validation of system outputs. This result aligns with existing HC-XAI work on the benefits of layered explanations in enabling trust calibration. However, the work also clearly shows that novice learning is served by supporting multiple explanation formats, even if they ultimately prefer one modality.

**Educational Implications**

The HC-XAI dashboard demonstrated strong value when it comes to helping students learn information systems and cybersecurity concepts. It can serve as an effective tool for both teaching and learning, as outlined below:

1. The dashboard is multimodal, catering to different learning preferences. The rule-based form is a recap of the textbook heuristics, while the natural language form is a chatbot analog of in-person instructor reasoning. Meanwhile, the visual form is more for the "Show, do not tell me!" crowd. Supporting multiple modalities means that the same phishing example can be explained from different angles, which may help reinforce understanding for a wider audience.

2. The dashboard can be used to promote active learning and critical engagement. It presents a phishing artifact and an associated AI classification. Students should be able to pause and take the time to evaluate and compare explanations to improve not only their ability to identify phishing, but also their understanding of AI's ability and limitations in explaining complex classification.

3. The dashboard may be of practical use to instructors. For example, it can be used for instructor demonstrations, lab exercises, or programming assignments to show the in-the-wild operation of XAI in a practical information security context. In a flipped classroom scenario, the system output can be used as an entry point into a class discussion. For example, students can be encouraged to "play devil's advocate" and argue the "shortcomings" of a given explanation. On the other hand, a more structured exercise can involve students in an attempt to find gaps in explanations and potentially fill them with human reasoning, for example, by being asked to reason as if they were the model. In both cases, students can be guided to map specific system reasoning to more general cybersecurity concepts.

4. The dashboard prepares students for their working future, where AI-powered tools will be an integral part of their professional activity. Exposure to and experience in working with explainable AI may allow students to calibrate trust in AI-powered decision support systems, which is important for any future professional cybersecurity analyst.

**Design Implications for HC-XAI**

In addition to the immediate learning insights, the evaluation also offers design implications for human-centered explainable AI (HC-XAI) systems more generally. Key takeaways include:

1. Value of multimodality. The results validate the notion that multimodal explanations provide confidence-boosting cross-validation. Even in the case where a modality (heatmaps) was consistently more difficult to interpret, its availability in the presence of rule-based and natural language outputs boosted overall confidence. This aligns with the HC-XAI principle that layered, multimodal explanations better calibrate trust.

2. Importance of simplicity and accessibility. The ambiguous performance of the rule-based outputs is an important reminder that technical jargon is a clear barrier to interpretability, particularly for novice users. HC-XAI system designers should use plain language whenever possible, and consider the user's domain knowledge level when crafting explanations (especially in learning or training scenarios).

3. Limitations of visual-only explanations. The heatmap results clearly show that visualizations on their own may not always be sufficient or clear, especially to non-expert users. Without scaffolding (legends, tooltips, instructor guidance), saliency maps may actually increase cognitive load. Designers should consider the need to support visual explanations with textual or narrative explanations to help prevent misinterpretation.

4. Onboarding and user guidance. The frequent requests from both experts and students for tutorials or quick-start guides suggest that HC-XAI systems should have integrated onboarding and user guidance features to assist users in understanding what the different outputs mean and how to interpret them, especially when using less familiar modalities (e.g., token-level heatmaps).

Overall, these design insights contribute to the HC-XAI literature by showcasing how multimodal explanation systems can be tailored to novice learners and adapted to a specific learning environment. They also offer concrete heuristics for future design efforts in this and similar contexts where transparency and accessibility are of utmost importance.

**Novelty/Contribution Statement**

To the best of our knowledge, we are not aware of any other validated HC-XAI dashboard for cybersecurity education. Prior research has proposed explainable AI dashboards for professional analysis tasks in operational environments. However, this study uniquely focuses on a novice learner use case by adapting multimodal explanations (rule-based logic, natural language rationales, and visual heatmaps) to the classroom context. This work bridges the gap between the XAI research community's focus on high-stakes industry applications and the development of XAI tools for applied education by grounding the artifact in DSR methodology and empirically validating it through expert walkthroughs and a student usability study. The novelty is in both the artifact and in showing that XAI can have instructional value, not just technical transparency.

