



International Journal of Science, Architecture, Technology and Environment

SuriSafe PH: An Explainable Scam and Phishing Assistance Platform for Digital Safety Literacy Among Filipino Users

Kheen Lester Barrozo^{1*}, Grace Marjorie B. Berber², Eliza B. Ayo³

^{1,2,3}*School of Science and Technology, Centro Escolar University*

^{*}*Corresponding author*

DOI: <https://doi.org/10.63680/ijstate102618.41>

Abstract

Filipino digital users regularly receive suspicious short message service (SMS) messages, chat messages, emails, links, screenshots, and quick response (QR) codes that impersonate banks, electronic wallets, couriers, employers, and public agencies, yet common spam filters and link checkers rarely explain why content is risky or what to do next. This study designed and prototyped SuriSafe PH, an explainable decision-support and incident-management platform that combines content analysis, privacy protection, safe-action guidance, incident preparation, and micro-learning in a single Filipino-focused workflow. Following an iterative, human-centered design approach, a literature and market review identified 10 research gaps that informed a target persona, 13 functional requirements, and a set of design guidelines. The prototype was refined through four staged design and implemented as a responsive web application with a Supabase back end and server-side OpenAI analysis. Evaluation comprised an evidence-keyed comparison with four existing services, an implementation-status audit, and a functional check on six researcher-labeled examples. The prototype provides 20 user and administrator interfaces. None of the compared services documented the combination of privacy review, separation of automated and verified evidence, emergency recovery guidance, and incident preparation. Only the interface layer was demonstrated; back-end persistence, access control, and live registry integrations remain incomplete. Separating automated assessment from verified source status is the central design contribution, and integration testing, a larger evaluation, and a usability study are recommended before deployment, in support of Sustainable Development Goals (SDGs) 4, 8, and 16.

Keywords: Phishing, scam prevention, explainable security warnings, human-centered design, Philippines

Introduction

Regional law-enforcement reporting describes escalating cyber threats across Asia and the South Pacific (INTERPOL, 2026). In the Philippines, digital users regularly receive suspicious short message service (SMS) messages, chat messages, emails, links, screenshots, and quick response (QR) codes that imitate banks, electronic wallets, couriers, employers, and public agencies. These messages often rely on urgency, impersonation, shortened links, unfamiliar domains, or small payment requests to obtain personal information, account credentials, or money. Generic spam filters and link checkers may provide a warning

without explaining which parts of the content caused concern, leaving a non-technical user uncertain about the safest next action.

The National Privacy Commission advises users to examine unknown messages and shortened links carefully and to avoid disclosing sensitive information (National Privacy Commission, 2021). The Bangko Sentral ng Pilipinas (BSP) reported that financial scams affected Filipino adults and that many incidents were not reported because reporting was viewed as time-consuming or inconvenient (Bangko Sentral ng Pilipinas, 2022). Together, these findings point to the need for a system that explains warning signs, protects private information, prepares a reusable incident summary, and gives the user clear steps for verification and recovery.

This study aimed to design and prototype SuriSafe PH, an explainable scam and phishing assistance platform for Filipino digital users, and to assess the prototype against existing services and its own implementation evidence. SuriSafe PH is positioned as a decision-support and incident-management companion. It does not replace a bank, regulator, law-enforcement agency, cybercrime office, browser security service, or antivirus product; it does not declare that content is completely safe; and it does not present automated analysis as an official government or financial-institution decision. Instead, it helps the user organize evidence, understand risk indicators, and reach the correct official channel. The study is also aligned with the United Nations 2030 Agenda for Sustainable Development (United Nations, 2015). Its primary contribution is to SDG 4 (Quality Education), Target 4.4, which seeks to increase the number of youth and adults with relevant skills, including information and communication technology skills, by building users' ability to recognize and respond to digital scams. It also supports SDG 8 (Decent Work and Economic Growth), Target 8.10, which promotes access to financial services for all, because trust and safety are preconditions for sustained use of electronic wallets and online banking. Finally, it supports SDG 16 (Peace, Justice and Strong Institutions), Target 16.4, which calls for combating all forms of organized crime, by helping victims prepare reports and reach the appropriate authorities.

Specifically, the study sought to (a) identify the gaps left by existing research and services; (b) translate these gaps into requirements and a working prototype; (c) compare the prototype with existing services using documented evidence; and (d) audit the implementation status of the prototype and the reliability of its automated assessment.

Literature Review

Explainable Phishing Warnings

Research on phishing warnings shows that users benefit when a warning explains the evidence and the recommended response instead of presenting an unexplained block (Aneke et al., 2021; Buono et al., 2023; Desolda et al., 2023). Explainable SMS phishing research has also demonstrated the value of walkthrough-style explanations generated by large language model (LLM) agents, although performance reported in a single evaluation cannot be treated as guaranteed real-world accuracy (Wang et al., 2025).

QR Code Threats

QR code attacks require additional safeguards because a visible code can hide its destination. In experiments reported by Kowalewski et al. (2025), participants had difficulty identifying fraudulent QR payment requests, and research on in-app QR scanning describes risks that arise when decoded content enters unsafe application flows (Han et al., 2023).

