RESEARCH-ARTICLE

Seek and You Shall SOC: Blending Human Expertise with Multimodal Generative AI for Scalable Threat Prevention

DAN XU

IQBAL GONDAL

XUN YI

TEO SUSNJAK

TIMOTHY MCINTOSH

Published: 26 February 2026
Accepted: 20 February 2026
Revised: 10 October 2025
Received: 21 March 2025

Seek and You Shall SOC: Blending Human Expertise with Multimodal Generative AI for Scalable Threat Prevention

DAN XU*, RMIT University, Melbourne, Australia
IQBAL GONDAL, RMIT University, Melbourne, Australia
XUN YI, RMIT University, Melbourne, Australia
TEO SUSNJAK, Massey University - Auckland Campus, Auckland, New Zealand
TIMOTHY MCINTOSH, RMIT University, Melbourne, Australia and Cyberoo Pty Ltd, Melbourne, Australia

Large language models (LLMs) are increasingly employed within Security Operations Centres (SOCs), including SOC for Digital Risk Protection (DRP), yet their outputs often exhibit partial coverage, hallucinations, verbosity, and lack of localized insights. This article proposes a hybrid reasoning pipeline that combines multimodal LLMs with stable human-curated references to mitigate these issues, and is distinct from standard retrieval-augmented generation because offline, human-curated references are applied as an explicit decision-time *override* rather than used solely as supportive retrieved context. We introduce a step-by-step process that incorporates multi-vantage crawling for evasive content, deterministic prompts to manage inconsistency, and a structured approach to override or refine the model's classifications when local brand knowledge contradicts global assumptions, together with an analyst-governed escalation loop that records when and why overrides occur in external-SOC DRP settings. Empirical evaluations with multiple commercial and open-source model providers show that this method significantly boosts scam detection accuracy, lowers token costs through caching, and reduces misleading outputs by adopting curated domain data, including comparisons against a RAG-only configuration and classical non-LLM baselines. Results underline how offline reference injection fosters a reliable collaboration pattern that harmonizes automated tasks with human expertise, thereby enhancing scalability and trust in real-world SOC environments.

CCS Concepts: • **Computer systems organization** → *Reliability*; • **Security and privacy** → **Information flow control**; *Vulnerability management*.

Additional Key Words and Phrases: Zero Trust, Generative AI, Cybersecurity, Adversarial Attacks, Trust Mechanisms, AI Auditing

1 Introduction

Digital Risk Protection (DRP) has become an important requirement for client organizations that rely heavily on public-facing digital assets and brand channels against cyber frausters impersonating the organizations, *e.g.*, the detection and removal of fraudulent websites [2, 16], verification of official SMS communications [39, 48], identification of deceptive advertising campaigns [52], elimination of counterfeit

*Corresponding Author

Authors' Contact Information: Dan Xu, RMIT University, Melbourne, Victoria, Australia; e-mail: s3152414@student.rmit.edu.au; Iqbal Gondal, RMIT University, Melbourne, Victoria, Australia; e-mail: iqbal.gondal@rmit.edu.au; Xun Yi, RMIT University, Melbourne, Victoria, Australia; e-mail: xun.yi@rmit.edu.au; Teo Susnjak, Massey University - Auckland Campus, Auckland, Auckland, New Zealand; e-mail: T.Susnjak@massey.ac.nz; Timothy McIntosh, RMIT University, Melbourne, Australia and Cyberoo Pty Ltd, Melbourne, VIC, Australia; e-mail: timothy.mcintosh@rmit.edu.au.

ACM 1557-6051/2026/2-ART
https://doi.org/10.1145/3799420

mobile apps [56, 66], and mitigation of social media impersonations [26]. Conventional Security Operations Centers (SOCs) [7, 29], which primarily safeguard internal network perimeters through intrusion detection and threat hunting, are not fully equipped to mitigate evolving external threats such as phishing websites, brand impersonations, and fraudulent social media accounts. These elusive threats often lie beyond the direct control of the protected infrastructure and spread rapidly through third-party platforms, user devices, and novel communication channels. Organizations entrust external SOCs, operating "above the attack surface," with the task of identifying and neutralizing rogue sites and malicious content at scale before reputational or financial damages grow severe. Yet, the rapid escalation of scam campaigns and the creativity of malicious actors demand a more adaptive approach than legacy methods can provide.

1.1 Challenges in DRP and Shortcomings of Traditional Approaches

Traditional machine learning pipelines and manually written rules have been deployed to detect known phishing keywords or recognize blacklisted domains (*e.g.*, [11, 12, 72]). However, many advanced threats remain unobserved due to their nuanced, multimodal nature. Fraudsters often repurpose legitimate logos, alter brand references, embed short or regional phone numbers, or even generate near-duplicate screenshots that confuse static detection models. An organization that receives an image of a suspicious SMS in a foreign language with only a slight change to the contact phone number might find their older rule-based systems struggling to identify the threat. Large numbers of alerts and limited staff bandwidth add further pressure. Repetitive low-level verifications, especially in multilingual environments, strain even well-funded SOC teams. Consequently, the manual overhead of verifying every potential scam and producing timely guidance for stakeholders becomes unsustainable as threat volumes grow.

1.2 Generative AI as a Catalyst for Human Collaboration

Recent advances in multimodal generative language models have opened pathways to a more scalable and context-aware detection framework. Systems that can analyze textual, visual, and structured logs simultaneously are better equipped to handle ambiguous inputs, including half-captured screenshots, masked phone numbers, and partially concealed URLs. Yet, generative solutions alone do not suffice when operating at industrial scales. Many large language models require supervision to ensure consistent outputs, controlled verbosity, and correct domain references. Moreover, end users external to the SOC (such as corporate customers or everyday mobile phone subscribers) frequently submit critical evidence that automated crawlers cannot retrieve or interpret alone. This evidence enriches detection but requires interpretive reasoning to confirm authenticity and assess language-specific or brand-specific quirks. A purely technical pipeline would struggle to incorporate these factors reliably. By fusing user submissions, onsite human expertise, and generative reasoning, it becomes possible to balance automation with targeted oversight.

1.3 Contributions and Article Structure

This article introduces a practical operational framework for external-SOC DRP with an *override-first* orchestration of human knowledge and generative models, tailored to an external SOC environment focused on DRP. Rather than limiting detection to traditional approaches or purely automated scans, we integrate contributions from two categories of human agents—end users who supply insider artifacts (such as suspicious SMS captures) and onsite SOC analysts who refine brand references and supervise regulatory guidelines, with a documented analyst-governed escalation loop that determines when offline references override model hypotheses. This synergy enables the SOC to proactively discover malicious behavior while scaling beyond the capacity of existing rule-based pipelines, and it is *distinct from standard*

RAG because decision-time overrides are enforced when conflicts arise rather than treating retrieved material as optional context. Related efforts that combine knowledge grounding with phishing detection include Liu et al. [38] and Li et al. [34], though their focus on architectural enhancements, whereas we emphasize external-SOC orchestration and override logic, and focus on cyber scam (not just phishing).. Our major contributions are:

(1) **Hybrid Human–AI Pipeline for DRP:** We propose an end-to-end system that merges user-submitted evidence with passive crawls and generative language models, with an explicit decision layer that prioritizes offline, human-curated references when they conflict with model outputs rather than using them only as retrieved context. This pipeline manages nuanced brand references and multilingual content to reduce the false negatives and false positives that hamper existing detection methods.
(2) **Mutual Enhancement Through Collaboration:** We detail how end users and SOC analysts jointly strengthen the generative AI. User submissions reveal dynamic threats that passive crawling alone cannot reach, while onsite experts update localized knowledge, control costs, and comply with governance or risk requirements, operationalized through an analyst-governed escalation loop that records override conditions and outcomes.
(3) **Technical and Operational Insights:** We demonstrate the effectiveness of this approach through empirical evaluations on challenging DRP tasks, including comparisons against a RAG-only configuration and classical non-LLM DRP baselines, and reporting stability and cost behaviors under the override-first orchestration. We highlight how technical oversight (managing references, re-prompting, caching) and strategic management (imposing regulatory, budgetary, and operational constraints) align to deliver scalable automation without sacrificing precision or user trust.

Our contribution is an operationalization of human–AI collaboration for external-SOC DRP rather than a new LLM architecture; the framework's distinctiveness lies in its override-first decision layer and analyst-governed escalation loop, which turn curated references into enforceable controls at inference time.

The remainder of this article is organized as follows. §2 defines the unique SOC and DRP challenges we address. §3 formalizes the human–generative-AI collaboration with a focus on bridging offline brand intelligence and real-time threat signals. §4 details our hybrid pipeline design, highlighting how end-user input, onsite analyst oversight, and generative modeling cooperate. §5 describes the experimental setup, including datasets and cross-validation, evaluation metrics and cost analysis, and the experimental procedures used to assess the pipeline under human oversight and human management. §6 presents experiments and operational metrics, categorizing the benefits that humans and AI bring to each other. §7 explores key findings, limitations, and future directions. Finally, §8 concludes with closing thoughts on the broader implications of human–AI collaboration for external SOCs. In the appendix: §**??** lists the abbreviations used in this article. Appendix §**??** surveys relevant literature and discusses prior work on LLMs and generative AI in cybersecurity. Appendix §**??** discusses caching and resource management. Appendix §**??** provides 10 masked examples.

2 Problem Statement and Domain Context

This section presents the problem statement and domain context.

2.1 The DRP Challenge

Organizations seeking to safeguard their external digital assets and reputations depend on comprehensive protection against phishing schemes, counterfeit websites, impersonated social media channels, and other malicious online activities. These client organizations often possess a large public footprint, including

multiple brand names, domain portfolios, social media pages, and mobile applications. Defensive strategies must therefore extend beyond conventional in-house monitoring to include external scanning of the open internet, as well as targeted takedown actions when threats are confirmed.

A core requirement in DRP is the ability to collect pertinent artifacts from diverse sources. Passive crawls of websites can uncover unauthorized copies of legitimate pages, while brand monitoring engines may reveal suspicious domains or visually deceptive logos. However, the threat landscape extends into end-user devices and private communication channels, such as SMS or instant messaging, which cannot be fully accessed through automated crawlers. End users, who often receive scam messages in multiple languages and formats, submit screenshots or forwarded texts that provide crucial evidence. Robust DRP consequently hinges on this dual-input model: passive crawling of publicly accessible resources and active contributions of private artifacts from external users.

2.2 What Our SOC Does

Unlike traditional SOCs focusing on threats within an organization's network perimeter, our external SOC operates "above the attack surface." Rather than deploying intrusion detection sensors on internal systems, we monitor and mitigate external-facing threats before they impact end users or corporate reputations. We continuously scan domains, social platforms, underground marketplaces, and certificate repositories to detect malicious impostors, phishing sites, or brand exploitation.

Once these external threats are identified, we coordinate with registrars, hosting providers, and relevant authorities to request prompt takedowns. This external posture requires a broad vantage over public cyberspace. It also brings a unique operational complexity because covert attacks, like smishing or social media impersonation, commonly bypass central logging or conventional perimeter defenses. Accordingly, external SOC teams rely on both user-generated inputs and open-source intelligence to strengthen detection and takedown workflows, which entails diverse data streams and further complicates classification tasks.

2.3 Why Traditional Methods Fail

Industry-standard approaches frequently rely on text-based machine learning classifiers or manually curated rules, which often prove inadequate for nuanced or multilingual threats. When scammers slightly revise a legitimate brand's messaging by substituting a phone number, conventional keyword-based detectors typically miss the alteration because the textual similarity is high. Computer vision methods like YOLO can similarly fail if a screenshot has only partial text or if the attacker disguises malicious hints within realistic brand imagery. Moreover, socially engineered attacks may be context-dependent: scammers might match local dialects or regional phone prefixes, further complicating detection.

