Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.085454

ARTICLE

Intelligent Risk Prioritization for Phishing Mitigation: A Human-Factor-Aware Framework for Healthcare SOCs

Chia-Nan Wang[1], **Tsei-Hsuan Chen**[2,*], **Syuan-Yun Wang**[3,*] **and Chung-Nan Cheng**[2]

[1]Department of Industrial Engineering and Management, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan

[2]Department of Medical Information, Kaohsiung Armed Forces General Hospital, Kaohsiung, Taiwan

[3]Department of Mechanical Engineering, National Cheng Kung University, Tainan, Taiwan

*Corresponding Authors: Tsei-Hsuan Chen. Email: i113143106@nkust.edu.tw; Syuan-Yun Wang. Email: e14124145@gs.ncku.edu.tw

Received: 11 May 2026; Accepted: 13 August 2026; Published: 28 September 2026

ABSTRACT: In email-centric healthcare environments, social engineering attacks increasingly exploit human psychology, organizational trust relationships, and persuasive communication strategies to bypass conventional cybersecurity defenses. While existing email security controls are effective at blocking many malicious messages, they remain vulnerable to whitelist-failure scenarios in which compromised or seemingly legitimate communications evade detection and reach end users. Under limited analyst capacity and increasing alert volumes, the operational challenge is no longer solely identifying phishing emails but determining which socially engineered communications should be reviewed first. To address this problem, this study proposes a governance-oriented human-factor risk prioritization framework that operates as a post-detection decision-support layer rather than a traditional phishing-filtering mechanism. The proposed P×F framework integrates persuasion tactics (P) and persona-based message-framing factors (F) to model social engineering influence patterns and transform psychological manipulation cues into interpretable risk representations for Top-K (top-ranked K alerts) alert prioritization. Using Bayesian smoothing and Logistic Regression, the framework converts semantic indicators into auditable risk scores that support analyst attention allocation under constrained Security Operations Center (SOC) resources. The framework was evaluated using Enron and Nazario email corpora, Large Language Model (LLM)-generated phishing scenarios, and a real-world Hospital A proof-of-concept dataset. Experimental results achieved mean Receiver Operating Characteristic Area Under the Curve (ROC-AUC) values of 0.9829 under subject-only conditions and 0.9871 using full-text content, while maintaining robust performance under distribution shift with performance degradation below 0.06 AUC. In operational evaluation, the Top-100 prioritization mechanism achieved 0.95 precision across 10,463 real SOC emails. Compared with Bidirectional Encoder Representations from Transformers (BERT), which exhibited recall collapse (0.03) under limited-context conditions, the proposed framework demonstrated greater stability, interpretability, and operational suitability for alert triage. Unlike conventional phishing detection approaches that primarily emphasize content classification, the proposed framework models psychological manipulation mechanisms derived from persuasion theory and human-factor literature. It operationalizes these mechanisms as interpretable P×F human-factor indicators and validates the resulting governance-oriented risk prioritization framework using real hospital phishing emails.

KEYWORDS: SOC alert prioritization; human-factor risk modeling; phishing mitigation; healthcare security operations; risk prioritization

1 Introduction

Phishing remains one of the most prevalent and damaging forms of social engineering, where attackers impersonate trusted entities to induce harmful actions such as credential compromise, malware infection, or data leakage [1,2]. Unlike existing phishing detection studies that primarily focus on improving classification accuracy through increasingly complex machine learning architectures, this study addresses a different operational problem: how to prioritize limited analyst attention after suspicious emails have already passed conventional security controls. Recent SOC research increasingly identifies alert prioritization, analyst workload management, and alert-fatigue mitigation as critical operational challenges beyond traditional detection accuracy [3,4]. This challenge is particularly important in healthcare environments, where privilege-sensitive workflows and time-critical communications make human errors operationally disruptive [5,6]. A particularly challenging situation arises when attackers compromise trusted accounts or exploit legitimate communication relationships to deliver socially engineered messages that successfully bypass blacklist-based filtering and Security Email Gateway (SEG) protection [7,8]. As illustrated in Fig. 1, the proposed framework is not intended to replace existing phishing detection mechanisms. Instead, it is deployed as a post-detection human-factor representation layer operating on trusted or whitelisted communications that have already passed conventional email security controls. The framework complements existing technical defenses by providing governance-oriented risk prioritization for Security Operations Center (SOC) analysts under whitelist-abuse scenarios.