**Usage Recommendations for Educators**

Educators may adopt HC-XAI to support classroom instruction of information systems and cybersecurity curricula in various ways. The tool may be introduced in a lecture setting as a demonstration artifact to illustrate phishing heuristics and AI reasoning. Instructors can show rule-based, natural language, and visual modes of explanation in the HC-XAI dashboard, revealing how the model makes a phishing prediction for a given email. For example, the instructor can then use this visualization to initiate a discussion on the relative interpretability and trustworthiness of each mode of explanation. Alternatively, the tool can be used for experiential learning in lab exercises by having students explore the model outputs for phishing emails they enter, supporting the calibration of trusting beliefs in AI systems through the comparative evaluation of outputs. Educators can also design flipped classroom activities by assigning students to explore phishing examples with the HC-XAI dashboard before class. The generated outputs then form the basis of a group discussion, case analysis, or "AI auditor" role-play activity in class. The tool can also be used to evaluate student understanding with formative assessments. For instance, an instructor may challenge a student to justify their classification decision in a sample phishing exercise using more than one modality of explanation. As students are not cybersecurity experts, instructors should also provide simple onboarding material for novice users to help them parse the tool's rule-based and heatmap outputs. This onboarding content may take the form of a brief tutorial video or a quick-start user guide for rule-based and heatmap outputs to help decrease extraneous cognitive load. These and other suggested uses will allow the HC-XAI dashboard to function not just as a demonstration artifact, but as an integrated pedagogical tool that can deepen student understanding of phishing while also building a critical explanation-awareness of explainable AI.

**Limitations and Future Work**

The evaluation strongly supports the potential educational value of the HC-XAI dashboard for cybersecurity training. However, several limitations of the current work should be noted.

Sample size and participant profile. The sample size of 23 students is relatively small and should be considered when interpreting the results. The limited size of the study may not provide sufficient statistical power to detect small effects and limits generalizability to other populations. Furthermore, the participants were student learners rather than professional security analysts. Therefore, the findings of the evaluation should be interpreted as supporting evidence for the value of the system in learning contexts rather than directly generalizable to industry or security operations center settings.

Scope of explanation modalities. The system was limited to three static explanation modalities, as shown in Table 1. While all modalities were found to be of value for students, the current implementation does not include adaptive or dynamic personalization of explanations to individual users or use cases. It could be interesting for future work to explore dynamic adaptation to different users or contexts. For example, what if the explanation types were customized for a user's individual cognitive preferences or cognitive styles, levels of expertise, or current learning objectives?

Evaluation methodology. The current work used subjective self-report ratings of trust, confidence, effort, and confusion. These metrics can provide insight into user experience, learning, and cognitive load, but do not measure performance or effectiveness directly. Objective performance-based evaluation criteria, such as error detection, decision-making speed, or accuracy improvements, would provide more robust evidence of the utility of different types of XAI in cybersecurity contexts.

Technical constraints. The current system uses a remote cloud-hosted API for system operations, as shown in Figure 4. This setup may not be feasible or appropriate in all contexts, particularly in secure or locked-down environments. Future work could explore the potential for local deployment, edge computing, or privacy-preserving AI models.

Sequencing of explanations. The system allowed free navigation between the three static explanation modalities in the HC-XAI dashboard. The evaluation did not log the specific sequences of modality switches or control the order of presentation. It would be interesting to see whether modality sequencing has an impact on the overall explanation quality. For example, does the order in which students are presented with the explanation modalities impact their interpretation? For instance, do they trust rules more if they are shown a heatmap that matches them first?

Open-source availability. We have made the entire source code for the HC-XAI dashboard available at https://github.com/sasajs/hc-xai-dashboard, which is hosted on a GitHub repository. This directly addresses one of the study's limitations: the single-site limitation. With the release of the code, other researchers and practitioners can easily reproduce the artifact, verify its capabilities, and use it in educational settings. The open-source nature of the dashboard will encourage further extensions, such as adaptive explanation strategies, novel visualizations, and global integration into cybersecurity training programs. Future research will extend this work by testing the dashboard with industry professionals and security analysts to validate its applicability beyond student populations.