Even when detection is partially successful, legacy methods rarely produce clear explanations or user-specific guidance. For instance, an internal rule-based system may flag suspicious content based on a known malicious URL, yet it cannot advise an end user, who submitted a screenshot, on the specific nature of the underlying scam or the risk that a newly observed phone number poses. As a result, organizations relying on these older workflows often see high false negatives and a debilitating volume of manual reviews.

2.4 Why Generative AI

The rapid expansion of advanced phishing and scam operations has created a demand for more adaptive detection and response mechanisms. Generative language models, capable of synthesizing text, images, and contextual metadata, offer a flexible platform for reasoning over varied evidence. They can detect

nuanced patterns by referencing brand documentation, recognized contact channels, or official subdomains, and they can produce coherent responses—even in multilingual scenarios—to educate external users.

However, deploying generative language models in DRP contexts raises practical concerns. Models can over-generate irrelevant text, hallucinate plausible but incorrect details, and omit region-specific constraints if not carefully directed. These shortcomings mandate a robust system of human oversight and organizational governance, or "management," that collectively refine prompts, verify critical outputs, and ensure compliance with industry regulations. Without such checks, generative technology risks creating noise or misdiagnoses that diminish trust and inflate operational costs.

2.5 Why Passive Crawling Alone Is Not Enough

Although large-scale crawling of public websites remains a cornerstone of DRP, it alone cannot uncover many scams or misuses concealed in private or limited-access environments. For example, a malicious URL may redirect desktop users to a benign corporate domain but display phishing content only to mobile visitors. Scammers also exploit ephemeral infrastructure—like disposable short links, one-time phone numbers, or region-specific channels—well before these resources are recognized by conventional threat databases.

End-user submissions thus become indispensable. By uploading screenshots, suspicious SMS text, or ephemeral links, users enrich the SOC's vantage. These artifacts reveal attack vectors that purely passive crawls fail to capture. The external SOC then combines vantage-based crawling with human-supplied evidence in real time to identify and takedown threats that would otherwise evade detection. As the next sections will illustrate, enabling this synergy requires a structured framework in which both onsite analysts and external users collaborate with a generative language model to form an adaptive, scalable DRP capability.

3 Theoretical Framework

This section establishes the theoretical underpinnings of our proposed human–generative-AI collaboration for Digital Risk Protection (DRP). We begin by identifying the three key entities in this collaboration: end users, onsite analysts, and the generative model. We then introduce two governing pillars—*human oversight* and *human management*—that ensure the system is both technically sound and aligned with strategic, regulatory, and financial imperatives. Following this conceptual setup, we formalise the collaboration through equations that address knowledge discrepancies, iterative alignment, offline references, and cost constraints. This framework provides a rigorous basis for the hybrid pipeline design in Section 4, highlighting how each generative inference is grounded in curated offline knowledge and guided by end-user submissions, onsite analyst expertise, and management-imposed constraints.

3.1 Entities in Collaborative DRP

The entities in our collaborative DRP include humans (end users, onsite analysts) and generative LLM.

End Users are external stakeholders or customers who submit private artifacts, such as suspicious SMS screenshots or app screenshots that passive web crawlers cannot access. Their inputs furnish real-world evidence of emerging scams, thereby exposing regional or ephemeral threats that are otherwise elusive.

Onsite Analysts, operating within the SOC, maintain and refine brand references, verify ambiguous evidence, and guide generative language model (LLM) outputs. They correct hallucinations, integrate region-specific phone prefixes, and enforce consistent risk scoring.

Generative LLMs synthesize textual, visual, and metadata cues at scale, dynamically adapting to the varied nature of DRP threats. Though highly flexible, they require structured guidance to control verbosity, minimise hallucinations, and incorporate evolving local knowledge.

3.2 The Two Pillars: Human Oversight and Human Management

Human Oversight and Human Management: Two Pillars of Effective SOC Collaboration

Human Oversight refers to the technical governance performed by cybersecurity analysts, data scientists, and other domain experts who refine the pipeline's references, craft and adapt prompt designs, examine LLM outputs for accuracy and threats, and integrate local brand or regulatory knowledge into the system's core logic. These experts address ambiguities (for instance, distinguishing a six-digit phone number from a verification code), perform iterative risk analysis, and maintain alignment between the generative model's reasoning and the SOC's security requirements. Their continuous refinements reduce misclassifications, update knowledge bases with newly uncovered short codes or domains, and correct hallucinations before they propagate to end-users.

Human Management denotes the oversight executed by governance, risk, and compliance (GRC) practitioners, along with project managers and policy makers, who ensure that the technical decisions made by oversight teams remain aligned with cost targets, customer-facing needs, regulatory obligations, and enterprise priorities. These managers approve vendor selection (including permissible LLMs in certain regions), define budgets for multi-pass or advanced inference, specify acceptable false positive rates, and incorporate user feedback on transparency and user experience. They also determine how the pipeline is scaled, how logs are stored, and how updates are operationalized within evolving compliance mandates.

Figure 1 illustrates the cyclical synergy among these two pillars, the generative model, and the ultimate operational or business outcomes. Human oversight yields refined technical solutions, while human management sets and enforces strategic objectives, resource allocations, and regulatory thresholds. This joint feedback loop allows generative language models to surpass static detection processes by assimilating context-specific references, user feedback, and evolving threat insights.

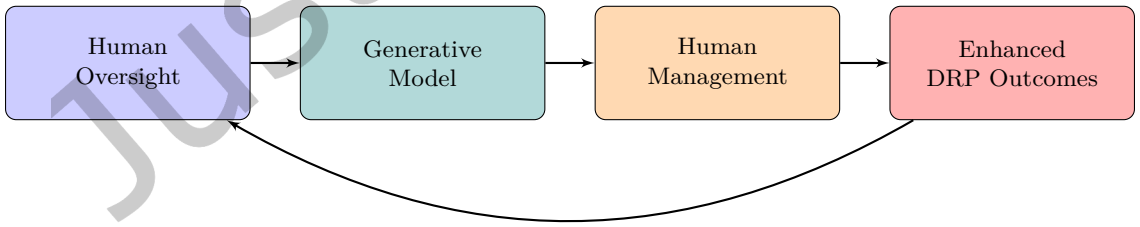

Fig. 1. Cyclical synergy among human oversight, the generative model, and human management, culminating in enhanced external SOC outcomes. This loop ensures that iterative refinements and resource constraints evolve in tandem, leading to sustainable, domain-specific DRP.

3.3 Formalising the Collaboration

The collaboration merges three streams of input: (1) *end-user evidence*, typically in the form of screenshots or text logs from private channels; (2) *onsite analyst inputs* such as validated whitelists, short codes, and brand references; and (3) *generative modeling* capable of reasoning over ambiguous or multilingual artifacts. The generative language model M processes these inputs, yet domain-specific knowledge is housed in a curated repository H under human oversight.

3.3.1 Human–AI Knowledge Discrepancies. LLMs often neglect localized or brand-specific facts unless explicitly provided. We define a discrepancy function between the curated knowledge H and the model's internal representation M:

$$\Delta H, M \;=\; \sum_{k=1}^{K} w_k \, \phi\big(f_k H,\, f_k M\big), \tag{1}$$

where K is the number of reference categories (e.g., phone prefixes, brand channels), w_k weights each category by its importance, and $\phi\cdot,\cdot$ measures mismatches in domain labels or brand attributes. A high $\Delta H, M$ suggests that M is missing or misrepresenting vital domain knowledge, prompting more targeted human oversight (e.g., injecting validated phone codes) or, in extreme cases, management directives to limit or reconfigure queries.

Decision-time gating and measurable outcome link. We operationalize $\Delta H, M$ with a logistic gate that blends raw and corrected scores: $\alpha\mathbf{x} = \sigma\big(\gamma\,\Delta H, M; \mathbf{x} - \tau\big)$ and $s^\star\mathbf{x} = 1 - \alpha\mathbf{x}\, s_M\mathbf{x} \;\; \alpha\mathbf{x}\, s_\Psi\mathbf{x}$, where s_M is the model score and s_Ψ is the score after knowledge correction via Ψ. The threshold τ and slope γ are fit on the training folds to minimize Brier loss subject to an FPR cap set by management. The resulting *override rate* (share of items with $\alpha\mathbf{x} > 0.5$) is logged and analyzed alongside FNR and F1 in §6 and in the ablations of §**??** (Table **??**).

3.3.2 Chain-of-Thought, Instability, and Iterative Re-Prompting. Generative language models can produce inconsistent answers if their sampling parameters (e.g., temperature, random seeds) are not tightly controlled. We quantify instability by sampling N outputs for the same input and measuring their pairwise divergences:

$$\text{Instability}\mathbf{x}, N \;=\; \frac{1}{NN-1} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} d\big(\Phi\mathbf{x}, r_i,\, \Phi\mathbf{x}, r_j\big). \tag{2}$$

Human oversight mitigates such fluctuations by curating prompts and temperatures, while human management caps the total number of repeated inferences to align with cost or latency requirements.

When the model's initial output $\mathbf{r}_1$ mismatches brand references in H, analysts initiate a re-prompting cycle:

$$\mathbf{r}_2 \;=\; \Phi\big(\mathbf{r}_1 \oplus H,\; \mathbf{p}'\big), \tag{3}$$

where $\oplus$ denotes explicit knowledge injection. Each subsequent re-prompt reduces misalignment by a fraction of the previous error, in a geometric convergence. Although repeated queries improve accuracy, they inflate resource usage, necessitating oversight for prompt tuning and management for cost-control thresholds.

Calibration of fusion to outcomes. We estimate β_s and β_u by nested cross-validation to maximize macro-F1 while constraining FPR to a management target. The fused signal is fed into the gate above: when reliable sources or well-corroborated user evidence raise $\omega\cdot$ and $\Delta H, M$ is high, $\alpha\mathbf{x}$ increases, making

the system adopt s_Ψ more often. This mechanism links the construct directly to observed reductions in FNR and increases in F1 reported in §?? (Table ??).

3.3.3 Confidence-Weighted Data Fusion. DRP tasks require synchronizing open-source intelligence (e.g., domain blocklists) with private user-submitted artifacts. We combine these signals via a weighted aggregator:

$$\mathbf{X} \;=\; \overset{n}{\underset{i=1}{}}\omega\big(s_i\big)\,\mathbf{S}_i \;\cup\; \overset{m}{\underset{j=1}{}}\omega\big(u_j\big)\,\mathbf{U}_j, \tag{4}$$

where each $\mathbf{S}_i$ is a structured source and each $\mathbf{U}_j$ is user evidence. Human analysts refine reliability scores Rs_i and cross-verification scores Cu_j, which feed logistic weighting functions:

$$\omega\big(s_i\big) = \frac{1}{1\;\; e^{-\beta_s Rs_i}}, \quad \omega\big(u_j\big) = \frac{1}{1\;\; e^{-\beta_u Cu_j}}. \tag{5}$$

This fusion ensures that generative inferences are anchored in brand-authoritative knowledge while still incorporating emerging signals from external submitters.

3.3.4 Offline References as Retrieval-Augmented Generation. Global LLMs can misclassify brand references if they have never encountered the specific phone formats or local domains. We treat the curated repository H as an offline retriever within a retrieval-augmented generation (RAG) workflow. Formally, if $\mathbf{r}_1 = \Phi\mathbf{x}$ conflicts with known facts in H, a corrective function Ψ reconciles these discrepancies:

$$\mathbf{r}_{\text{corrected}} \;=\; \Psi\big(\mathbf{r}_1,\, H\big), \tag{6}$$

overriding or clarifying generative outputs that deviate from verified knowledge. Onsite analysts update H as threats evolve, while management maintains usage policies and monitors the overhead of large reference lookups.