Figure 1: Operational deployment of the proposed P×F human-factor representation framework.

Rather than attempting to further strengthen blacklist-based detection, the proposed framework complements existing email security mechanisms by operating as a post-detection governance layer. This study introduces a theory-driven human-factor threat modeling framework for whitelist-failure scenarios. Persuasion tactics (P) represent the influence strategies embedded in social engineering communications, whereas persona framing factors (F) are derived from Hogshead's fascination framework to characterize message framing styles and audience susceptibility [9–11]. The proposed P×F framework does not operate as a conventional keyword-based detector. Instead, it encodes theory-driven human-factor representations derived from persuasion theory and social engineering literature, transforming observable semantic manipulation patterns into interpretable governance-oriented risk features. Although previous studies have incorporated influence-related features into phishing detection [12], limited research has translated human-factor representations into interpretable risk structures suitable for operational governance and analyst decision-support [13,14].

Using publicly reproducible datasets together with LLM-generated emails to evaluate robustness under evolving language styles [15], this study develops a governance-oriented P×F risk representation capable of

identifying high-risk psychological manipulation patterns [16]. The resulting framework supports alert prioritization, security awareness training, and governance-aligned risk assessment consistent with international information security control frameworks and risk management standards [17,18]. Accordingly, the primary contribution of this study is not a new phishing detection engine, but a governance-oriented human-factor representation framework that complements existing email security mechanisms by enabling interpretable post-detection risk prioritization for Security Operations Centers (SOCs).

2 Human-Factor Decision-Support Literature Review

As social engineering attacks continue to evolve, threat intelligence and industry reports consistently identify phishing and credential abuse as among the most frequent threats with the highest operational impact, indicating that purely technical defenses are insufficient for real-world risk mitigation [19,20]. Cybercrime loss statistics also support increased investment in human-factor-oriented protection and training, particularly in high-risk operational environments, as reported by the Federal Bureau of Investigation Internet Crime Complaint Center (IC3) [21]. Rapid changes in attack narratives and linguistic styles further highlight the need to enhance semantic-level sensitivity to psychological manipulation strategies, as discussed by Moura et al. and related phishing trend studies [22].

At the organizational level, empirical studies by the Health Information Sharing and Analysis Center (Health-ISAC) and related public healthcare threat intelligence reports emphasize both the prevalence and operational disruption caused by social engineering, making governance-oriented threat modeling essential for defensive prioritization [23]. Government anti-fraud platforms and law enforcement statistics further summarize common scam patterns, providing practical references for feature design and training scenarios [24]. Recent work by Al-Subaiey et al. integrates explainable artificial intelligence (XAI) into phishing email detection to support alert interpretation and user trust, positioning explainability as a core design requirement [25]. Feature attribution methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), proposed by Lundberg and Lee, enhance transparency and risk understanding without degrading detection performance [26]. Egelman and Peer highlighted the growing importance of human-factor-oriented phishing research and the need to consider organizational and contextual variability in phishing detection analysis [27]. Related studies by Gupta et al. show that explainable design can balance operational deployment and analytical needs [28]. Transforming model outputs into interpretable and comparable cues facilitates governance and audit adoption, as discussed in recent studies on phishing detection optimization by Bari et al. [29], while context-aware and hybrid approaches further improve real-world usability. Standardized machine learning pipelines such as scikit-learn have therefore become a common foundation for reproducible security research [30,31].