### Directions for Further Work

Beyond the caveats described, several concrete avenues for further work were enabled by this study. Foremost among these is the validation of the HC-XAI dashboard in the workplace among practicing cybersecurity analysts, and especially with industry threat response teams. In this study, we showed educational impact using the artifact with students, but expanded deployment can shed light on whether the artifact also supports trust calibration and decision-making in operational settings where accountability and timeliness are of the essence. Future validation efforts should also scale up classroom use of the dashboard, evaluating the tool in other institutions with larger cohorts and in diverse curricular contexts, to better understand the generalizability and durability of learning gains. To this end, all source code for the dashboard is publicly available, allowing researchers, educators, and other practitioners to reproduce the system, adapt it to new use cases, and extend its design. This open-source approach enables a collaborative path forward for both industry and academic partners to refine, extend, and validate the artifact in complementary ways that have the potential to improve its impact and sustainability.

## 6. Appendices

### Appendix A. Design Theory Table

Table 5 outlines the expert heuristic criteria used for the artifact evaluation.

*Table 10: Design Theory Table (Gregor & Jones, 2007)*

| **Component** | **Description** |
|---|---|
| Purpose and Scope | The XAI interface design needs to enable users to understand and trust AI-generated cybersecurity threat alerts. |

| | |
|---|---|
| Constructs | The explanation modes — rule-based, LLM, and heatmap — affect users' perceived trust and clarity, while influencing their cognitive effort. |
| Principles of Form and Function | The artifact should enable users to switch between different explanation types, ensuring that explanations align with their cognitive models and decision requirements. |
| Artifact Mutability | The system can be adjusted to support new methods of explanations and emerging threat domains. |
| Testable Propositions | Multiple explanation modalities enhance users' understanding and trust levels, while non-expert individuals tend to favor explanations provided by large language models (LLMs). |
| Justificatory Knowledge | Research by Gregor and Benbasat (1999), together with Gunning and Aha (2019) and Wang et al. (2020), informs the development of cognitive psychology theories. (2020), cognitive psychology theories. |
| Principles of Implementation | Developed through Python and Streamlit technology, with separate modules designed for each explanation method. |
| Expository Instantiation | Developed through Python and Streamlit technology, with separate modules designed for each explanation method. |

## Appendix B. Knowledge Contribution Matrix

Table 6 reports the participant demographics from the user study.

*Table 11: Knowledge Contribution Matrix (Gregor & Hevner, 2013)*

| Problem Domain Maturity | Solution Domain Maturity | Contribution Type | Description |
|---|---|---|---|
| Established: Research has extensively demonstrated the requirement for AI explainability and | Nascent: The field of cybersecurity lacks sufficient research and development into | Invention | The artifact introduces a unique system that merges three different explanation methods into one adaptive interface to address a |

| | | | |
|---|---|---|---|
| trustworthiness in cybersecurity applications. | multimodal XAI dashboards. | | missing component in HC-XAI for cybersecurity. |

**Appendix C. Survey Instrument (Post-IRB)**

After completing their session with the HC-XAI dashboard, participants took the completed post-intervention survey. The University of Wisconsin–Oshkosh IRB authorized the survey tool that contains Likert-scale questions and open-ended queries to assess usability quality alongside explanation clarity, trustworthiness, cognitive workload, and professional utility.

**Post-Intervention Survey**

- Consent and IRB compliance section
- Demographics: Age, gender, experience with cybersecurity
- Dashboard Usability
- Explanation Evaluation: Rule-Based, AI-Generated, Heatmap
- Trust and Cognitive Effort
- Open-ended survey items ask participants to evaluate the dashboard in terms of clarity and trustworthiness while assessing its professional utility.

A complete item list can be reviewed in the supplemental materials and through a request process. Table 7 displays the usability and trust metrics collected during the evaluation phase.

*Table 12: Post Test Survey*

| Construct | Survey Item |
|---|---|
| **Dashboard Usability** | The dashboard was easy to use. |
| | It was clear how to select different explanation modes. |
| | I felt comfortable navigating the interface. |

| | |
|---|---|
| **Rule-Based Explanation** | The rule-based explanation helped me understand the classification. |
| | The explanation was clear and concise. |
| | I would feel confident relying on the rule-based explanation. |
| **AI-Generated Explanation** | The AI-generated explanation helped me understand the classification. |
| | The explanation was clear and concise. |
| | I would feel confident relying on the AI-generated explanation. |
| **Heatmap Explanation** | The heatmap helped me understand the classification. |
| | The explanation was clear and concise. |
| | I would feel confident relying on the heatmap explanation. |
| **Trust & Effort** | The explanations increased my trust in the classification. |
| | I had to think hard to interpret the explanations. |

*Note: Following the Likert-based items, participants were presented with open-ended questions to provide qualitative feedback on the dashboard's clarity, usefulness, and potential areas for improvement.*

Instructions for Use: Experts should complete this checklist after interacting with the artifact. Experts can evaluate items by providing scores and comments, which will assist in progressive refinement.