Repository governance signals. Each repository entry $e \in H$ carries provenance and validation attributes used to compute a trust score $Te = \sigma\big(a\,\text{prove} \;\; b\,\text{agree}e - c\,\text{conflict}e - d\,\text{age}e\big)$. Entries with Te below a quarantine threshold are withheld from Ψ and do not influence $\alpha\mathbf{x}$. This guards against accidental or adversarial insertions and provides an auditable link from H to decision-time behavior.

3.4 Cost and Resource Constraints

Due to frequent queries and iterative re-prompts, LLMs can accumulate significant resource consumption. We model the cost of a query $\mathbf{q}$ under random seed r as

$$\mathcal{C}\mathbf{q}, r \;=\; c_t \cdot \text{Tokens}\mathbf{q}, r \quad c_\ell \cdot \text{Latency}\mathbf{q}, r, \tag{7}$$

where c_t and c_ℓ represent per-token and per-latency costs, respectively. Caching mechanisms store partial chain-of-thought outputs and classification decisions, preventing redundant compute for repeated user submissions. Human oversight fine-tunes caching strategies to maintain consistent accuracy, while human management enforces budgetary guidelines, specifying how many re-prompts or retrieval operations are permissible. This cost-aware structure ensures that generative analyses can expand in scope without surpassing the enterprise's operational or regulatory thresholds.

4 Hybrid Reasoning Pipeline Design and Implementation

We propose a multi-phase solution for automating Digital Risk Protection (DRP) within a Security Operations Centre (SOC) by integrating the theoretical constructs from Section 3 with two complementary pillars: **human oversight** and **human management**. The former governs technical feasibility—implementing strategies such as prompt refinement, iterative alignment, and reference injection—while the latter

addresses cost control, regulatory compliance, and strategic decision-making. Each pipeline step leverages human oversight for technical optimization (for example, verifying local brand references or screening LLM hallucinations) and human management for determining vendor viability, resource allocation, and acceptable operational thresholds. Together, these functions ensure that the pipeline remains both technically robust and operationally sustainable.

4.1 Overall Architecture for a DRP-Focused SOC

The proposed pipeline consists of five principal stages (Figure 2): *(1) passive crawling* to harvest global threat signals, *(2) user submissions* that capture private or localised evidence, *(3) multimodal analysis* incorporating OCR and brand-referencing logic, *(4) knowledge injection and generative reasoning* that leverages curated domain references while minimising hallucinations, and *(5) final classification or actionable output*. Human oversight steers each technical phase by refining prompts, verifying ambiguous cases, and checking model responses against domain-specific knowledge; the broader strategic and governance aspects of human management, including cost control and regulatory compliance, are subsequently layered on top (see Sections that follow).

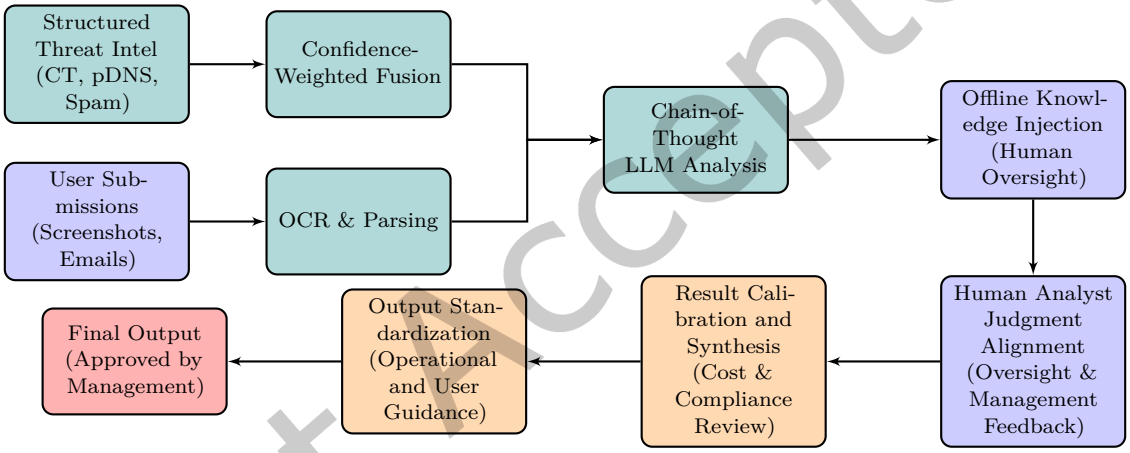

Fig. 2. Hybrid pipeline integrating technical (human oversight and enhanced technical outcomes) and strategic (improved human management and refined business outcomes) controls. Operational constraints from management directly shape technical decisions, and technical refinements feed back into management processes.

Figure 2 sketches the end-to-end workflow, revealing how modules reliant on human oversight (blue and green) interconnect with feedback loops (yellow and red) that reflect enterprise priorities. Proactive crawling gathers external threat intelligence, which is fused with user-submitted evidence of emerging scams. A chain-of-thought generative language model then analyzes each item, referencing an offline knowledge base when ambiguous attributes (e.g., phone prefixes or short codes) might otherwise lead to incorrect inferences. Any problematic output triggers prompt revision or brand-specific injections overseen by onsite analysts. This mechanism ensures that new, regionally tailored intelligence is continuously integrated without retraining the generative model.

Although the illustration distinguishes certain blocks—such as the OCR and fusion modules for text normalisation or the alignment step that handles borderline classifications—the pipeline is highly modular. The interplay of human expertise and generative reasoning evolves with each new threat vector or brand

Table 1. Representative Types of End User Submissions, Their Approximate Frequency, and Difficulty of Obtaining the Same Information via Public Crawling

Submitted Content Type	Approx % of Total	Difficulty via Public Crawling
Suspicious SMS	42%	Nearly impossible (private channel)
Phishing Emails	26%	Low, unless leaked on forums
Social Media Posts	14%	Moderate, often behind personal privacy settings
Malvertisements	9%	High, ephemeral ads not always indexed
Misc. Documents/Images	9%	Variable, many stored on private drives

requirement. By allowing local references and re-prompts, the pipeline keeps false positives manageable and reduces the risk of overlooking novel scams. Crucially, the final decisions or takedown actions are only issued once the aggregated insights pass scrutiny from both technical analysts (human oversight) and strategic management teams (human management), creating a balanced environment for sustainable Digital Risk Protection.

4.2 End User Collaboration

In addition to passive crawling, external users submit private artifacts that are inaccessible to standard web scrapers. This submission channel, typically exposed through a secure portal or mobile application, enables everyday consumers, enterprise clients, or internal help desk staff to upload suspicious messages, images, or partial domain references. Such evidence often pinpoints emerging scams that redirect traffic to malicious pages only under specific conditions, such as a mobile device's user agent. By tapping into these user submissions, the SOC significantly broadens its detection range beyond the data gleaned from public-facing crawls.

User Input Categories and Challenges. Table 1 illustrates the major types of submissions commonly received from end users. These categories highlight both the importance of direct user reports and the difficulty of intercepting such data via purely public crawls. For instance, short-lived phishing SMSes are seldom publicly indexed, yet they regularly fool recipients by mimicking official notifications from banks or government agencies. Without user cooperation, these private messages remain invisible to most detection strategies.

Triggering Modules and Real-Time Guidance. When a user uploads an image or text snippet, the pipeline automatically performs optical character recognition (OCR) and brand matching to convert the file into structured data suitable for subsequent analysis. The generative language model then examines the extracted content. If further disambiguation is necessary, it prompts the user through a short interactive exchange that provides simple instructions or questions in the user's preferred language. For instance, if the text mentions a bank's short code that does not match existing references, the model may request clarity: *"Does this code appear in other legitimate messages from your bank?"* Such tailored interactions foster trust and facilitate more accurate classification.

Plain-Language Explanations and User Trust. Once the pipeline completes its evaluations, a concise narrative is returned, giving the user immediate insight into whether the submission likely constitutes a scam, phishing attempt, or legitimate variation of an official message. These explanations avoid obscure technical jargon and include, if warranted, recommended steps (e.g., "Contact your bank's verified hotline" or "Disregard this message as it is harmless"). This plain-language communication, refined by human oversight and validated by management for correctness and compliance, reinforces end-user confidence in

the SOC's capabilities. As a result, overall engagement increases, feeding further high-quality data back into the pipeline and strengthening the broader generative detection framework.

4.3 Onsite Analyst Collaboration

The SOC's onsite analysts oversee brand references, compliance requirements, and iterative refinements to the generative model's decision flow. Rather than relying solely on static or global information, these experts continuously adapt local knowledge bases to reflect new short codes, updated domain patterns, or recently uncovered phishing styles. When discrepancies emerge—for example, a bank's name that appears with an unconventional spelling in a suspicious message—analysts validate or correct the classification outcome before the pipeline escalates it to final takedown or user notifications.

Brand-Specific Oversight and Offline References. As described in Section 4, analysts integrate brand data into the pipeline through an offline knowledge repository that contains official contact numbers, validated subdomains, and region-specific whitelists or blacklists. They monitor the generative language model for hallucinations or inconsistent handling of local phone prefixes, intervening with direct injections when the model's chain-of-thought conflicts with verified references. This oversight mechanism prevents misinterpretations that might arise if a brand changes its short code or if a foreign-language domain closely resembles a legitimate local service.

Cost and GRC Constraints. Onsite analysts also coordinate closely with governance, risk, and compliance (GRC) teams to ensure that each detection step is cost-effective and meets regulatory standards. If repeated re-prompts or resource-intensive vantage crawls exceed the allocated budget, analysts either refine prompt strategies or place explicit caps on certain categories of queries. Conversely, if certain phishing patterns are deemed high-risk, management may permit additional scrutiny or re-prompt passes, trusting that analyst oversight will keep false alarms under control.

4.4 Repository H: schema, update cadence, and anti-poisoning safeguards

Schema. H stores typed entities `{phone_number, domain, short_code, brand_alias, locale}` with fields for provenance (source, submitter role, timestamp), evidence hashes, and validation votes. Phone numbers follow E.164 with locale tags; domains include registrar and WHOIS snapshots.
Update cadence. Daily batched merges with per-brand hotfixes; all changes are signed and versioned with reversible diffs.
Two-person control. Any addition that can change a decision path (e.g., short codes, hotline numbers) requires dual approval (analyst and reviewer).
Conflict resolution. Deterministic precedence rules (newer evidence with higher trust Te supersedes stale entries); contradictions trigger quarantine until adjudicated.
Automatic checks. Format and range checks (e.g., regional prefix validity), near-duplicate detection, and black-box canaries that assert benign templates must remain benign.
Integration to scoring. The trust score Te feeds Rs_i and Cu_j in the fusion (§3), which in turn modulates the gate $\alpha\mathbf{x}$. This closes the loop between data governance and the empirical gains reported in §**??**.

4.5 Generative AI Components

The generative language model at the core of this pipeline combines chain-of-thought reasoning with explicit brand reference checks to balance adaptability and domain precision. The system operates at

a low temperature to minimize randomness, generating deterministic and concise rationales that allow analysts to trace how the model arrived at a given classification.

Chain-of-Thought Reasoning. Chain-of-thought outputs reveal the intermediate steps the model uses to interpret suspicious text or screenshots. While such transparency helps analysts identify points of confusion—e.g., incorrectly mapping a telephone number onto a known short code—it can also increase verbosity. Human oversight thus calibrates prompt templates and temperature settings to keep explanations as short, factual, and actionable as possible. In multilingual scenarios, especially when end-user submissions include partial translations or brand keywords in multiple languages, these prompts simplify the generative model's reasoning by providing explicit brand or language cues.