Within a broader governance context, quantitative risk assessment models and risk management standards proposed in International Organization for Standardization (ISO) 31000 provide mechanisms for translating detection outputs into actionable decisions [32]. Multi-criteria decision-making (MCDM) approaches have been applied to cybersecurity risk assessment to support structured threat evaluation and prioritization in complex decision environments [33]. Bayesian approaches proposed by Efron provide theoretical foundations for risk estimation under limited data conditions [34]. Machine learning and pattern recognition theory developed by Bishop further underpin classification model design and generalization analysis across heterogeneous data distributions [35]. Empirical Bayes methods extend Bayesian risk modeling by enabling data-driven prior estimation in practical settings, as discussed by Gelman et al. [36]. In addition, MCDM-based security control prioritization provides a complementary perspective for structured security governance and organizational decision support [37]. In human-factors and safety engineering, Failure Mode and Effects Analysis (FMEA) remains a foundational risk analysis method in healthcare

and other high-reliability systems, as summarized by Garosi et al. [38]. Human reliability analysis (HRA) frameworks further support quantitative human-factor risk assessment in complex sociotechnical systems, as discussed by Wang et al. [39]. Cognitive- and system-level human-factor models form the theoretical basis of resilience engineering and the Safety-II paradigm [40].

Recent research has increasingly emphasized alert prioritization and alert-fatigue mitigation as critical operational challenges in Security Operations Centers (SOCs), particularly under conditions of limited analyst capacity and increasing alert volumes [41,42]. While existing studies have investigated phishing detection, explainable artificial intelligence, cybersecurity risk assessment, and human-factor analysis separately, limited research has integrated persuasion-based tactics and persona-oriented framing factors into an interpretable governance-oriented risk prioritization framework. This research gap motivates the development of the proposed P×F framework, which focuses on post-detection human-factor risk prioritization rather than conventional phishing classification.

3 Related Work

The objective of the proposed P×F framework is not to improve phishing classification accuracy, but to operationalize psychological manipulation patterns into interpretable human-factor features for governance-oriented risk prioritization. The proposed P×F framework is not intended to replace advanced Human Reliability Analysis (HRA) methods but to address human-factor risk inference under the severe data constraints commonly encountered in cybersecurity operations. As summarized in Table 1, established Human Reliability Analysis (HRA) approaches such as Healthcare Failure Mode and Effects Analysis (HFMEA), Human Error Assessment and Reduction Technique (HEART), Cognitive Reliability and Error Analysis Method (CREAM), and Functional Resonance Analysis Method (FRAM) typically require explicit task structures, observable operator behavior, and detailed operational process data [38–40]. Such requirements limit their applicability to phishing and social-engineering scenarios, where analysts often have access only to email content and sparse contextual information. In contrast, the proposed P×F framework operates directly on linguistic and semantic evidence extracted from email messages and translates psychological manipulation strategies into governance-oriented risk indicators for Security Operations Center (SOC) alert prioritization.

Table 1: Comparison between advanced human reliability analysis methods and the proposed P×F framework.

Comparison Dimension	Advanced HRA Methods (FMEA/HEART/CREAM/FRAM)	Proposed P×F Framework
Primary Objective	Modeling human error mechanisms and supporting accident prevention	Modeling social engineering risk to support cybersecurity governance and policy decisions
Analytical Level	Task-, action-, cognition-, and system-interaction levels	Email semantic level and abstracted victim susceptibility mechanisms based on Natural Language Processing (NLP) and Machine Learning (ML)
Observable Data Requirements	Explicit task workflows, operational logs, and human-machine interaction data	Email text content and limited contextual cues available in Security Operations Center (SOC) workflows

(Continued)

Table 1 (continued)

Comparison Dimension	Advanced HRA Methods (FMEA/HEART/CREAM/FRAM)	Proposed P×F Framework
Quantitative Depth	High (error probabilities, cognitive control states, system resonance effects)	Moderate (risk matrix + persuasion psychology + Bayesian smoothing)
Applicable Domains	High-reliability physical systems (healthcare operations, nuclear power, aerospace)	Text-based communication and operational security logs (SOC/healthcare email contexts)
Real-Time Deployability	Low (high dependence on expert knowledge and detailed process data)	High (can be embedded directly into email analysis pipelines and Security Information and Event Management (SIEM)/Security Orchestration, Automation and Response (SOAR) workflows)
Decision Outputs	Accident prevention recommendations and system design improvements	Operational prioritization, training focus, and governance policy adjustments (persona-aware)
Research Positioning	Human-factors engineering/Safety engineering	Governance-oriented threat modeling with explicit human-factor awareness for phishing detection