Reference-Based Verification and Human-Centered Design

Reference-based phishing detection can improve explainability when claims about brands and domains are checked against external evidence (Liu et al., 2024), which suggests that model-generated interpretation should be distinguished from evidence obtained from an identified reference source. Human-centered design, in turn, requires explicit attention to users, tasks, context of use, evaluation, and iteration (International Organization for Standardization, 2019); for security tools, this includes plain language, accessible interaction, and recovery from failure states.

Existing Services

Existing products address parts of the problem. Google Messages provides spam protection inside a messaging product (Google, n.d.-a). Google Safe Browsing checks uniform resource locators (URLs) against lists of unsafe resources (Google, n.d.-b). VirusTotal provides specialist analysis of files, URLs, domains, and Internet Protocol (IP) addresses (VirusTotal, n.d.). BSP Verifier helps users determine whether a financial institution is regulated by the Bangko Sentral ng Pilipinas (Bangko Sentral ng Pilipinas, n.d.). These services have different purposes, and none provides a single Filipino-focused journey that combines privacy review, explainable evidence, safe actions, incident preparation, follow-up, and micro-learning.

Research Gaps

Table 1 summarizes the 10 research gaps identified from the literature and market review, the basis for each gap, and the corresponding design response.

Table 1: Research Gaps and Corresponding SuriSafe PH Design Responses

No.	Research gap	Basis	SuriSafe PH design response
1	Generic warnings may not explain the specific cues behind a decision.	(Aneke et al., 2021; Desolda et al., 2023; Wang et al., 2025)	Highlights suspicious phrases and explains urgency, impersonation, destination, and payment cues.
2	Scanning, reporting, learning, and recovery tools are often separated.	(Bangko Sentral ng Pilipinas, 2022, n.d.)	Combines inspection, safe actions, incident preparation, history, and safety lessons.
3	General services may not reflect Filipino scam scenarios and mixed-language use.	(Bangko Sentral ng Pilipinas, 2022, n.d.; National Privacy Commission, 2021)	Uses familiar e-wallet, parcel, bank, employment, and government-impersonation examples with English (EN) and Filipino (PH) controls.
4	QR destinations can remain hidden until the user scans the code.	(Han et al., 2023; Kowalewski et al., 2025)	Decodes first, displays the destination, and never opens it automatically.
5	Users may expose personal or financial data while asking for help.	(National Privacy Commission, 2021)	Adds a privacy review and redaction step for one-time passwords (OTPs), personal identification numbers (PINs), card verification values (CVVs), account numbers, and other sensitive data.

No.	Research gap	Basis	SuriSafe PH design response
6	Artificial intelligence (AI) interpretation can be confused with verified institutional evidence.	(Liu et al., 2024)	Separates automated assessment from verified source status and identifies unconfigured integrations.
7	Reporting may be avoided when users must reconstruct the incident repeatedly.	(Bangko Sentral ng Pilipinas, 2022)	Prepares a reusable redacted incident summary and evidence checklist.
8	Users need status and reply visibility after submitting a case.	Project requirement	Connects incident cases, administrator replies, notifications, and a status timeline to Scan History.
9	Precise percentages may imply scientific certainty without calibration.	(OpenAI, n.d.-a; Wang et al., 2025)	Uses Low, Medium, or High qualitative confidence that describes indicator consistency, not scam probability.
10	Security tools can be difficult to use on mobile devices or with assistive technology.	(International Organization for Standardization, 2019)	Provides responsive navigation, 44-pixel targets, text and icon status cues, keyboard focus, and 200% zoom support.

Method

Research Design

The study followed an iterative, human-centered design and development approach (International Organization for Standardization, 2019) in four stages: (a) problem analysis and literature and market review; (b) definition of the target user, functional requirements, and design guidelines; (c) iterative, prototyping and implementation; and (d) evaluation through an evidence-keyed feature comparison, an implementation-status audit, and a functional check of the automated assessment. No human-participant data were collected for the work reported here.

Functional Requirements

Thirteen functional requirements (FRs) were derived from the research gaps and the persona (Table 2).

Table 2: Functional Requirements of SuriSafe PH

ID	Requirement	Description
FR 01	Authentication	Register, sign in, recover password, optional Google sign-in, session handling, and role-based routing.
FR 02	Dashboard	Provide the primary scan action, scan modes, recent activity, and navigation.
FR 03	Content analysis	Accept message, screenshot, URL, and QR input; validate, redact, and analyze the content.
FR 04	Explainable result	Show High Risk, Needs Verification, or No Clear Threat Detected with qualitative confidence and highlighted evidence.
FR 05	Verify sender	Record the sender claim, method, result, source status, warning signs, and recommended action.
FR 06	Emergency help	Present prioritized steps based on whether the user opened a link, paid, installed an app, or disclosed information.
FR 07	Scan history	Store authorized scans, sender verifications, and incidents with filters, details, and permitted deletion.
FR 08	Safety center	Provide guides, official agencies, and fictional scam simulations.
FR 09	Incident management	Create a case number, store a redacted report, show status, support replies, and notify the user.
FR 10	Administrator operations	Provide incident queue, user management, verified sources, AI evaluation, audit logs, and system settings.
FR 11	Privacy	Prevent the collection or display of passwords, OTPs, PINs, CVVs, and unnecessary personal information.
FR 12	Accessibility	Support responsive layouts, keyboard operation, visible focus, text and icon states, large targets, themes, and bilingual controls.
FR 13	Error handling	Show loading, empty, offline, permission-denied, no-readable-text, application programming interface (API) failure, and database-failure states without a blank page.