Selective Re-Prompting. When the system encounters ambiguous signals (e.g., local phone numbers that differ from known references by one digit), it may generate contradictory outputs over multiple inference passes. Onsite analysts employ re-prompting selectively, merging partial chain-of-thought evidence with brand-specific references until the discrepancy is resolved or deemed too minor to warrant further cost. This iterative process—described in Section 3—commonly concludes after one or two passes in production workflows, minimizing overhead while preserving accurate judgments.

*Caching and Resource Management (Appendix **??**).* The pipeline minimizes repeated token usage by caching partial results for near-duplicate user inputs, which, along with a balanced cost–benefit strategy, ensures large-scale efficiency and stable runtime performance. Full details, including the algorithm and specific resource-management considerations, are provided in Appendix **??**.

5 Experimental Setup

This section details our experimental setup, which evaluates the pipeline from a dual perspective: technical performance under human oversight and operational feasibility under human management. Such a dual framework is essential for a holistic assessment of human–generative-AI collaboration in Security Operations Centres (SOCs).

5.1 Datasets and Cross-Validation

Our study relies on a curated dataset of 10,000 labeled samples spanning multiple threat categories and benign cases, thereby reflecting real-world conditions in DRP. Importantly, **the dataset is proprietary; our human cybersecurity analysts had previously classified past incidents, ensuring high-quality labels without incurring additional cost or time overhead during the development of our automated generative AI–based system**. The dataset was split into 90% for training and 10% for testing. We employed a 10-fold cross-validation strategy on the 9,000 training samples to capture variations across different data folds. Each fold reinitialized the entire pipeline—from chain-of-thought inference to offline knowledge injection—thereby tracking cost metrics and convergence rates in an unbiased manner.

Why not use an existing public dataset? Public corpora for phishing or fraud are typically URL- or text-centric and do not capture the broader DRP surface that we target (e.g., social media impersonations, fake mobile apps, brand abuse, and regional SMS short codes). They also rarely provide analyst-adjudicated, case-level ground truth spanning multimodal artifacts and escalation outcomes. In contrast, DRP operations frequently involve heterogeneous evidence (screenshots, short codes, app manifests, posts, and takedown artifacts) and governance steps that public sets do not encode, leading to an evaluation mismatch if used as the sole benchmark.

Why a dedicated "cyber scam" corpus is hard to assemble. Our focus is broader than classic web phishing: we detect contemporary cyber scams across channels where ground truth is often undecidable in real time. Scammers regularly rotate infrastructure or shutter command-and-control servers, so independent re-validation can become impossible days later. Victim confirmation and platform responses are intermittent, and legitimate campaigns may resemble abuse (e.g., near-identical SMS short codes across regions). These factors yield inherently ambiguous items (e.g., when only an image is available—such as a logo screenshot—without accompanying context, and it is infeasible to determine legitimacy); our adjudication policy therefore excludes $\approx 15\%$ uncertain cases from headline scores (reported below) while retaining their metadata to improve references and future audits.

Table 2. Dataset Distribution with 90/10 Split and 10-Fold Cross-Validation

Threat Type	**Total**	**Training (90%)**	**Validation (10%)**	**Examples of Key Indicators**
Phishing Websites	3500	3150	350	Forged login screens, suspicious domains
Fake Mobile Apps	2000	1800	200	Counterfeit APK links, cloned branding
Social Media Impersonations	500	450	50	Fake profiles, impersonated brand pages
Brand Abuse	2000	1800	200	Typosquatted URLs, unauthorized usage
Benign Websites (*non-threat*)	2000	1800	200	Legitimate brand channels, verified sites
Total	**10000**	**9000**	**1000**	

Labeling was performed by onsite human cybersecurity analysts, whose independent verdicts on each sample's malicious or benign status (as part of routine screening and audit work) were reconciled via majority voting. Ambiguous items (approximately 15% of the dataset) were omitted from final performance scoring. Notably, this process embodies human oversight: analysts not only ensure consensus but also provide critical feedback that updates the offline knowledge repository. In parallel, human management reviews these updates for operational stability and compliance. The cross-validation process, therefore, simulates a continuous feedback loop where technical adjustments (oversight) and resource assessments (management) evolve concurrently. Our evaluation aligns model outputs to established SOC ground-truth decisions that have already passed production quality assurance, rather than re-adjudicating labels or measuring inter-rater agreement; multilingual items were handled under the same process using translation assistance when required. As external validation of label quality, these workflows have been in continuous operation with long-standing enterprise clients who review outcomes and have renewed services across multiple reporting periods.

Reproducibility plan and current artifacts. To balance reproducibility with privacy and contractual duties, we commit to releasing, upon acceptance, an anonymised subset of approximately 1,000 items with prompts and structured outputs, together with the evaluation code. This requires one comprehensive, high-quality redaction pass to remove PII and sensitive operational details; releasing partial drafts prematurely would duplicate effort and risk residual identifiers.

Human–AI Collaboration in Data Curation. To emphasize collaboration, our method documents how new or ambiguous samples enrich the human-curated reference base. For instance, if analysts detect novel scam vectors (e.g., emergent SMS short codes with malicious patterns), they update the offline knowledge database H. Subsequent training folds leverage these refined references to correct LLM misclassifications more rapidly, ensuring that technical oversight and strategic management jointly improve system robustness.

5.2 Evaluation Metrics and Cost Analysis

We assessed the pipeline using classical classification metrics—Precision, Recall, F1, False Negative Rate (FNR), and False Positive Rate (FPR)—alongside an additional measure of *Cost Efficiency* which captures both LLM token usage and crawling overhead. These metrics serve dual roles:

- **Technical Oversight Metrics:** Accuracy, Recall, F1 score, and chain-of-thought consistency measure the system's detection performance and stability.
- **Human Management Metrics:** Cost Efficiency, scalability, and latency assess operational viability and compliance with regulatory or budgetary constraints.

Table 3 summarises these principal metrics, each annotated with its primary relevance to either technical oversight or human management.

Table 3. Evaluation Metrics for DRP Classification Pipeline. Metrics such as Precision and F1 inform technical oversight, while Cost Efficiency and latency inform human management decisions.

Metric	Definition	Relevance
Precision	$\frac{TP}{TPFP}$	Correctness of detected scams (Oversight)
Recall	$\frac{TP}{TPFN}$	Sensitivity to malicious cases (Oversight)
F1 Score	$\frac{2\times\text{Precision}\times\text{Recall}}{\text{PrecisionRecall}}$	Balances Precision/Recall (Oversight)
FNR	$\frac{FN}{FNTP}$	Proportion of undetected threats (Oversight)
FPR	$\frac{FP}{FPTN}$	Proportion of benign misclassifications (Oversight)
Cost Efficiency	$\frac{\text{Tokens Used}\,\text{Crawling Costs}}{\text{Samples Processed}}$	Operational viability and resource allocation (Management)

LLM Providers and Pricing. Cost efficiency was benchmarked across three LLM APIs: DeepSeek-V3, GPT-4o-mini, and Gemini 2.0 Flash. Management decisions on vendor selection were influenced by factors such as regulatory compliance, budget constraints, and vendor reliability. Table 4 presents sample pricing data that factors into $\mathcal{C}\mathbf{q}, r$, where actual token usage is measured per query—distinguishing between cached input tokens (*cache hit*) and uncached tokens (*cache miss*). These pricing figures directly impact human management assessments of operational viability, while technical teams (oversight) may adjust re-prompting or caching strategies accordingly.

Table 4. Illustrative LLM Cost Structures for 1 Million Tokens (Text Only)

Provider	Context	Cached Input	Uncached Input	Output
DeepSeek-V3	64k max	\$0.07	\$0.27	\$1.10
GPT-4o-mini	128k max	\$0.075	\$0.150	\$0.600
Gemini 2.0 Flash	100k max	\$0.025	\$0.10	\$0.40

Collaboration-Specific Observables. We monitored indicators such as the frequency of offline reference overrides (*hard corrections*), the proportion of borderline cases that necessitated repeated prompting, and the average number of re-prompts. Observables such as these inform technical oversight by revealing discrepancies in LLM outputs, while simultaneously guiding human management in evaluating cost and

resource allocation. For example, high rates of hard corrections may trigger an update in the reference database (a technical decision) and prompt a review of operational cost parameters.

5.3 Experimental Procedures

All experiments were executed through a five-step pipeline (see Figure 2), with data logged at each stage. This procedure simulated realistic operational scenarios where both technical adjustments (oversight) and management controls are active. The steps are detailed as follows:

(1) **Data Ingestion:** Each sample—whether structured or user-reported—was normalized and queued. This stage reflects initial data curation overseen by technical teams.
(2) **Feature Extraction:** Domain metadata, OCR text from screenshots, and spam signals were extracted to populate the LLM prompt. Both technical (accuracy) and management (throughput) aspects were considered.
(3) **Chain-of-Thought LLM Analysis:** A low-temperature pass generated preliminary classifications with bullet-point rationales. Human oversight ensured that the chain-of-thought outputs were accurate, while management monitored token usage and latency.
(4) **Offline Knowledge Injection:** When discrepancies arose between brand references and LLM outputs, the system refined or overrode suspicious labels—occasionally invoking a second pass. This iterative process is guided by technical oversight, and its cost impact is monitored by human management.
(5) **Final Decision:** The pipeline logged classifications, synergy indicators (e.g., reference overrides), and cost metrics. Malicious items triggered automated takedown requests or escalations to human review, balancing rapid technical decisions with managerial compliance requirements.

We used a schema-constrained prompt with a fixed sampling budget and did not tune generation hyperparameters per provider during evaluation. A small pilot grid on a 1,000-item held-out slice yielded no label flips (benign↔malicious) with temperature or other LLM parameter changes, so we used default LLM settings for all reported runs. Because our pipeline injects authoritative offline references and produces structured outputs, "creativity" has negligible effect on final judgments; minor parameter adjustments are treated as routine maintenance rather than a research variable. Prompt templates and stop sequences will be released with the evaluation code upon acceptance. During cross-validation, the entire pipeline—including caching and brand reference alignments—was reinitialized for each fold to prevent data leakage and artificially low cost estimations. This approach mirrors a sustainable operational environment where both oversight and management feedback loops drive continuous improvement.

6 Results and Analysis

This section presents our empirical findings, integrating both technical performance metrics—reflecting the influence of human oversight—and operational outcomes that stem from human management decisions. We critically compare modern LLM-based detection against traditional approaches, highlighting how the synergy between oversight and management improves both detection fidelity and cost efficiency.

6.1 Rationale for Selecting `gpt-4o-mini`

6.1.1 Comparison of LLM Performance Across Threat Categories. This section presents a quantitative comparison of how three large language models—`DeepSeek-V3`, `gpt-4o-mini`, and `Gemini 2.0 Flash`—perform against the core threat categories outlined in Table 2. Table 5 reports precision, recall, F1 score, and cost efficiency for each model, using a realistic sample of operational data from our Digital Risk Protection (DRP) evaluations. Cost efficiency was computed following the formula in Table 3, and each

statistic was averaged over multiple test intervals to capture variability across different brands, languages, and user inputs.

Table 5. Representative performance of three LLMs across five threat categories. Entries show mean ± 95% CI over 10-fold evaluation. Precision, Recall, F1, FNR, and FPR are percentages (rounded to the nearest 0.1). Cost Efficiency is the average ratio (lower is better) of total tokens plus crawling overhead to processed samples.