Although the F dimension originates from communication and marketing literature [11], it is adopted in this study as an interpretable message-framing representation rather than a consumer-behavior model. Prior phishing-susceptibility research has shown that attacker success frequently depends on trust building, authority projection, emotional triggering, and persuasion-based message framing embedded in phishing content [12,13]. Accordingly, the seven F factors (Trust, Prestige, Power, Alarm, Passion, Mystique, and Rebellion) are used as structured framing categories that capture how phishing messages present psychological influence cues to potential victims. Unlike traditional HRA frameworks that focus on task execution and human error mechanisms, the proposed F layer focuses on attacker message framing and victim susceptibility signals observable directly from email content. This design enables psychological influence patterns to be represented in an interpretable and auditable form suitable for operational cybersecurity environments. Rather than estimating precise human error probabilities, the proposed framework operationalizes these signals as governance-oriented risk indicators that support Top-K alert prioritization under limited SOC resources.

4 Research Framework

As illustrated in Fig. 2, the proposed P×F framework is not intended to compete with advanced Human Reliability Analysis (HRA) methods but to operationalize human-factor risk inference under the severe data constraints typical of cybersecurity operations. The workflow begins with a conceptual layer that defines the principles of P×F human-factor threat modeling, followed by a modeling layer that constructs a scenario taxonomy and encodes persuasion tactics and persona indicators from social engineering messages. In the risk quantification stage, Exposure (E) and Task Vulnerability (V) represent likelihood and impact under

auditable organizational evidence. These components are integrated in the threat amplification layer, where persona-weighted and smoothed P×F matrices highlight governance-relevant risk hotspots. Finally, the framework is validated through model integration by combining P×F outputs with email semantic detection models for operational evaluation.

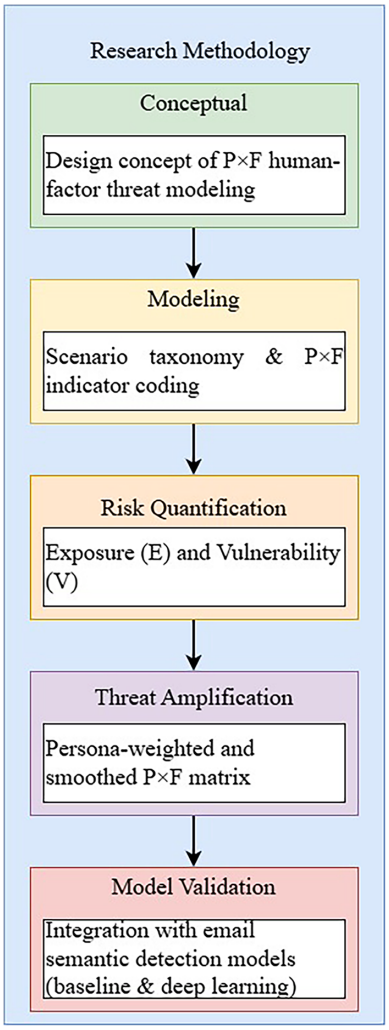

Figure 2: P×F human-factor threat modeling with machine learning integration.

As summarized in Table 1, established HRA approaches such as Healthcare Failure Mode and Effects Analysis (HFMEA), Human Error Assessment and Reduction Technique (HEART), Cognitive Reliability and Error Analysis Method (CREAM), and Functional Resonance Analysis Method (FRAM) provide greater theoretical depth and quantitative rigor for modeling human error mechanisms, cognitive control states, and system-level safety risks and have been widely applied in healthcare and other high-reliability domains [38,39]. However, these methods generally assume well-defined task structures, observable operator behavior, and detailed human-machine interaction data, which are rarely available in phishing and social engineering contexts where analysis is limited to message content and sparse contextual cues [40].