**Appendix D. Suspicious Email (generated by ChatGPT)**

**Subject:** Urgent: Account Suspension Notification

**Body:**

Dear User,

We have detected unusual activity in your account that violates our terms of service. As a result, your access has been temporarily suspended.

To restore access, please verify your credentials immediately by clicking the secure link below:

http://account-verification-secure.com/login

Failure to act within 24 hours will result in permanent suspension of your account.

Thank you for your prompt attention to this matter.

Sincerely,

Account Security Team

**Appendix E. Rule-based Model Response to Suspicious Email**

The rule-based response identifies key indicators of a phishing email using predefined logical rules, as demonstrated in **Fig. 3**.

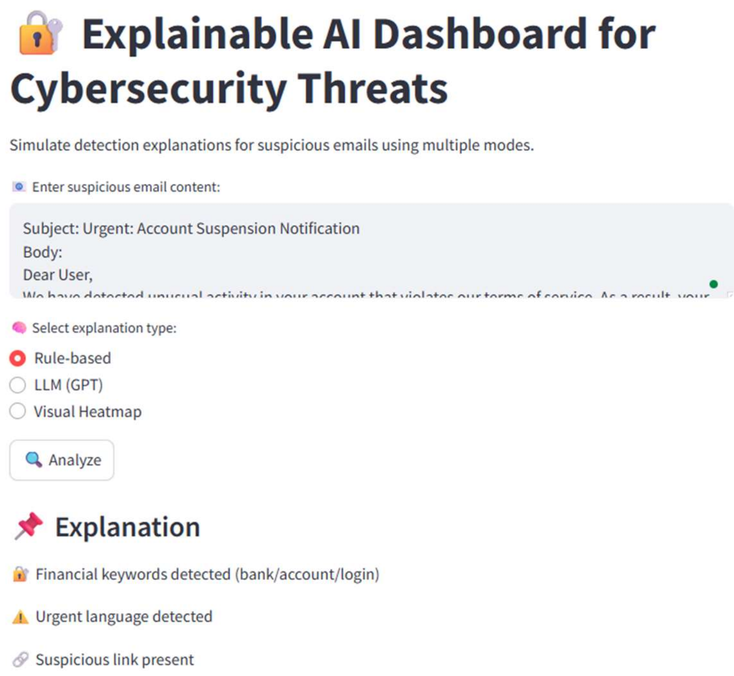


*Figure 3: Rule-based Response to Phishing Email*

In **Fig. 3**, the rule-based response is shown, where the system identifies key indicators of a phishing email using predefined logical rules.

### Appendix F. LLM (GPT) Model Response to Suspicious Email

The LLM (GPT) generates a natural language explanation for why the email is classified as phishing, providing a more intuitive, human-like justification, as shown in **Fig. 4**.