Design Principles

Five principles guided the design. First, the interface presents the risk state, the supporting evidence, and the safest next action in a fixed sequence, because users need to understand why a warning appears before deciding how to respond (Buono et al., 2023; Desolda et al., 2023). Second, privacy review occurs before

analysis to reduce accidental disclosure (National Privacy Commission, 2021). Third, automated assessment is kept separate from verified source checks (Liu et al., 2024): an official source is shown only when its name, URL, and connection status are recorded, and unconnected registries are labeled STANDBY or Integration Not Configured. Fourth, QR scanning requests camera permission only after user intent, decodes once, shows the destination, and requires a separate analysis action before any navigation (Han et al., 2023; Kowalewski et al., 2025). Fifth, confidence is expressed qualitatively (Low, Medium, or High) rather than as a precise percentage because the underlying model is not calibrated (OpenAI, n.d.-a).

Accessibility and context of use were addressed through plain language; bilingual EN and PH controls; keyboard focus; text labels beside icons; large touch targets; responsive layouts; and error recovery for camera, network, optical character recognition (OCR), and API failures (International Organization for Standardization, 2019). Both desktop and mobile browsers are supported because suspicious content may be received on a phone but reviewed or reported from a computer. The visual design uses a near-black background, dark cards, an orange accent, sans-serif body text of at least 16 pixels, a fixed desktop sidebar that becomes a drawer on mobile, and one dominant action per screen; no status is communicated by color alone, and the label "Completely Safe" is never used. Administrator functions are grouped in a protected workspace, separate from ordinary user tasks.

Prototyping and Implementation

The interface was generated and refined with generative design tools through four staged natural-language designs, reproduced in the Appendix. The initial design specified the core screens, risk labels, plain language, privacy reminders, and bilingual controls. The first refinement added responsive navigation, a permission-gated QR scanner, and the verification, emergency, history, learning, reporting, and settings screens. The second refinement added authentication, role-based routing, and the administrator workspace, restricting the Quick Role Access control to isolated Evaluation Mode data. The technical refinement fixed the implementation stack, the server-side handling of the AI service, and the risk and confidence vocabulary.

The client was implemented with React, TypeScript, and Vite. Supabase Database, a managed PostgreSQL service, was selected for persistence, and Supabase authentication (Supabase, n.d.-a), private storage (Supabase, n.d.-d), and Edge Functions (Supabase, n.d.-b) form the target back end, to be relied upon only after they are connected and tested. Image and text understanding is performed by the OpenAI API (OpenAI, n.d.-b), called exclusively from a server-side function so that the API key never reaches browser code. PostgreSQL Row Level Security (RLS) (Supabase, n.d.-c) was not implemented; cross-user isolation and administrator authorization must therefore be enforced through server-side role and ownership checks.

Evaluation Procedures

Evidence-Keyed Feature Comparison

SuriSafe PH was compared with Google Messages (Google, n.d.-a), Google Safe Browsing (Google, n.d.-b), VirusTotal (VirusTotal, n.d.), and BSP Verifier (Bangko Sentral ng Pilipinas, n.d.) on 10 functions derived from the research gaps. Judgments for existing services were keyed to their official documentation, and judgments for SuriSafe PH were keyed to the prototype screen implementing the function. Six ratings were used: YES (the official source directly documents the feature), PARTIAL (only part of the workflow is present or an additional integration is required), LIMITED (protection is narrower than the complete SuriSafe PH workflow), TECHNICAL (output is aimed at specialists), ADVISORY (guidance rather than automated checking), and NO (not documented in the reviewed official scope, which is not a claim that the feature cannot exist).

Implementation-Status Audit

Each architectural component was classified according to the available evidence. A screenshot was treated as evidence that an interface exists, not that authentication, persistence, an external integration, or a button action executed successfully.

Functional Check of the Automated Assessment

The automated assessment was run on six researcher-labeled examples, including Filipino-language content, through the administrator AI Evaluation module, which compares each ground-truth label with the predicted label and tabulates a confusion matrix of threat and non-threat outcomes. This check was designed as a sanity check of the evaluation pipeline, not as an accuracy study.

Results and Discussion

System Architecture and Implementation Status

Figure 1 shows the technical architecture. The client communicates with Supabase authentication, database, and storage services and with an Edge Function that validates and sanitizes input before calling the OpenAI API server-side. Verified sources and safety content are stored separately from AI output.