Accordingly, the proposed P×F framework is positioned as a governance-oriented intermediate human-factor modeling layer operating primarily on linguistic and semantic evidence. Rather than estimating

precise human error probabilities or reconstructing complete cognitive task models, the framework translates psychological manipulation strategies into auditable risk indicators that support risk aggregation, Security Operations Center (SOC) alert prioritization, and training scenario planning in data-constrained cybersecurity environments. This design therefore complements, rather than replaces, advanced HRA methods operating at higher levels of organizational safety and human reliability analysis. The resulting formulation operationalizes a human-factor risk amplification model in which persuasion tactics and persona framing jointly modulate the underlying technical risk.

4.1 Concept of P×F Threat Modeling Matrix and Indicator Design

In practical cybersecurity risk assessments, risk ratings are based on expert judgment and subjective scales, limiting traceability and institutional auditability [18,19]. To address this limitation, this study adopts interpretability and auditability as primary design principles and structurally encodes psychological pressure tactics and message-framing patterns commonly used in social engineering attacks to construct a governance-ready P×F human-factor threat modeling matrix [10,12]. Rather than focusing solely on external indicators such as malicious senders or suspicious links, the proposed approach emphasizes semantic-level manipulation cues to address scenarios in which trusted accounts are compromised or whitelisted identities are abused. The healthcare threat context is grounded in publicly available threat intelligence, including Health-ISAC reports and government anti-fraud statistics, real-world operational observations. These sources consistently identify authority abuse, urgency induction, and credential solicitation as dominant attack patterns, which closely align with the high-risk P×F combinations observed in this study [14].

To ensure operational relevance, the P×F matrix is aligned with social engineering patterns commonly encountered in healthcare organizations by consolidating publicly documented fraud scenarios into a set of 30 representative scam types as the foundational threat modeling space [24]. A two-layer indicator encoding scheme is then applied. The P layer is derived from Cialdini's influence principles and defines five persuasion-based psychological pressure indicators, each binary-coded to form computable manipulation cues suitable for feature engineering and matrix construction [10,12]. The F layer adopts Hogshead's fascination-based persona framework to characterize attacker message framing styles and victim susceptibility triggers, likewise, encoded in a structured binary form to address the limited treatment of message-style variation in prior research [11,13]. By integrating the P and F layers, the proposed framework establishes a traceable mapping between semantic-level social engineering cues and governance-actionable risk hotspots, directly supporting SOC alert prioritization, training design, and rule development [15,16].

$$Risk_{base} = PL \times E \times V$$

Note: The base risk score follows the likelihood-impact risk concept proposed in National Institute of Standards and Technology (NIST) Special Publication (SP) 800-30 and is extended by incorporating persuasion leverage (PL), exposure (E), and task vulnerability (V) to form a governance-oriented multiplicative human-factor risk indicator [18].

Persuasion Leverage (PL) represents the intensity of psychological pressure on decision heuristics; Exposure (E) captures organizational exposure and occurrence frequency as a likelihood proxy; Task Vulnerability (V) reflects security and operational impact together with susceptibility to procedural deviation [18,19]. In healthcare settings, Business Email Compromise (BEC) attacks often occur within trusted boundaries, limiting indicator-based defenses; incorporating E and V improves governance relevance and operational applicability [15,23].

4.2 Hospital-Wide Exposure (E) and Task Vulnerability (V): Operational Definitions

The E and V are operationalized using auditable SOC records and mapped to a 1–5 scale (Tables 2 and 3), ensuring traceability and cross-scenario comparability.

Table 2: Operational definition and grading rules for exposure (E, 1–5).