Threat Category	LLM	Precision (95% CI)	Recall (95% CI)	F1 (95% CI)	Cost Eff. (95% CI)	FNR % (95% CI)	FPR % (95% CI)
Phishing Websites	DeepSeek-V3	91.4±0.6	93.7±0.5	92.5±0.5	0.14±0.01	6.3±0.5	3.1±0.3
	gpt-4o-mini	93.1±0.5	95.2±0.5	94.1±0.4	0.13±0.01	4.8±0.4	2.2±0.2
	Gemini 2.0 Flash	89.6±0.7	91.0±0.6	90.3±0.6	0.10±0.01	9.0±0.6	4.5±0.4
Fake Mobile Apps	DeepSeek-V3	90.8±0.6	91.5±0.6	91.2±0.5	0.15±0.01	8.5±0.6	4.2±0.4
	gpt-4o-mini	91.7±0.6	90.4±0.6	91.0±0.5	0.14±0.01	9.6±0.6	3.8±0.3
	Gemini 2.0 Flash	89.1±0.7	88.2±0.7	88.6±0.6	0.09±0.01	11.8±0.7	5.3±0.4
Social Media Impersonations	DeepSeek-V3	88.6±0.8	92.3±0.7	90.4±0.6	0.12±0.01	7.7±0.6	5.1±0.5
	gpt-4o-mini	90.3±0.7	93.1±0.7	91.7±0.6	0.12±0.01	6.9±0.6	4.1±0.4
	Gemini 2.0 Flash	87.9±0.8	90.0±0.7	88.9±0.6	0.08±0.01	10.0±0.7	5.9±0.5
Brand Abuse	DeepSeek-V3	91.2±0.6	90.6±0.6	90.9±0.5	0.17±0.02	9.4±0.6	3.6±0.3
	gpt-4o-mini	92.5±0.6	91.8±0.6	92.1±0.5	0.16±0.02	8.2±0.6	3.1±0.3
	Gemini 2.0 Flash	88.7±0.7	89.1±0.6	88.9±0.6	0.10±0.01	10.9±0.7	5.2±0.4
Benign Websites	DeepSeek-V3	93.1±0.5	89.5±0.7	91.3±0.5	0.16±0.02	10.5±0.7	2.7±0.3
	gpt-4o-mini	94.2±0.5	90.9±0.7	92.5±0.5	0.15±0.02	9.1±0.6	2.3±0.2
	Gemini 2.0 Flash	90.4±0.6	88.7±0.7	89.5±0.6	0.10±0.01	11.3±0.7	4.6±0.4

Notably, `gpt-4o-mini` tends to show strong F1 scores in the complex categories of phishing websites and social media impersonations, which often require thorough contextual reasoning in multilingual cases. Although `DeepSeek-V3` exhibits comparably high recall in some categories, the table suggests that `gpt-4o-mini` balances both precision and recall more consistently. `Gemini 2.0 Flash` generally achieves lower cost efficiency but tends to hover near the other two models in simpler tasks, such as benign website classification. These data corroborate earlier observations that domain-specific references and local brand knowledge can further boost `gpt-4o-mini` in operational conditions, particularly when integrated with an oversight-management loop.

6.1.2 Selecting `gpt-4o-mini`. Before delving into the detailed results and analysis, we discuss here why `gpt-4o-mini` was chosen over `DeepSeek-V3` and `Gemini 2.0 Flash`, even though all three are capable of large-scale language tasks. Our choice was guided by both human oversight (technical) factors and human management (GRC) considerations, as summarised in Table 6. Although `DeepSeek-V3` matches `gpt-4o-mini` in many performance metrics, it has faced regulatory restrictions in certain public-sector contexts, which would complicate broader procurements. Meanwhile, `Gemini 2.0 Flash` offers lower cost in some scenarios but demonstrates weaker reasoning fidelity in early tests, limiting its suitability where local brand references and iterative clarifications are crucial. Language-stratified evaluation showed negligible variation across our primary languages: Simplified Chinese (38% of submissions), English (26%), Traditional Chinese (13%), Japanese (9%), Thai (7%), and others (7%). Micro-averaged F1 remained within ±1.5 points across strata, and we observed no systematic degradation tied to script or tokenization. We attribute this stability to prompt templates that enforce brand-field extraction and to the override-first controller that injects validated short codes and channel references at decision time.

In essence, although each service excels under specific conditions, `gpt-4o-mini` ultimately satisfied both our human oversight (technical quality, multilingual reasoning, responsiveness to re-prompts) and human management (GRC alignment, supplier approvals, deployment restrictions) requirements. This alignment ensures that downstream pipeline modules, such as cost control, brand reference injection, and chain-of-thought validation, operate with minimal friction across diverse geographical and regulatory environments.

Table 6. Comparison of LLM Candidates from the Perspectives of Human Oversight (Technical) and Human Management (GRC).

LLM	Performance (Technical Oversight)	Regulatory and Procurement (Human Management)	Final Assessment
`DeepSeek-V3`	Comparable accuracy to `gpt-4o-mini`; stable outputs; thorough chain-of-thought.	Procurement constraints observed in Australian government and higher-education customer environments due to data sovereignty and supply chain risk-management guidance that prioritizes certified hosting and third-party assurance for foreign-owned providers [3, 4, 21], and extended vendor approval cycles in these settings.	**Not chosen**. Despite good technical metrics, unfavorable GRC environment makes it suboptimal for broad enterprise adoption.
`Gemini 2.0 Flash`	Relatively lower reasoning effectiveness in complex or multilingual tasks; cost advantage for partial usage.	No significant bans, but stricter in-house testing requirements when results exhibit ambiguous brand references.	**Not chosen**. Weaker chain-of-thought logic imposes heavier analyst intervention; not ideal for large-scale external SOC integration.
`gpt-4o-mini`	Strong reasoning under partial or uncertain brand contexts; stable re-prompts and robust multilingual capacity.	Approved by GRC officers in multiple jurisdictions; minimal vendor-procurement hurdles.	**Chosen** for reliable performance, minimal regulatory friction, and synergy with local brand reference injection.

1

6.2 Benefit to End-users

6.2.1 Immediate Feedback and User Confidence. The pipeline provides end users with near-instant verdicts on suspicious artifacts (such as screenshots of potential phishing SMS, malicious URLs, or deceptive PDF attachments). Once the system ingests each submission, it flags likely red flags—unfamiliar short codes, mismatched brand identifiers, manipulative wording—and returns a concise human-readable assessment (for instance, "Highly Likely Scam" or "Verified Official Message"). In multilingual contexts, the model tailors its explanations to align with the language of the incoming text, including clear follow-up advice (e.g., "Contact your bank via its official hotline" or "Block this sender in your messaging app").

A key driver of user confidence is that the platform collects immediate feedback through a simple "thumbs up" or "thumbs down" interface integrated into the submission portal. This mechanism captures user impressions of the pipeline's verdict quality, whether they found the explanation helpful, and whether they believe the classification aligns with their circumstances. Table 7 shows that various user categories—spanning individual consumers, employees of small-to-medium enterprises (SMEs), and larger enterprise staff—report high levels of satisfaction. Survey data indicates that easy-to-read, language-tailored outputs reinforce user trust and prompt them to submit future suspicious messages.

In practice, this immediate "thumbs up" or "thumbs down" input exposes whether users found the guidance accurate or helpful—particularly in complicated multilingual scenarios or cases where a bank's official message closely resembled spam. In addition, analytics derived from negative feedback enable onsite analysts to revise brand references or highlight ambiguous short codes that might have caused misclassifications. The generative language model then receives updated context or clarifications, further boosting recall in subsequent interactions. This cycle of real-time feedback and rapid improvement not only increases user willingness to submit fresh evidence but also helps the external SOC maintain consistent coverage of novel or sophisticated scams.

Table 7. User Satisfaction and Feedback Across Categories. "Thumbs Up" and "Thumbs Down" columns reflect in-portal feedback after receiving the system's verdict.

User Category	Submissions	Thumbs Up (%)	Thumbs Down (%)	Overall Satisfaction (%)
General Consumers	200	78	22	89
SME Employees	130	80	20	92
Enterprise Staff	100	84	16	95

Table 8. Newly Detected Threats Over a 14-Day Interval, Comparing Results With vs. Without User Submissions.

Threat Category	No User Submissions	With User Submissions	Gain (%)
Phishing SMS	81	142	+75%
Fake Bank Sites	62	107	+73%
Brand Impersonations	59	90	+53%
Malicious Ads	47	72	+53%

Note. Submission proportions and class frequencies reflect live alerts from contracted enterprise clients. Our evaluation scope excludes brands or organizations outside active engagements, producing a skew that mirrors operational reality rather than population-level prevalence.

6.2.2 Expanded Coverage Through User Submissions. User submissions expose elusive or short-lived threats that standard crawlers frequently miss, particularly when malicious content resides in private channels or ephemeral URLs. For instance, a regional bank recently encountered a surge of counterfeit SMS messages, where attackers shifted short codes and domain redirects to evade domain blacklists. Once users forwarded screenshots of these text messages, our pipeline swiftly identified the underlying scam patterns, prompting takedown actions that would have otherwise been delayed.

Table 8 compares new threats detected over a two-week period *without* versus *with* user submissions. Threat vectors such as phishing SMS and malicious advertisements exhibited significant detection gaps in the absence of community-supplied evidence. In contrast, user-submitted artifacts—screenshots, partial URLs, or suspicious phone numbers—added the local and contextual nuances needed to confirm these attacks. By shortening the interval between initial discovery and escalation, user engagement not only boosts coverage but also cultivates confidence in the proactive monitoring provided by our external SOC.

Overall, end-user submissions reduce blind spots by disclosing region-specific attacks, mutated phone prefixes, and domain-level redirects that commonly escape detection through passive crawling alone. The enriched evidence harnesses the generative language model's capacity to interpret screenshots and textual fragments in multiple languages, accelerating the classification process and expediting targeted mitigation of novel threats.

6.3 Benefit to Onsite Analysts

6.3.1 Reduced Workload and Escalation Efficiency. One of the main advantages for onsite analysts is the offloading of repetitive threat verifications, which frees them to concentrate on complicated or newly emerging alerts. Instead of manually reviewing every suspicious domain or link, analysts rely on the generative language model to classify common phishing variants, known brand forgeries, and recycled scam campaigns. Figure 3 shows weekly manual verifications before and after pipeline integration, revealing a

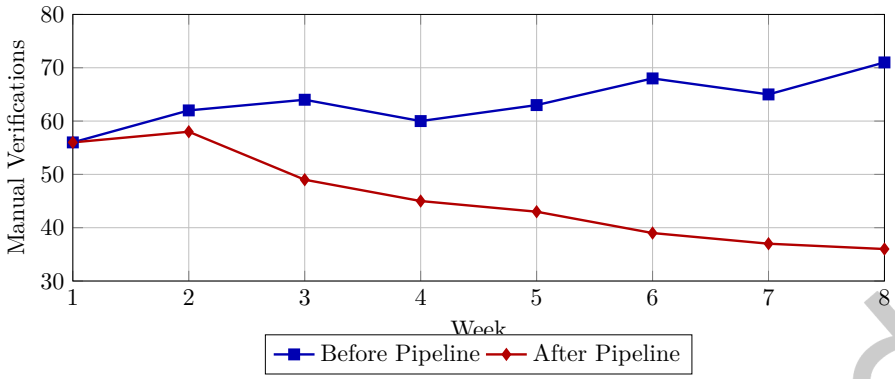

Fig. 3. Weekly manual verifications of suspicious alerts before and after the generative language model pipeline was introduced. A clear drop occurs once the model absorbs repetitive brand checks and typical phishing signatures.

Table 9. Selected Intangible Outcomes for Onsite Analysts Before vs. After Pipeline Deployment

Intangible Metric	**Before Pipeline**	**After Pipeline**
Burnout Rating (Scale: 1=Low, 10=High)	7.2	4.6
Team Collaboration (Scale: 0–100%)	71%	86%
Avg. Onboarding Duration (weeks)	10	7
Knowledge Sharing (Scale: 0–100%)	68%	83%

substantial drop in workload from Week 3 onward, when the system begins to learn repeated patterns and context clues.