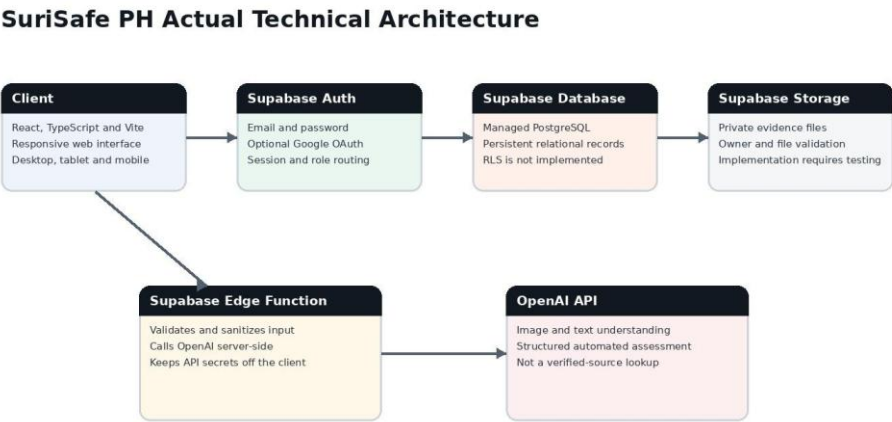


Figure 1: Technical Architecture of SuriSafe PH and Its Integration Boundaries

The implementation-status audit (Table 3) shows that only the interface layer was demonstrated; the back-end components were configured, selected, or not implemented, and no live connection to a Philippine registry was claimed.

Table 3: Implementation Status of the Main Architectural Components

Component	Status	Interpretation
React, TypeScript, and Vite interface	Demonstrated	The user and administrator screenshots show the responsive interface.
OpenAI secret and screenshot Edge Function	Configured; retest required	Successful end-to-end behavior requires a recorded test.

Component	Status	Interpretation
Supabase Database (managed PostgreSQL)	Selected; integration incomplete	The data model is documented; persistence and access-control behavior must still be verified.
PostgreSQL RLS	Not implemented	RLS is not claimed to be active.
External Philippine registries	Standby	No live CICC, BSP, PNP ACG, or DTI feed is claimed.

Note. CICC = Cybercrime Investigation and Coordinating Center; PNP ACG = Philippine National Police Anti-Cybercrime Group; DTI = Department of Trade and Industry; RLS = Row Level Security.

Data Model

The relational data model (Figure 2) supports registered users, scan history, evidence, sender verification, incident cases, administrator replies, notifications, and audit events. Verified sources and safety content are held in separate tables so that source status and educational material are never embedded inside an AI result.

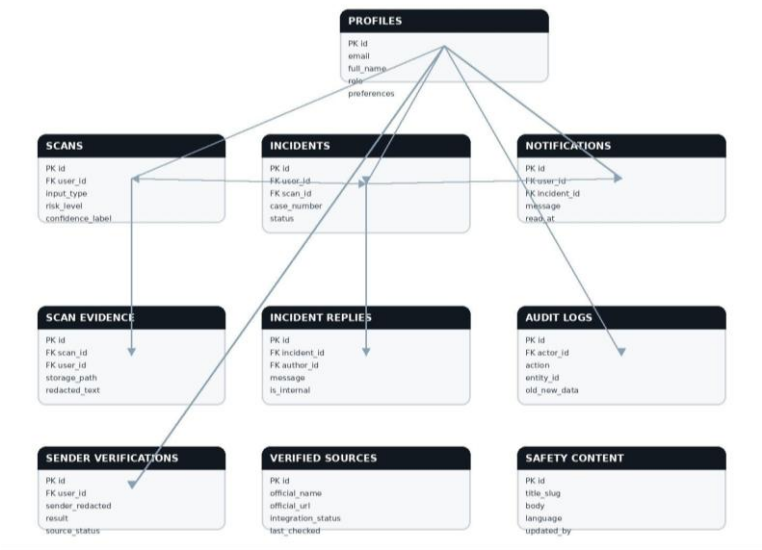


Figure 2: Entity-Relationship Diagram of the Proposed Supabase Database Schema

User and Administrator Workflow

- Figure 3 presents the combined workflow. The primary user journey consists of seven steps:
- Sign in with a registered account or enter the clearly labeled Evaluation Mode.
 - Open the dashboard and choose a message, screenshot, URL, or QR scan, or verify a sender.
 - Review and remove sensitive information before analysis.
 - Read the automated assessment, qualitative confidence, warning signs, and source status.

Choose a safe action, request emergency help, or prepare an incident summary.

View the saved scan, sender verification, or incident in Scan History once persistence succeeds.

Receive administrator replies, information requests, status changes, and resolution notifications

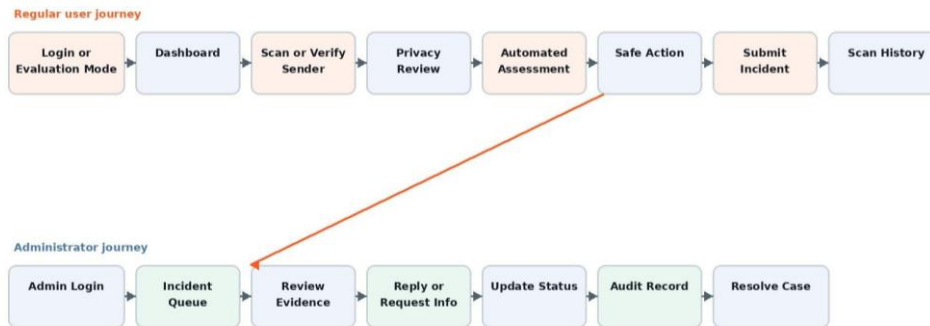


Figure 3: User and Administrator Workflow

The use case model (Figure 4) involves five actors. The regular user submits suspicious content, reviews evidence, chooses safe actions, manages personal history, and communicates through incident cases. The administrator, through a protected route, reviews redacted cases, replies, requests information, updates status, maintains approved content, and records sensitive actions. SuriSafe PH itself validates input, applies privacy rules, stores authorized records, and presents uncertainty and integration status. The OpenAI API provides server-side image and text understanding but no official verification, and a verified source is treated as live only when retrieval succeeds.