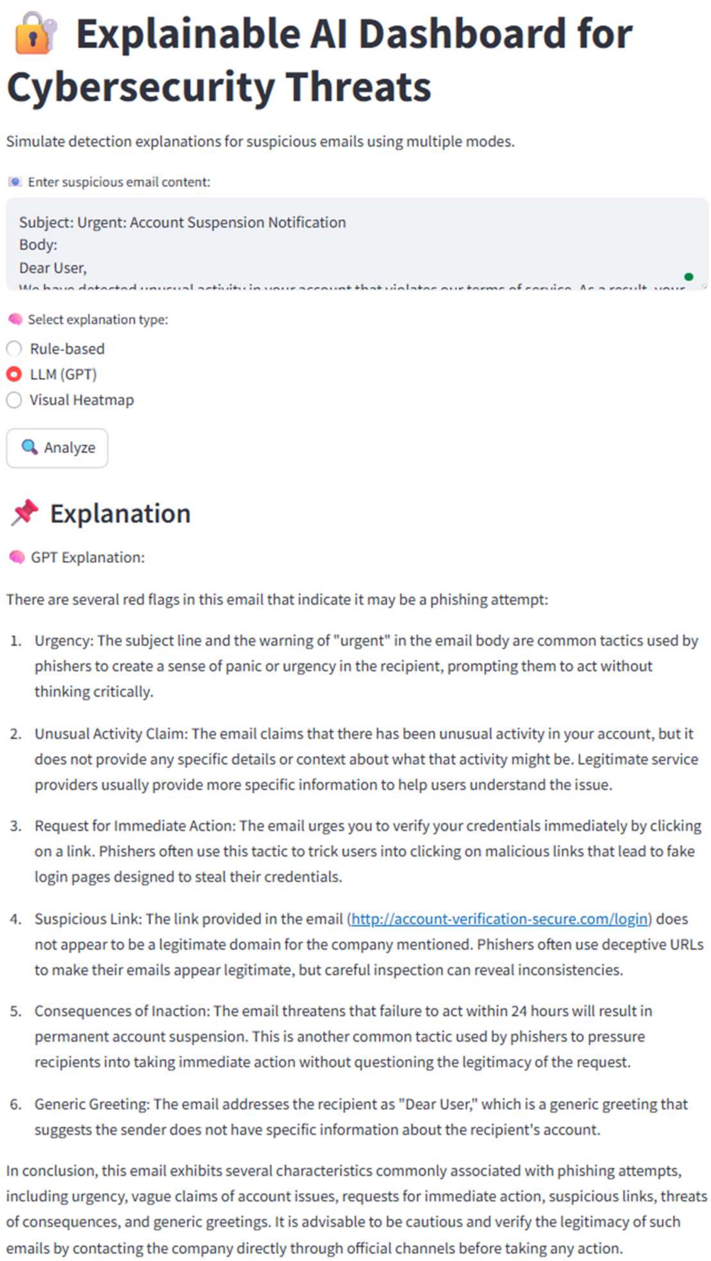


*Figure 4: LLM (GPT) Generated Explanation for Phishing Email*

As depicted in **Fig. 4**, the LLM (GPT) generates a natural language explanation for why the email is classified as phishing, providing a more intuitive, human-like justification.

## Appendix G. Visual Heatmap Model Response to Suspicious Email

The visual heatmap generated by the system highlights suspicious language patterns identified in the phishing email, aiding users in understanding the AI's attention areas, as depicted in **Fig. 5**.