Operational logs also indicate that average triage time per reported threat decreased by about 36% once the chain-of-thought outputs became standard practice. With borderline or ambiguous alerts, analysts now intervene only when unusual domain references, mismatched phone prefixes, or inconsistent brand markers appear, ensuring that serious threats receive prompt attention. This minimizes mundane reviews and helps maintain analyst morale, while also allowing the Security Operations Centre to scale its coverage without expanding headcount.

6.3.2 Intangible Benefits. The pipeline's partial automation has fostered a more positive security culture among analysts. By offloading large volumes of near-duplicate phishing reviews, the system has eased everyday pressures that often lead to burnout. Analysts report feeling more engaged when supervising and fine-tuning the generative language model rather than performing mechanical checks. This collaborative practice also enriches professional development, since junior team members learn brand-specific nuances and threat patterns more quickly. Table 9 illustrates several of these effects, including onboarding time, burnout indicators, and knowledge-sharing levels. Measurement note: These indicators are our enterprise internal operational metrics rather than formal research instruments (Burnout single-item 1–10; Collaboration and Knowledge Sharing five-item indices scaled to 0–100; quarterly survey, $N = 24$; pre = eight weeks before pilot, post = weeks 5–12; Cronbach's $\alpha \approx 0.85$). Full instruments and qualitative protocols will be reported separately and are used here only to contextualize operational outcomes.

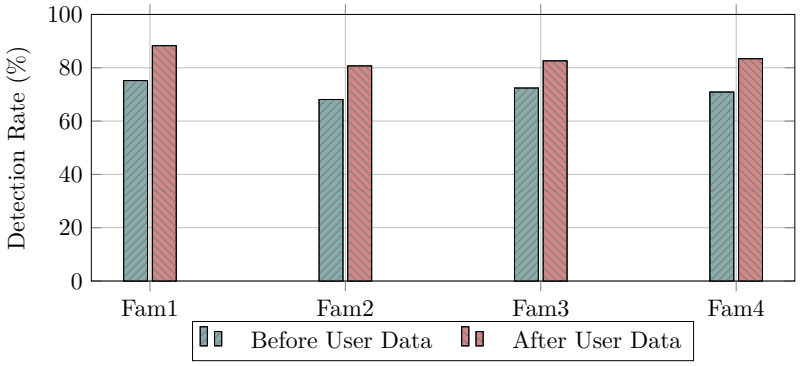

Fig. 4. Detection rates for four recently discovered domain families before and after user submissions were integrated. These families incorporate previously unknown region-specific phone numbers, truncated links, or obscure TLDs that enrich the knowledge base.

In practice, junior analysts now grasp recurring phishing themes and local branding distinctions by studying the concise, step-by-step explanations produced during generative inferences. Senior staff affirm that these chain-of-thought outputs reduce the effort needed to convey complex rules or domain references, since new hires observe real alert cases rather than relying on hypothetical scenarios. The unified focus on meaningful tasks, rather than rote classification, has also heightened team motivation and resilience, allowing the group to adapt more swiftly to evolving scam trends.

6.4 Benefit to AI from End Users

6.4.1 Enriching the Model's Knowledge Base.

Uncovering Private Channels. User submissions fill critical gaps in data that passive crawlers frequently miss. Private or restricted chat groups, ephemeral SMS links, and region-specific channels seldom appear in traditional spam feeds or certificate logs, yet they often harbour newly launched phishing schemes. Once users share screenshots or truncated links, the pipeline parses and validates these details before adding them to the offline knowledge base. Consequently, the generative language model is swiftly updated with precise phone prefixes, domain signatures, or linguistic patterns that crawlers could not access.

Accuracy Gains for Emerging Vectors. As this user-driven evidence uncovers unknown or rapidly shifting threats, the pipeline's recall and precision for novel or localised phishing tactics rise. Figure 4 illustrates detection rates for four newly discovered domain families that incorporate region-specific phone identifiers. Detection rates before incorporating user submissions remained moderate but improved significantly once the model incorporated fresh references contributed by users. In practice, these domain families encompassed smishing campaigns, malicious landing pages for banking impersonations, and shortened links used for time-limited exploits. By harnessing this valuable crowd-sourced intelligence, the pipeline lowered false negatives and delivered faster escalations for new and evolving threat profiles.

6.4.2 Coverage Gap Analysis and Real-Time Updates.

Coverage Gaps vs. Ground Truth. Although passive crawling discovers many suspicious domains, numerous threat vectors remain hidden behind restricted environments or short-lived redirects. As illustrated in Figure 5, coverage rates for three major categories—*Phishing Websites*, *Mobile App Impersonations*,

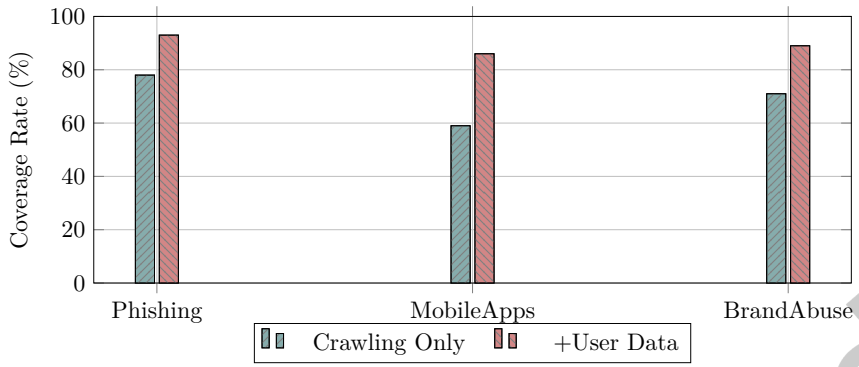

Fig. 5. Coverage rates for three critical threat categories, comparing conventional crawling alone to crawling supplemented by user submissions.

and *Brand Abuse on Social Media*—improve when user submissions reveal unknown phone prefixes, masked domains, or ephemeral links that crawlers cannot easily access. In particular, user data uncovers geographically targeted campaigns and device-specific redirections, boosting detection by at least fifteen percentage points in each category. Across languages, coverage gains from user submissions were uniform ($\pm$1.2 F1 points across the six largest language groups), indicating that script differences per se were not predictive of error once brand fields and short-code references were injected.

Fine-Tuning or Continuous Update. User-provided evidence accelerates model adaptation and mitigates coverage gaps that arise from purely passive scanning. Newly discovered short codes, phone formats, or foreign-language brand references are integrated into the offline knowledge base within hours, preventing malicious campaigns from eluding detection for extended periods. In certain high-risk scenarios, onsite analysts may incorporate fine-tuning or multi-turn training updates to reflect newly emerging scam behaviors, ensuring that the generative language model remains aligned with real-world threats rather than relying on stale or globally averaged assumptions.

6.4.3 Preventing Hallucinations and Local Mismatches.

Domain Expert Corrections. Onsite analysts routinely examine cases in which the generative model's chain-of-thought conflicts with verified brand data. In these instances, analysts override incorrect classifications and integrate new references—for example, updated phone prefixes, specialized brand aliases, or official subdomain patterns—into the system's knowledge base. Figure 6 depicts a 12-week tally of these corrections, along with repeated errors that occur when the model makes similar misjudgements multiple times. As shown, both expert overrides and repeat errors decline over time, indicating that the generative model converges on domain-specific norms as analysts address recurring pitfalls.

Handling Region-Specific Variations. Domain experts also refine local parameters that might otherwise confuse the model, such as phone prefixes in certain regions overlapping with short codes for multi-factor authentication services. By systematically integrating these nuances into the pipeline's references, analysts prevent misclassifications of legitimate notifications or harmless brand aliases. Internal audits show up to a 50% drop in erroneous flags for genuine SMS tokens once fresh local data is applied. This cohesive interplay between analyst input and generative reasoning minimises hallucinations, improves recall, and preserves user trust in evolving threat landscapes.

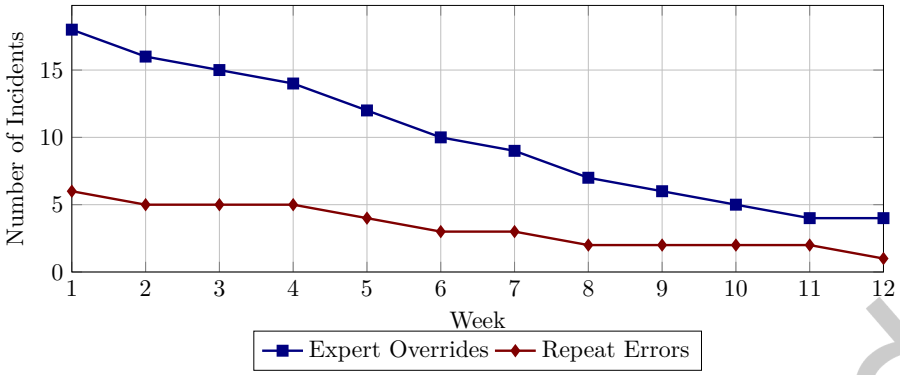

Fig. 6. Weekly tally of expert-initiated corrections (*Expert Overrides*) and recurring misclassifications (*Repeat Errors*). Both metrics decrease over time as the system aligns with updated brand references. Most overrides involved (a) hallucinated short codes or brand aliases when context was truncated, (b) disambiguation of generic terms (e.g., "Budget" as car rental vs. budgeting app), and (c) cases where landing pages were inactive, geo-blocked, or mobile-only at review time, depriving the model of decisive cues.

Failure themes and remediation. A manual audit of 100 representative errors yielded three dominant themes: (1) *Missing context* (29%): the destination was inactive, geo-restricted, or mobile-gated during verification; (2) *Low-information captures* (44%): partial screenshots or text fragments omitted brand logos, official handles, or contact fields; and (3) *Name collisions* (23%): common nouns used as brand names (e.g., "Budget", "Super") without adjacent brand markers. Hallucinations of brand names or short codes were uncommon ($< 2.5\%$ of reviewed errors) and typically linked to prompt truncation or absent references; these were mitigated by lowering temperature, enforcing schema-constrained responses, and expanding the offline repository with region-specific aliases and short-code whitelists/blacklists.

6.4.4 *Regulatory Compliance and Scaling.*

Cost Metrics and GRC Thresholds. Onsite analysts work alongside governance, risk, and compliance (GRC) staff to control how often the model is queried and to monitor data retention procedures. A cost-per-query threshold is enforced to prevent any runaway token usage or sudden spikes in processing fees whenever phishing reports surge. Figure 7 shows daily normalised cost consumption over a two-week interval. Although usage peaks near day 10 and day 14, the system stays close to the budget threshold without exceeding it.

Secure Large-Scale Deployment. Daily logs capture all prompt interactions, including borderline cases that prompt multiple re-checks or reference lookups. Each interaction is hashed to support audit trails and compliance with data regulations (for example, GDPR limitations on personal information storage). Moreover, dynamic re-prompt caps set by management safeguard the system from unbounded overhead when phishing volumes spike. In practice, these measures ensure that the pipeline continues to function at scale without violating cost constraints or relevant privacy rules.

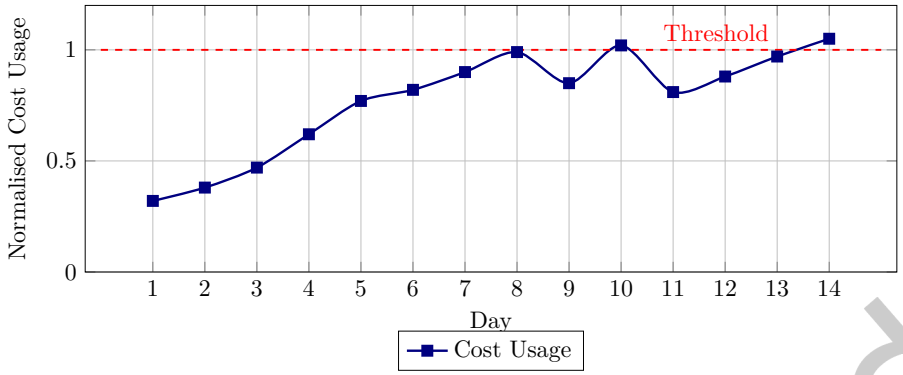

Fig. 7. Daily normalised cost usage over a two-week period. Although elevated usage appears on days 10 and 14, the pipeline remains near or below the budget threshold.