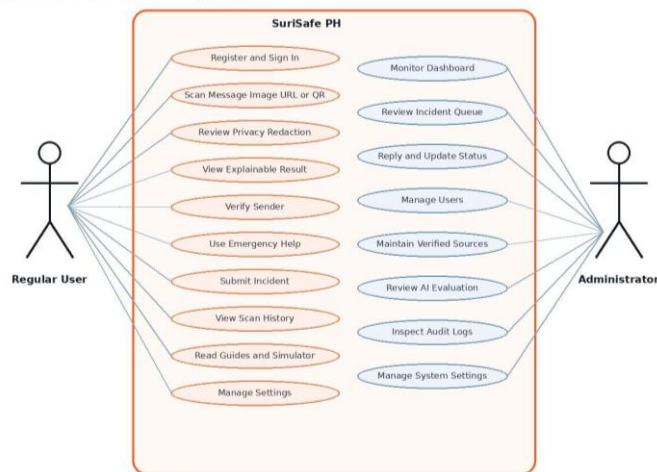


Figure 4: Use Case Diagram for Regular Users and Administrators

Prototype Interfaces

The prototype comprises 20 user interface (UI) screens, UI 01 to UI 20, shown in Figures 5 to 8. Consistent with the audit criterion, these screenshots confirm that each interface exists but do not demonstrate that the underlying authentication, persistence, integration, or action executed successfully.

Account Entry, Scanning, and Verification. Users sign in by email and password or Google sign-in; administrator access depends on the stored account role, and public registration never assigns an administrator role. The dashboard presents a dominant Start New Scan action, and input is reviewed for privacy before transmission or storage. The QR scanner requests camera permission only after user intent, decodes once, displays the destination, and never opens it automatically, while sender verification records the claimed identity, method, warning signs, and source status (Figure 5).

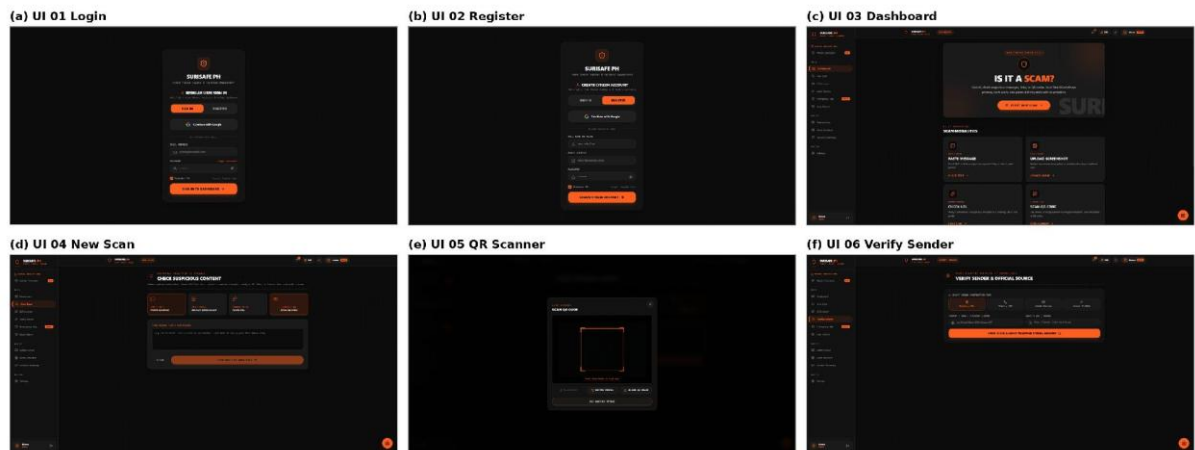


Figure 5: Account Entry, Scanning, and Verification Interfaces

Response, Learning, Reporting, and Preferences. Emergency Help asks what the user has already done and prioritizes containment and official-contact steps, and the unified Scan History holds scans, verifications, and incident reports with type, risk, and status filters. The Safety Center provides lessons on common scams in a Filipino context, and the official-agency directory lists reporting channels separately from AI output. The simulator uses only fictional scenarios with no active malicious destinations, and the incident form prepares a redacted summary intended to create a case number, appear in Scan History, and enter the administrator queue (Figure 6)