E Level	Operational Definition (Exposure Frequency and Organizational Coverage)	Auditable Evidence Sources (Hospital A)	Representative Healthcare Scenarios (Mapped to 30 Fraud Types)
5	Near-daily exposure across multiple clinical and administrative roles via routine email channels	Recurrent daily detections or SOC reports from email gateways; security training feedback indicating high familiarity	Official hospital notices, credential login prompts, system alerts, HIS/EMR-related phishing
4	Multiple exposures per week through routine operational workflows	Stable occurrence in SOC tickets or reports across multiple departments	Shift scheduling, training notifications, administrative announcements, internal portal links
3	Weekly to monthly exposure; not a daily workflow but encountered by multiple roles	Moderate SOC reporting frequency with cross-role visibility	Government or insurance notifications, document approval requests, medical report-related emails
2	Occasional exposure, primarily limited to specific workflows or specialized roles	Events concentrated in specific roles with limited historical records	Procurement, maintenance, vendor communications, or project-related emails
1	Rare or negligible exposure; unrelated to routine healthcare workflows	Survey and interview results indicating minimal exposure; very few SOC reports	Non-work-related scams such as investment, dating, or lottery fraud

Table 3: Task vulnerability V (1–5) grading rules.

V Level	Operational Definition (Impact + Susceptibility)	Auditable Evidence (Available at Hospital A)	Representative Healthcare Scenarios (Mapped to 30 Categories)
5	Execution directly causes severe and often irreversible damage, typically under time pressure	Credential/privilege/financial traceability; high-severity incident reports	Password reset, OTP/credential disclosure, fund transfer/payment, privilege activation, VPN credentials, system configuration changes

(Continued)

Table 3 (continued)

V Level	Operational Definition (Impact + Susceptibility)	Auditable Evidence (Available at Hospital A)	Representative Healthcare Scenarios (Mapped to 30 Categories)
4	High-risk actions that may trigger security incidents; require strong judgment to avoid attack chains	EDR alerts and forensics linking download-execution-exfiltration	File download/software update, personal data submission, password less login redirection, approval/signature links
3	Moderate impact; workflow deception or initial data exposure causing localized disruption	Tickets indicating misclicks or misresponses without organization-wide impact	Fake announcements, fake training notices, fake duty rosters
2	Low impact; mitigable through multiple procedural or technical controls	Review/approval or ticketing mechanisms can interrupt execution	General awareness scams, non-critical administrative workflows
1	Negligible impact; no sensitive operations involved	Clicks without login, download, or authorization; or offline content	Lifestyle-related messages without links or actions

Note: Exposure (E) and Task Vulnerability (V) are defined using auditable SOC records at Hospital A and mapped to a unified 1–5 scale (Tables 2 and 3), ensuring traceability, cross-scenario comparability, and minimized subjectivity in governance-oriented risk assessment.

4.3 Persona-Weighted Risk Amplification and Smoothed P×F Matrix Construction

Building upon the base risk score, this study first defines $Risk_{\text{base}}$ as follows:

$$Risk_{\text{base}} = PL \times E \times V$$

where Persuasion Leverage (PL) represents psychological influence intensity, Exposure (E) represents organizational exposure likelihood, and Task Vulnerability (V) denotes operational susceptibility and potential impact. Consistent with NIST SP 800-30 and ISO-aligned risk assessment practices, the base risk formulation follows a likelihood-impact structure adapted for human-factor-oriented phishing risk assessment. In this study, Exposure (E) and Task Vulnerability (V) serve as auditable SOC-based proxies for likelihood and impact, while Persuasion Leverage (PL) captures the intensity of psychological pressure embedded in the phishing scenario. The 1–5 scales for E and V are derived from hospital SOC operational records, documented email exposure frequency, incident reports, and observed impact severity. To improve transparency, the revised manuscript now further clarifies the rationale for adopting a 1–5 scale. The selected scale represents a practical compromise between granularity and operational usability, providing sufficient discrimination among phishing scenarios while avoiding excessive subjectivity and annotation inconsistency. Consistent with NIST SP 800-30 and ISO-aligned risk-assessment practices, E and V serve as auditable proxies for likelihood and impact. While the specific scoring criteria may be locally calibrated according to organizational exposure patterns and operational requirements, the underlying likelihood–impact structure remains

applicable across different organizational contexts. Documented email exposure frequency, incident reports, and observed impact severity, thereby providing a governance-oriented semi-quantitative implementation of risk assessment principles. The 1–5 scale was selected as a practical compromise between granularity and operational usability. A coarser scale may not provide sufficient discrimination among phishing scenarios, whereas a finer scale may introduce excessive subjectivity and annotation inconsistency.