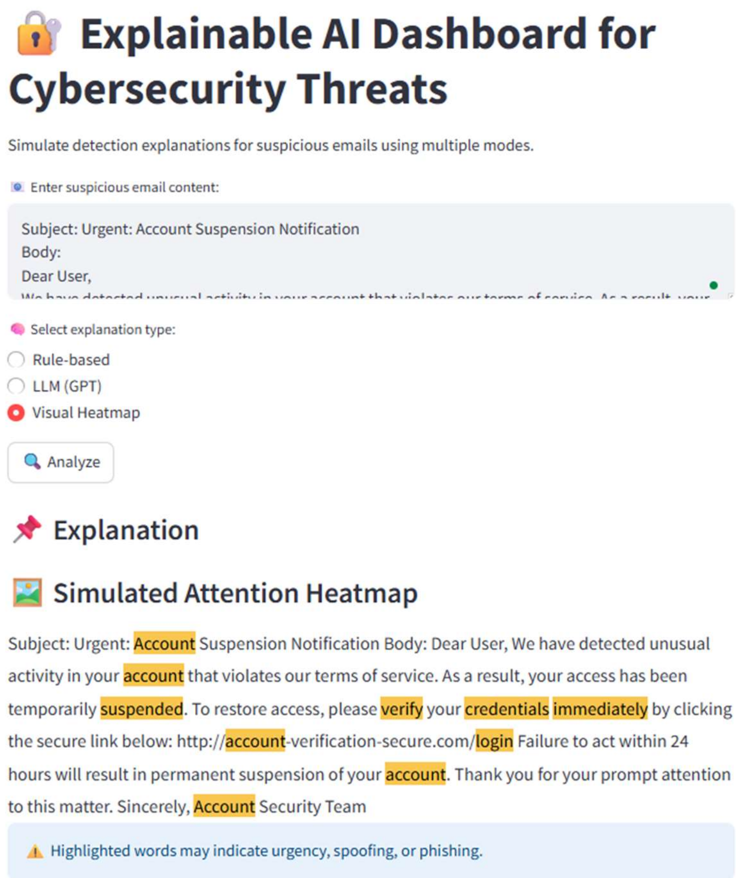


*Figure 5: Visual Heatmap Model Response to Phishing Email*

In **Fig. 5**, the visual heatmap generated by the system highlights the suspicious language patterns identified in the phishing email, helping users understand the AI's attention areas.

## 7. Declarations

**Ethics approval and consent to participate**

This study was reviewed and approved by the Institutional Review Board (IRB). All participants provided informed consent prior to participation.

**Consent for publication**

Yes, not applicable.

**Availability of data and material**

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

**Competing interests**

The author declares no competing interests.

**Funding**

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**Authors' contributions**

Dr. Steven A. Schilhabel was solely responsible for the conceptualization, design, implementation, data analysis, and manuscript preparation.

**Acknowledgments**

The author thanks the participating cybersecurity students for their contributions, as well as the expert evaluators who assisted with the heuristic walkthrough.

## 8. References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6,* 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alqaraawi, A., Schuessler, M., Weiß, P., & Kulesza, T. (2020). Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM. https://doi.org/10.1145/3313831.3376218

Arora, A., & Rahman, Z. (2017). Information Dashboards for IS Education: A Teaching Note. *Journal of Information Systems Education, 28*(3), 193–202.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics, 8*(8), 832. https://doi.org/10.3390/electronics8080832

Davis, J. G., Dehlinger, J., & Hilburn, T. (2020). Using Dashboards to Teach Software and Systems Concepts. *Information Systems Education Journal, 18*(5), 20–29.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. *arXiv*. https://arxiv.org/abs/1702.08608

Ehsan, U., & Riedl, M. O. (2020). Human-Centered Explainable AI: Toward a Reflective Sociotechnical Approach. *arXiv*. https://doi.org/10.48550/arXiv.2001.00093

Ehsan, U., Rai, A., & Riedl, M. O. (2021). Explainability in Human–AI Interaction: Toward a Reflective Sociotechnical Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). ACM. https://doi.org/10.1145/3411764.3445188

Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine, 40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Holstein, K., Wortman Vaughan, J., Halpern, M., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. https://doi.org/10.1145/3290605.3300830

Kraemer, S., Carayon, P., & Clem, J. (2009). Human and Organizational Factors in Cybersecurity: A Literature Review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 53*(4), 197–201. https://doi.org/10.1177/154193120905300420

Liao, Q. V., & Varshney, K. R. (2022). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *ACM Transactions on Interactive Intelligent Systems, 12*(2), 1–45. https://doi.org/10.1145/3495244

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes, 151,* 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Molnar, C. (2019). *Interpretable Machine Learning.* Independently published. https://christophm.github.io/interpretable-ml-book

Mohale, O., & Obagbuwa, I. (2025). A Systematic Review of Explainable AI in Cybersecurity. *Journal of Cybersecurity Research, 12*(1), 1–20. https://doi.org/10.1080/23742917.2025.000000

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). Multidisciplinary Perspectives on Human-Centered AI: A Survey. *Frontiers in Artificial Intelligence, 4,* 26. https://doi.org/10.3389/frai.2021.608463

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International*

*Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Speith, T. (2022). A Review of Explainable Artificial Intelligence in Practice: Barriers and Opportunities. *Philosophy & Technology, 35*(4), 1–29. https://doi.org/10.1007/s13347-022-00541-7

Verma, R., Das, A., Hasan, S., & Swetapadma, A. (2017). Detecting Phishing Emails Using Natural Language Processing and Machine Learning. In *2017 IEEE 16th International Conference on Machine Learning and Applications (ICMLA)* (pp. 1106–1113). IEEE. https://doi.org/10.1109/ICMLA.2017.00021

Whitman, M. E. (2019). Design Science Research in IS Education: Reflections and Opportunities. *Journal of Information Systems Applied Research, 12*(1), 4–12.