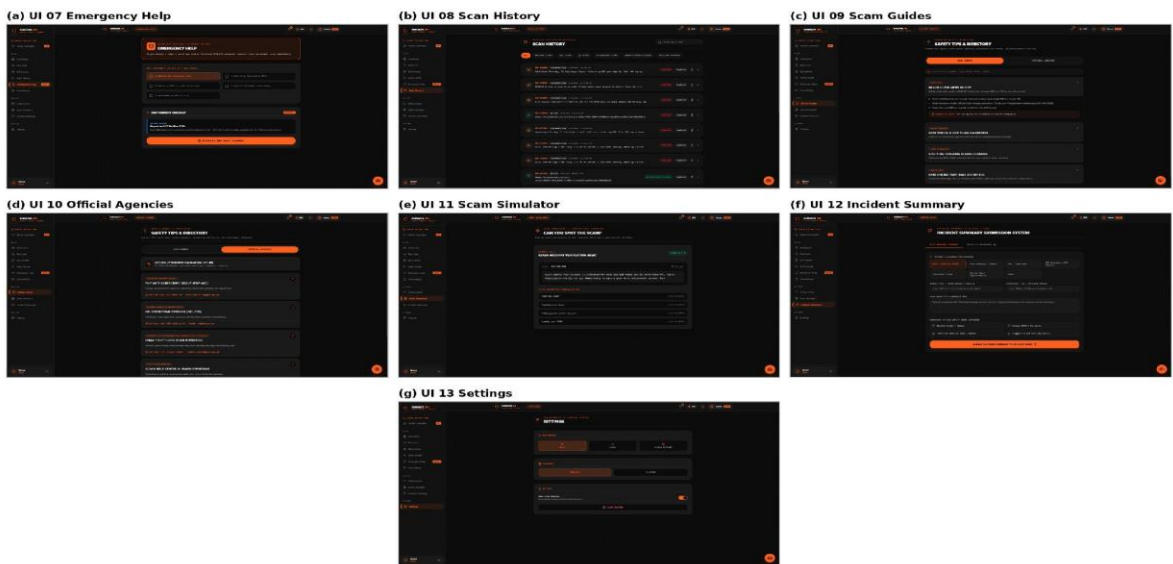


Figure 6: Response, Learning, Reporting, and Preference Interfaces

Note. (a) Emergency help; (b) scan history; (c) scam guides; (d) official agencies; (e) scam simulator; (f) incident summary; (g) user settings.

Administration. Administrators review redacted evidence, assign cases, reply, and update status, with user-visible replies kept separate from internal notes. User management never displays credentials. The Verified Sources screen currently shows STANDBY and therefore does not demonstrate live registry access. Audit logs record the actor, action, record, timestamp, and previous and new state, and secret values in System Settings remain server-side (Figure 7).

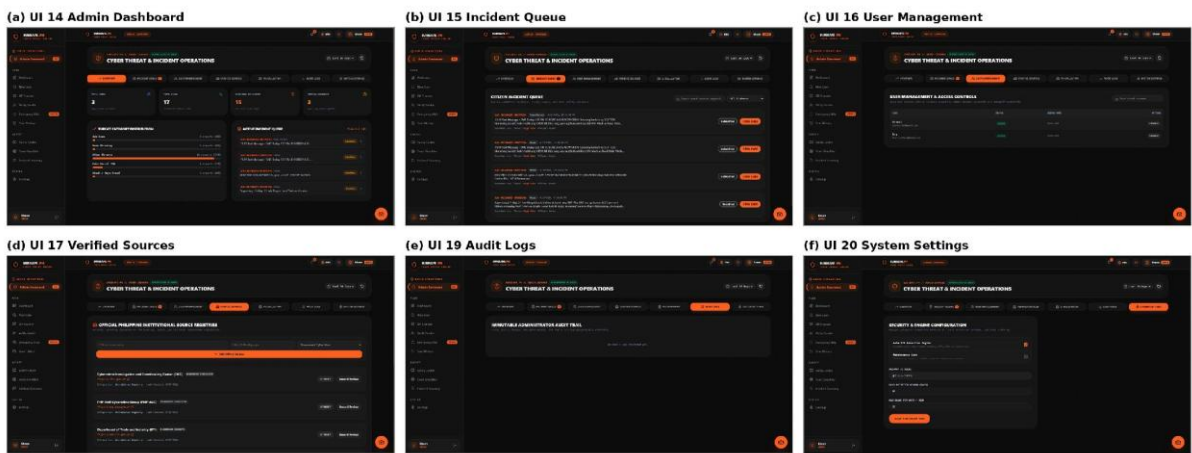


Figure 7: Administrator Interfaces

Note. (a) Dashboard; (b) incident queue; (c) user management; (d) verified sources; (e) audit logs; (f) system settings.

Comparison With Existing Services

Within the reviewed documentation (Table 4), Google Safe Browsing and VirusTotal offered the most complete URL and domain checking, and Google Messages offered native SMS or chat scam detection. None of the four existing services documented privacy review and redaction, separation of AI and verified-source evidence, or Filipino-context learning with a simulator, and only BSP Verifier partially supported emergency recovery and incident preparation. SuriSafe PH implemented all 10 functions at the interface level but was rated PARTIAL for URL checking and screenshot analysis, reflecting its dependence on an AI function and registry integrations that were not yet live.