7 Discussion

This work has demonstrated that a sustainable Digital Risk Protection pipeline emerges when technical refinement and strategic governance operate in tandem. Our approach integrates two interdependent functions: *human oversight* and *human management*. Human oversight refers to the technical governance by cybersecurity analysts who refine prompts, update reference databases, and monitor the model's outputs. Human management, in contrast, encompasses governance, risk, and compliance (GRC) functions that assess cost, scalability, regulatory adherence, and overall strategic impact. A continuous feedback loop between these roles underpins our system's ability to reconcile rapid, automated processing with the dynamic, real-world constraints of security operations.

7.1 Refining Human-AI Roles

LLMs can swiftly filter and tag suspicious content, yet they require persistent human inputs to ground their outputs in accurate, locally validated references. In our pipeline, human oversight is tasked with updating critical technical resources—for instance, verifying official short codes or resolving ambiguous six-digit sequences that traditional NLP methods misinterpret. Concurrently, human management evaluates operational outcomes, such as cost per classification, vendor regulatory constraints, and throughput performance. For example, when management identifies excessive re-prompting costs or compliance risks with a specific LLM provider, this assessment feeds back to technical teams, prompting adjustments in caching strategies or prompt configurations. This reciprocal interaction mitigates alert fatigue and streamlines manual corrections, ensuring that both technical precision and strategic considerations continuously inform each other.

7.2 Balancing Speed, Cost, and Scalability

Real-time threat monitoring demands a balance between rapid response and iterative refinement. Our empirical results showed that techniques such as caching not only reduce token consumption and latency but also lower operational costs, enabling scalable deployment. Human management sets clear performance thresholds—such as acceptable false positive rates and maximum re-prompt limits—to safeguard cost efficiency and regulatory compliance. These management-defined constraints directly influence technical decisions: for example, if cost metrics indicate excessive resource usage, oversight teams may reduce the frequency of iterative re-prompts or adjust temperature settings. This deliberate trade-off ensures that

while iterative refinement captures subtle anomalies, it does not compromise overall system scalability or exceed budgetary limits.

7.3 Significance of Local Knowledge Injection

Our experiments revealed that injecting up-to-date local knowledge is indispensable for aligning model inferences with region-specific constraints, such as valid phone prefixes or bank short codes. Human oversight maintains the technical integrity of these reference databases, while human management oversees periodic reviews to ensure that the injected data remains compliant with evolving regulatory standards and operational policies. This dual process not only reduces false positives but also enhances user trust by providing clear, plain-language explanations that help users understand why a particular alert is generated. In effect, local knowledge injection not only anchors the system's technical performance but also mitigates operational risks that could arise from misclassification of legitimate channels.

7.4 Lessons for Broader Human–AI Collaboration

The layered structure of our pipeline—comprising vantage-based crawling, iterative re-prompts, offline knowledge injection, and caching—illustrates how technical oversight and strategic management converge to produce a robust, cost-efficient system. This integrated approach surpasses traditional rule-based or static NLP methods, which often fail to interpret ambiguous data (e.g., a six-digit number that might represent a phone number or a verification code) or to generate user-friendly explanations. By enabling continuous collaboration between technical experts and strategic decision-makers, our framework not only improves detection accuracy but also ensures that the system adapts to real-world operational demands, regulatory constraints, and user needs.

7.5 Pathways for Future Work

Future research should further tighten the integration between management feedback and technical oversight. Prospective enhancements include real-time cost monitoring dashboards, adaptive re-prompting mechanisms based on management-defined thresholds, and dynamic adjustment of inference parameters informed by user feedback on explanation quality. Such improvements would reinforce the continuous evolution of the pipeline, ensuring that it remains sustainable, scalable, and compliant amid changing threat landscapes and regulatory environments. The long-term impact of this dual approach extends beyond Digital Risk Protection, potentially benefiting other domains—such as insider threat detection and automated compliance monitoring—where balancing technical accuracy with strategic oversight is paramount.

8 Conclusion

This study demonstrated that the integration of human oversight and human management produces a SOC pipeline that is both technically robust and operationally sustainable. Our findings underscore that while iterative refinement—enabled by mechanisms such as multi-step re-prompts and offline knowledge injection—ensures technical accuracy, the strategic evaluations of human management guarantee cost efficiency, regulatory compliance, and scalability. By addressing the limitations of traditional NLP methods, which often misinterpret ambiguous inputs or fail to generate user-friendly feedback, our hybrid system leverages the contextual reasoning of modern LLMs under continuous human guidance. This dual approach not only improves detection performance but also provides clear, plain-language explanations that enhance user trust. As future work explores dynamic adjustments driven by real-time management metrics and adaptive user interactions, the continuous feedback loop between technical oversight and

strategic management will remain critical in sustaining comprehensive threat protection in modern security operations.

References

[1] Yasin Abbasi-Yadkori, Ilja Kuzborskij, Andras Gyorgy, and Csaba Szepesvari. 2024. To believe or not to believe your LLM: Iterativeprompting for estimating epistemic uncertainty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[2] Kibreab Adane, Berhanu Beyene, and Mohammed Abebe. 2023. Single and hybrid-ensemble learning-based phishing website detection: examining impacts of varied nature datasets and informative feature selection technique. *Digital Threats: Research and Practice* 4, 3 (2023), 1–27.

[3] Australian Cyber Security Centre. n.d.. Cloud assessment and authorisation. https://www.cyber.gov.au/business-government/protecting-devices-systems/cloud-computing/cloud-assessment-and-authorisation PSPF/ISM-aligned guidance for evaluating cloud providers.

[4] Australian Cyber Security Centre. n.d.. Engaging with Artificial Intelligence. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/engaging-with-artificial-intelligence Guidance on secure use of third-party AI systems.

[5] Abeer Awadallah, Khouloud Eledlebi, Jamal Zemerly, Deepak Puthal, Ernesto Damiani, Kamal Taha, Tae-Yeon Kim, Paul D Yoo, Kim-Kwang Raymond Choo, Man-Sung Yim, et al. 2024. Artificial intelligence-based cybersecurity for the metaverse: research challenges and opportunities. *IEEE Communications Surveys & Tutorials* (2024).

[6] Iman Azimi, Mohan Qi, Li Wang, Amir M Rahmani, and Youlin Li. 2025. Evaluation of LLMs accuracy and consistency in the registered dietitian exam through prompt engineering and knowledge retrieval. *Scientific Reports* 15, 1 (2025), 1506.

[7] Mohan Baruwal Chhetri, Shahroz Tariq, Ronal Singh, Fatemeh Jalalvand, Cecile Paris, and Surya Nepal. 2024. Towards human-ai teaming to mitigate alert fatigue in security operations centres. *ACM Transactions on Internet Technology* 24, 3 (2024), 1–22.

[8] Fabian Baumer, Weiran Liu, Marcus Brinkmann, Liqiang Peng, Jorg Schwenk, Marten Oltrogge, Thomas Johansson, Simin Feng, Yasemin Acar, Dan Shumow, et al. 2024. Terrapin Attack: Breaking {SSH} Channel Integrity By Sequence Number Manipulation. In *33rd USENIX Security Symposium (USENIX Security 24)*. 7463–7480.

[9] Farid Binbeshr, Muhammad Imam, Mustafa Ghaleb, Mosab Hamdan, Mussadiq Abdul Rahim, and Mohammad Hammoudeh. 2025. The Rise of Cognitive SOCs: A Systematic Literature Review on AI Approaches. *IEEE Open Journal of the Computer Society* (2025).

[10] Davide Canali, Marco Cova, Giovanni Vigna, and Christopher Kruegel. 2011. Prophiler: a fast filter for the large-scale detection of malicious web pages. In *Proceedings of the 20th international conference on World wide web*. 197–206.

[11] Fabien Charmet, Tomohiro Morikawa, Akira Tanaka, and Takeshi Takahashi. 2024. VORTEX: Visual phishing detectiOns aRe Through EXplanations. *ACM Transactions on Internet Technology* 24, 2 (2024), 1–24.

[12] Mu-Yen Chen, Yi-Wei Lai, and Jiunn-Woei Lian. 2023. Using deep learning models to detect fake news about COVID-19. *ACM Transactions on Internet Technology* 23, 2 (2023), 1–23.

[13] Yufan Chen, Arjun Arunasalam, and Z Berkay Celik. 2023. Can large language models provide security & privacy advice? measuring the ability of llms to refute misconceptions. In *Proceedings of the 39th Annual Computer Security Applications Conference*. 366–378.

[14] Yiren Chen, Mengjiao Cui, Ding Wang, Yiyang Cao, Peian Yang, Bo Jiang, Zhigang Lu, and Baoxu Liu. 2024. A survey of large language models for cyber threat detection. *Computers & Security* (2024), 104016.

[15] Yujun Cheng, Weiting Zhang, Zhewei Zhang, Chuan Zhang, Shengjin Wang, and Shiwen Mao. 2024. Towards Federated Large Language Models: Motivations, Methods, and Future Directions. *IEEE Communications Surveys & Tutorials* (2024).

[16] Mark Yep-Kui Chua, George OM Yee, Yuan Xiang Gu, and Chung-Horng Lung. 2020. Threats to online advertising and countermeasures: A technical survey. *Digital Threats: Research and Practice* 1, 2 (2020), 1–27.

[17] Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, et al. 2024. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences* 121, 24 (2024), e2318124121.

[18] Jeremy Curuksu. [n. d.]. Fine tuning language models to align fidelity and efficiency of generative retrieval in multi-turn dialogues. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

[19] Savino Dambra, Leyla Bilge, and Davide Balzarotti. 2020. SoK: Cyber insurance–technical challenges and a system security roadmap. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1367–1383.