Table 4: Evidence-Keyed Comparison of SuriSafe PH With Existing Services

Function	Google Messages	Google Safe Browsing	VirusTotal	BSP Verifier	SuriSafe PH
SMS or chat scam detection	YES	NO	NO	NO	YES (UI 04)
URL or domain checking	LIMITED	YES	YES	PARTIAL	PARTIAL (UI 04)
Screenshot or OCR analysis	NO	NO	LIMITED	NO	PARTIAL (UI 04)
QR destination safety	NO	PARTIAL	PARTIAL	NO	YES (UI 05)
Plain-language warning cues	LIMITED	LIMITED	TECHNICAL	ADVISORY	YES (UI 03, 09)
AI and verified-source separation	NO	NO	NO	NO	YES (UI 17, 18)
Privacy review and redaction	NO	NO	NO	NO	YES (UI 04, 12)
Emergency recovery checklist	NO	NO	NO	PARTIAL	YES (UI 07)
Editable incident summary	NO	NO	NO	PARTIAL	YES (UI 12)
Filipino learning and simulator	NO	NO	NO	ADVISORY	YES (UI 09, 11)

Note. Ratings for existing services are keyed to their official documentation (Bangko Sentral ng Pilipinas, n.d.; Google, n.d.-a, n.d.-b; VirusTotal, n.d.); ratings for SuriSafe PH are keyed to the interface in parentheses and reflect interface-level implementation, not verified end-to-end function.

Functional Check of the Automated Assessment

In the six-example check (Figure 8), the automated assessment produced four true positives and two true negatives, with no false positives or false negatives, so the screen reports 100% accuracy, precision, and recall and an F1 score of 1.000. Because the sample is very small, unbalanced, and not held out, this result confirms only that the evaluation pipeline runs; it does not estimate real-world accuracy.

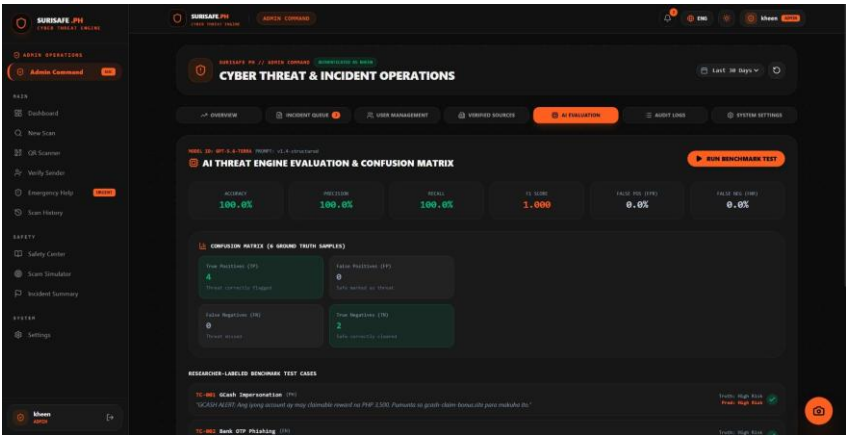


Figure 8: Administrator AI Evaluation Screen (UI 18) for Six Researcher-Labeled Examples

Note. The displayed metrics apply only to these six examples and are not calibrated real-world accuracy.

Discussion

The strongest design decision in SuriSafe PH is the separation between automated assessment and verified source status, which prevents a model-generated explanation from being presented as an official institutional check. The design operationalizes findings on explainable warnings (Aneke et al., 2021; Buono et al., 2023; Desolda et al., 2023; Wang et al., 2025) by pairing the specific cues behind each result with a recommended action. Its QR safeguards respond to evidence that users struggle to recognize fraudulent QR requests and that decoded content can enter unsafe flows (Han et al., 2023; Kowalewski et al., 2025), and its reusable incident summary addresses the reporting burden identified in national survey data (Bangko Sentral ng Pilipinas, 2022). Consistent with reference-based detection (Liu et al., 2024), institutional verification is treated as a distinct evidence source that must be identified and live before it is displayed as verified.

SuriSafe PH complements, rather than replaces, the services in Table 4. Google Safe Browsing and VirusTotal remain stronger for URL and domain reputation, and BSP Verifier remains the authoritative channel for checking whether a financial institution is regulated. The contribution of SuriSafe PH lies in organizing evidence, explaining risk indicators, and directing the user to the correct official channel within a single journey. In relation to the Sustainable Development Goals, the Safety Center, scam guides, and fictional simulator directly serve SDG Target 4.4 by providing digital-safety micro-learning in a Filipino context. The explainable warnings and privacy review protect users of electronic wallets and online banking, supporting the inclusive financial access sought under Target 8.10. The redacted incident summary and official-agency directory lower the reporting burden identified by the Bangko Sentral ng Pilipinas (2022) and connect users with law-enforcement channels, supporting Target 16.4. These contributions remain potential until confirmed through the planned usability evaluation.

Several limitations apply. The technical implementation is not production-complete: authentication, persistence, storage, and cross-user access control require controlled testing; RLS was not implemented; registries remain on standby; and the OpenAI screenshot function requires an end-to-end retest. The six-example check cannot support any accuracy claim, and confidence levels are uncalibrated (OpenAI, n.d.-a). The SuriSafe PH ratings in Table 4 rest on interface-level evidence, and the ratings of existing services are limited to their reviewed documentation. Finally, no user-based evaluation has been conducted, so usability and behavioral effects remain to be established.

Conclusions and Recommendations

SuriSafe PH demonstrates how an explainable, privacy-preserving, and Filipino-focused workflow can help users inspect suspicious content, understand its warning signs, and act safely before sharing money or private information. Its central contributions are the separation of automated assessment from verified source status, qualitative rather than numeric confidence, privacy review before analysis, safe QR handling, and an integrated path from assessment to incident reporting. The prototype is not yet production-ready. Nonetheless, its design demonstrates how a locally contextualized digital-safety tool can contribute to SDGs 4, 8, and 16 by combining scam-awareness education, safer access to digital financial services, and easier incident reporting.

Software Access and Supporting Materials

Software Access: <https://suri-safe-frontend.vercel.app/>

User Guide: Available through the software interface.

Demonstration Video: <https://tinyurl.com/2tv8j6kd>

References:

- Aneke, J., Ardito, C., & Desolda, G. (2021). Help the user recognize a phishing scam: Design of explanation messages in warning interfaces for phishing attacks. In *HCI for cybersecurity, privacy and trust* (pp. 403–416). Springer. https://doi.org/10.1007/978-3-030-77392-2_26
- Bangko Sentral ng Pilipinas. (2022). 2021 financial inclusion survey report. <https://www.bsp.gov.ph/Inclusive%20Finance/Financial%20Inclusion%20Reports%20and%20Publications/2021/2021FISToplineReport.pdf>
- Bangko Sentral ng Pilipinas. (n.d.). BSP verifier. <https://www.bsp.gov.ph/SitePages/FinancialStability/BSPVerifier.aspx>
- Buono, P., Desolda, G., Greco, F., & Piccinno, A. (2023). Let warnings interrupt the interaction and explain: Designing and evaluating phishing email warnings. In *Extended abstracts of the 2023 CHI conference on human factors in computing systems*. Association for Computing Machinery. <https://doi.org/10.1145/3544549.3585802>
- Desolda, G., Ardito, C., & Buono, P. (2023). Explanations in warning dialogs to help users defend against phishing attacks. *International Journal of Human-Computer Studies*, 176, 103056. <https://doi.org/10.1016/j.ijhcs.2023.103056>
- Google. (n.d.-a). Google Messages spam protection. Google Support. <https://support.google.com/messages/answer/9327903>
- Google. (n.d.-b). Google Safe Browsing. <https://safebrowsing.google.com/>

- Han, X., Zhang, Y., Zhang, X., Chen, Z., Wang, M., Zhang, Y., Ma, S., Yu, Y., Bertino, E., & Li, J. (2023). Medusa attack: Exploring security hazards of in-app QR code scanning. In 32nd USENIX Security Symposium (pp. 4607–4624). USENIX Association.
<https://www.usenix.org/conference/usenixsecurity23/presentation/han-xing>
- International Organization for Standardization. (2019). Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems (ISO Standard No. 9241-210:2019).
<https://www.iso.org/standard/77520.html>
- INTERPOL. (2026, June 17). New INTERPOL report highlights escalating cyber threats across Asia and South Pacific. <https://www.interpol.int/News-and-Events/News/2026/New-INTERPOL-report-highlights-escalating-cyber-threats-across-Asia-and-South-Pacific>
- Kowalewski, M., Lassak, L., Dürmuth, M., & Schnitzler, T. (2025). Scanned and scammed: Insecurity by ObsQRity? Measuring user susceptibility and awareness of QR code-based attacks. In 34th USENIX Security Symposium (pp. 1415–1434). USENIX Association.
<https://www.usenix.org/conference/usenixsecurity25/presentation/kowalewski>
- Liu, R., Lin, Y., Teoh, X., Liu, G., Huang, Z., & Dong, J. S. (2024). Less defined knowledge and more true alarms: Reference-based phishing detection without a pre-defined reference list. In 33rd USENIX Security Symposium (pp. 523–540). USENIX Association.
<https://www.usenix.org/conference/usenixsecurity24/presentation/liu-ruofan>
- National Privacy Commission. (2021, October 7). NPC PHE Bulletin No. 21: Preventive data privacy practices against smishing. <https://privacy.gov.ph/npc-phe-bulletin-no-21-preventive-data-privacy-practices-against-smishing/>
- OpenAI. (n.d.-a). Evaluation best practices. <https://developers.openai.com/api/docs/guides/evaluation-best-practices>
- OpenAI. (n.d.-b). Images and vision. <https://developers.openai.com/api/docs/guides/images-vision>
- Supabase. (n.d.-a). Auth documentation. <https://supabase.com/docs/guides/auth>
- Supabase. (n.d.-b). Edge Functions documentation. <https://supabase.com/docs/guides/functions>
- Supabase. (n.d.-c). Row Level Security documentation.
<https://supabase.com/docs/guides/database/postgres/row-level-security>
- Supabase. (n.d.-d). Storage access control. <https://supabase.com/docs/guides/storage/security/access-control>
- United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development* (A/RES/70/1). <https://sdgs.un.org/2030agenda>
- VirusTotal. (n.d.). VirusTotal API v3 overview. <https://docs.virustotal.com/reference/overview>
- Wang, Y., Zhai, H., Wang, C., Hao, Q., Cohen, N. A., Foulger, R., Handler, J. A., & Wang, G. (2025). Can you walk me through it? Explainable SMS phishing detection using LLM-based agents. In Proceedings of the 21st Symposium on Usable Privacy and Security (SOUPS 2025) (pp. 37–56). USENIX Association.
<https://www.usenix.org/conference/soups2025/presentation/wang>

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship and publication of this article.

Funding

The author received no financial support for the research, authorship and publication of this article.