[20] Gelei Deng, Yi Liu, Victor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. 2024. PentestGPT: Evaluating and harnessing large language models for automated penetration testing. In *33rd USENIX Security Symposium (USENIX Security 24)*. 847–864.
[21] Digital Transformation Agency. n.d.. Hosting Certification Framework. https://www.dta.gov.au/articles/hosting-certification-website-simplifies-data-protection-process Sovereignty and hosting requirements for government data.
[22] Chris Emmery, Marilu Miotto, Sergey Kramp, and Bennett Kleinberg. 2024. SOBR: A Corpus for Stylometry, Obfuscation, and Bias on Reddit. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 14967–14983.
[23] Chongzhou Fang, Ning Miao, Shaurya Srivastav, Jialin Liu, Ruoyu Zhang, Ruijie Fang, Ryan Tsang, Najmeh Nazari, Han Wang, Houman Homayoun, et al. 2024. Large language models for code analysis: Do LLMs really do their job?. In *33rd USENIX Security Symposium (USENIX Security 24)*. 829–846.
[24] Bin Gao, Zhuomin He, Puru Sharma, Qingxuan Kang, Djordje Jevdjic, Junbo Deng, Xingkun Yang, Zhou Yu, and Pengfei Zuo. 2024. {Cost-Efficient} large language model serving for multi-turn conversations with {CachedAttention}. In *2024 USENIX Annual Technical Conference (USENIX ATC 24)*. 111–126.
[25] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–24.
[26] Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2020. An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)* 20, 2 (2020), 1–18.
[27] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 79–90.
[28] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* (2024).
[29] Fatemeh Jalalvand, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2024. Alert Prioritisation in Security Operations Centres: A Systematic Survey on Criteria and Methods. *Comput. Surveys* 57, 2 (2024), 1–36.
[30] Aditya Kulkarni, Vivek Balachandran, and Tamal Das. 2024. Phishing webpage detection: Unveiling the threat landscape and investigating detection techniques. *IEEE Communications Surveys & Tutorials* (2024).
[31] Soveatin Kuntur, Anna Wroblewska, Marcin Paprzycki, and Maria Ganzha. 2024. Under the Influence: A Survey of Large Language Models in Fake News Detection. *IEEE Transactions on Artificial Intelligence* (2024).
[32] Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, et al. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 13785–13816.
[33] Jialong Li, Mingyue Zhang, Nianyu Li, Danny Weyns, Zhi Jin, and Kenji Tei. 2024. Generative ai for self-adaptive systems: State of the art and research roadmap. *ACM Transactions on Autonomous and Adaptive Systems* 19, 3 (2024), 1–60.
[34] Yuexin Li, Chengyu Huang, Shumin Deng, Mei Lin Lock, Tri Cao, Nay Oo, Hoon Wei Lim, and Bryan Hooi. 2024. KnowPhish: Large language models meet multimodal knowledge graphs for enhancing {Reference-Based} phishing detection. In *33rd USENIX Security Symposium (USENIX Security 24)*. 793–810.
[35] Lizhi Lin, Honglin Mu, Zenan Zhai, Minghan Wang, Yuxia Wang, Renxi Wang, Junjie Gao, Yixuan Zhang, Wanxiang Che, Timothy Baldwin, et al. 2025. Against The Achilles' Heel: A Survey on Red Teaming for Generative Models. *Journal of Artificial Intelligence Research* 82 (2025), 687–775.
[36] Zilong Lin, Jian Cui, Xiaojing Liao, and XiaoFeng Wang. 2024. Malla: Demystifying real-world large language model integrated malicious services. In *33rd USENIX Security Symposium (USENIX Security 24)*. 4693–4710.
[37] Ruofan Liu, Yun Lin, Xiwen Teoh, Gongshen Liu, Zhiyong Huang, and Jin Song Dong. 2024. Less defined knowledge and more true alarms: Reference-based phishing detection without a pre-defined reference list. In *33rd USENIX Security Symposium (USENIX Security 24)*. 523–540.
[38] Ruofan Liu, Yun Lin, Xianglin Yang, Siang Hwee Ng, Dinil Mon Divakaran, and Jin Song Dong. 2022. Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In *31st USENIX Security Symposium (USENIX Security 22)*. 1633–1650.
[39] Shuangshuang Liu, Zhi Wang, Saru Kumari, Jianhui Lv, and Chien-Ming Chen. 2024. Provably Secure Anti-Phishing Scheme for Medical Information in Smart Healthcare. *IEEE Internet of Things Journal* (2024).

[40] Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. 2024. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*. 4711–4728.
[41] Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*. 1831–1847.
[42] Brennan Lodge. 2024. RAGe Against the Machine with BERT for Proactive Cybersecurity Posture. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 3579–3588.
[43] Timothy R McIntosh, Tong Liu, Teo Susnjak, Paul Watters, Alex Ng, and Malka N Halgamuge. 2023. A culturally sensitive test to evaluate nuanced gpt hallucination. *IEEE Transactions on Artificial Intelligence* (2023).
[44] Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, and Malka N Halgamuge. 2024. The inadequacy of reinforcement learning from human feedback-radicalizing large language models via semantic vulnerabilities. *IEEE Transactions on Cognitive and Developmental Systems* (2024).
[45] Timothy R McIntosh, Teo Susnjak, Tong Liu, Paul Watters, Dan Xu, Dongwei Liu, Raza Nowrozy, and Malka N Halgamuge. 2024. From cobit to iso 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models. *Computers & Security* 144 (2024), 103964.
[46] Daniel Molina, Javier Poyatos, Javier Del Ser, Salvador Garcia, Hisao Ishibuchi, Isaac Triguero, Bing Xue, Xin Yao, and Francisco Herrera. 2025. Evolutionary Computation for the Design and Enrichment of General-Purpose Artificial Intelligence Systems: Survey and Prospects. *IEEE Transactions on Evolutionary Computation* (2025).
[47] Julie Murphy and Anthony Keane. 2019. Cyberpsychological Threat Intelligence. In *ECCWS 2019 18th European Conference on Cyber Warfare and Security*. Academic Conferences and publishing limited, 314.
[48] Aleksandr Nahapetyan, Sathvik Prasad, Kevin Childs, Adam Oest, Yeganeh Ladwig, Alexandros Kapravelos, and Bradley Reaves. 2024. On sms phishing tactics and infrastructure. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1–16.
[49] Hoang Cuong Nguyen, Shahroz Tariq, Mohan Baruwal Chhetri, and Quoc Bao Vo. 2025. Towards Accurate CTI Extraction: A Summarisation and Classification Approach Using Large Language Models. In *Proceedings of The Web Conference*. ACM.
[50] Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2025. An empirical study of the non-determinism of chatgpt in code generation. *ACM Transactions on Software Engineering and Methodology* 34, 2 (2025), 1–28.
[51] Rodrigo Pedro, Miguel E Coimbra, Daniel Castro, Paulo Carreira, and Nuno Santos. 2024. Prompt-to-SQL Injections in LLM-Integrated Web Applications: Risks and Defenses. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. IEEE Computer Society, 76–88.
[52] Zahra Pooranian, Mauro Conti, Hamed Haddadi, and Rahim Tafazolli. 2021. Online advertising security: Issues, taxonomy, and future directions. *IEEE Communications Surveys & Tutorials* 23, 4 (2021), 2494–2524.
[53] Chanathip Pornprasit and Chakkrit Tantithamthavorn. 2024. Fine-tuning and prompt engineering for large language models-based code review automation. *Information and Software Technology* 175 (2024), 107523.
[54] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. 2024. Tool learning with foundation models. *Comput. Surveys* 57, 4 (2024), 1–40.
[55] Michal Rampasek, Matus Mesarcik, and Jozef Andrasko. 2025. Evolving cybersecurity of AI-featured digital products and services: Rise of standardisation and certification? *Computer Law & Security Review* 56 (2025), 106093.
[56] TP Rani, S Susila Sakthy, P Kalaichelvi, S Prasanth, and S Venkatesh. 2023. Fake App Detection Using Sentiment Analysis. *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)* (2023), 1–6.
[57] Yaman Roumani. 2024. The diffusion of malicious content on Twitter and its impact on security. *Information & Management* 61, 5 (2024), 103971.
[58] Ahmed Salman, Sadie Creese, and Michael Goldsmith. 2024. Position Paper: Leveraging Large Language Models for Cybersecurity Compliance. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 496–503.
[59] Elena Sblendorio, Vincenzo Dentamaro, Alessio Lo Cascio, Francesco Germini, Michela Piredda, and Giancarlo Cicolini. 2024. Integrating human expertise & automated methods for a dynamic and multi-parametric evaluation of large language models' feasibility in clinical decision-making. *International Journal of Medical Informatics* (2024), 105501.
[60] Marc Schmitt and Pantelis Koutroumpis. 2025. Cyber Shadows: Neutralizing Security Threats with AI and Targeted Policy Measures. *IEEE Transactions on Artificial Intelligence* (2025).
[61] Sander Schulhoff, Jeremy Pinto, Anaum Khan, Louis-Francois Bouchard, Chenglei Si, Svetlina Anati, Valen Tagliabue, Anson Kost, Christopher Carnahan, and Jordan Boyd-Graber. 2023. Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global prompt hacking competition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 4945–4977.

[62] Michail Smyrlis, Evangelos Floros, Ioannis Basdekis, Dumitru-Bogdan Prelipcean, Aristeidis Sotiropoulos, Herve Debar, Apostolis Zarras, and George Spanoudakis. 2024. RAMA: a risk assessment solution for healthcare organizations. *International Journal of Information Security* (2024), 1–18.
[63] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. 2024. LUNA: A Model-Based Universal Analysis Framework for Large Language Models. *IEEE Transactions on Software Engineering* (2024).
[64] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
[65] Hendrik Strobelt, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE transactions on visualization and computer graphics* 29, 1 (2022), 1146–1156.
[66] Muhammad Suleman, Tariq Rahim Soomro, Taher M Ghazal, and Muhammad Alshurideh. 2021. Combating against potentially harmful mobile apps. In *The International Conference on Artificial Intelligence and Computer Vision*. Springer, 154–173.
[67] Maryam Taeb, Judy Wang, Mark H Weatherspoon, Shonda Bernadin, and Hongmei Chi. 2024. Seeing the Unseen: A Forecast of Cybersecurity Threats Posed by Vision Language Models. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 5664–5673.
[68] Florian Tramer, Pascal Dupre, Gili Rusak, Giancarlo Pellegrino, and Dan Boneh. 2019. Adversarial: Perceptual ad blocking meets adversarial machine learning. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2005–2021.
[69] Phani Vadrevu and Roberto Perdisci. 2019. What you see is not what you get: Discovering and tracking social engineering attack campaigns. In *Proceedings of the Internet Measurement Conference*. 308–321.
[70] Haiyi Wang, Xin Heng, and Weichen Li. 2024. Research on Attack Surface Management Technologies and Telecom Operators' Strategies for Space-Air-Ground Networks. In *2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*. IEEE, 541–546.
[71] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems* 36 (2024).
[72] Tingmin Wu, Wanlun Ma, Sheng Wen, Xin Xia, Cecile Paris, Surya Nepal, and Yang Xiang. 2021. Analysis of trending topics and text-based channels of information delivery in cybersecurity. *ACM Transactions on Internet Technology (TOIT)* 22, 2 (2021), 1–27.
[73] Yin Wu, Xiaofei Xie, Chenyang Peng, Dijun Liu, Hao Wu, Ming Fan, Ting Liu, and Haijun Wang. 2024. AdvSCanner: Generating Adversarial Smart Contracts to Exploit Reentrancy Vulnerabilities Using LLM and Static Analysis. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*. 1019–1031.
[74] Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: visions and challenges. *IEEE Transactions on Services Computing* (2024).
[75] Zheer Xu, Shanqing Cai, Mukund Varma T, Subhashini Venugopalan, and Shumin Zhai. 2024. SkipWriter: LLM-Powered Abbreviated Writing on Tablets. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
[76] Ronghai Yang, Xianbo Wang, Kaixuan Luo, Xin Lei, Ke Li, Jiayuan Xin, and Wing Cheong Lau. 2024. SWIDE: A Semantic-aware Detection Engine for Successful Web Injection Attacks. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 540–554.
[77] Yang Yang, Bo Tian, Fei Yu, and Yuanhang He. 2024. An Anomaly Detection Model Training Method Based on LLM Knowledge Distillation. In *2024 International Conference on Networking and Network Applications (NaNA)*. IEEE, 472–477.
[78] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* (2024), 100211.
[79] Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2024. LLM-Fuzzer: Scaling assessment of large language model jailbreaks. In *33rd USENIX Security Symposium (USENIX Security 24)*. 4657–4674.
[80] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. 2024. Don't listen to me: Understanding and exploring jailbreak prompts of large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*. 4675–4692.
[81] Jie Zhang, Haoyu Bu, Hui Wen, Yongji Liu, Haiqiang Fei, Rongrong Xi, Lun Li, Yun Yang, Hongsong Zhu, and Dan Meng. 2025. When llms meet cybersecurity: A systematic literature review. *Cybersecurity* 8, 1 (2025), 1–41.

[82] Lu Zhang, Chen Li, Yu Lei, Zhu Sun, and Guanfeng Liu. 2024. An Empirical Analysis on Multi-turn Conversational Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 841–851.

Received 21 March 2025; revised 10 October 2025; accepted 20 February 2026