This study further incorporates persona factors to capture the amplifying effects of combined message framing and psychological persuasion strategies, yielding a human-factor-aware risk score denoted as $Risk_{\text{with}}\text{F}$ [32]:

$$Risk_{\text{with}}\text{F} = Risk_{\text{base}} \times \left(1 + min\left(0.4, 0.1 \times Persona_{Sum}\right)\right)$$

where $Persona_{Sum}$ denotes the cumulative persona evidence score obtained by aggregating the encoded persona-related evidence associated with Trust, Prestige, Power, Alarm, Passion, Mystique, and Rebellion. Larger values indicate stronger cumulative psychological manipulation evidence rather than merely a greater number of activated persona categories. A semi-quantitative modifier is therefore introduced to adjust the base risk according to the accumulated human-factor influence. Detailed mathematical derivations, matrix aggregation procedures, and smoothing estimators are provided in Appendix A.

The incremental coefficient (0.1) represents the marginal contribution of cumulative persona evidence to risk amplification, whereas the upper bound (0.4) limits the maximum amplification effect to preserve ranking stability and interpretability. Rather than universal constants, these parameters are treated as governance-oriented calibration parameters. Their robustness was further examined through sensitivity analysis, demonstrating stable Top-10 P×F rankings across moderate parameter variations. The proposed P×F taxonomy was developed through a theory-driven knowledge-engineering process grounded in persuasion theory and human-factor literature. The P layer comprises five persuasion tactics (Consistency, Authority, Social Proof, Urgency, and Liking), whereas the F layer consists of seven persona-framing factors (Trust, Prestige, Power, Alarm, Passion, Mystique, and Rebellion).

For each email, the message content is interpreted according to predefined semantic interpretation rules derived from the P×F taxonomy. Observable psychological manipulation evidence is interpreted and encoded into persuasion tactics (P) and persona framing factors (F), which are subsequently aggregated into the P×F representation for risk amplification analysis. When multiple semantic cues belonging to the same category are identified, their semantic evidence is accumulated into category-specific scores, whereas evidence corresponding to different predefined P/F categories contributes independently to each applicable category. This design preserves reproducibility, interpretability, and traceability by linking each encoded indicator to explicit semantic evidence rather than latent model representations. The P×F representation operates independently of the TF-IDF + Logistic Regression classifier and serves as an interpretable human-factor representation layer for governance-oriented risk prioritization. BERT was included only as a comparative baseline during model evaluation and was not involved in P×F feature construction.

Risk amplification (ΔRisk) represents the incremental risk attributable to persona-related influence under identical exposure and vulnerability conditions. Aggregated ΔRisk values are subsequently summarized across persuasion tactics (P) and persona-framing factors (F) to construct the P×F risk matrix, providing an interpretable representation of relative psychological risk hotspots for SOC alert prioritization. Fig. 3 illustrates the smoothed P×F risk amplification matrix constructed from the average ΔRisk values computed using the calibrated $Risk_{\text{with}}\text{F}$ formulation described above. Each matrix cell represents the mean incremental risk associated with a specific combination of persuasion tactics (P1–P5) and persona framing factors (F1–F7). Shrinkage smoothing is applied to mitigate sparse observations and extreme values, enabling the matrix to provide a governance-oriented representation of relative psychological risk amplification

rather than a per-instance risk estimate. Risk amplification is primarily concentrated on Authority (P2), Urgency (P4), and Liking (P5), particularly when combined with persona framing factors such as Power, Prestige, Alarm, and Passion. In contrast, Consistency (P1) and Social Proof (P3) exhibit relatively limited amplification. Cells with near-zero values indicate low occurrence frequency or weak amplification rather than the absence of risk. Consequently, the matrix highlights relative psychological risk hotspots that support SOC alert prioritization under constrained operational resources.

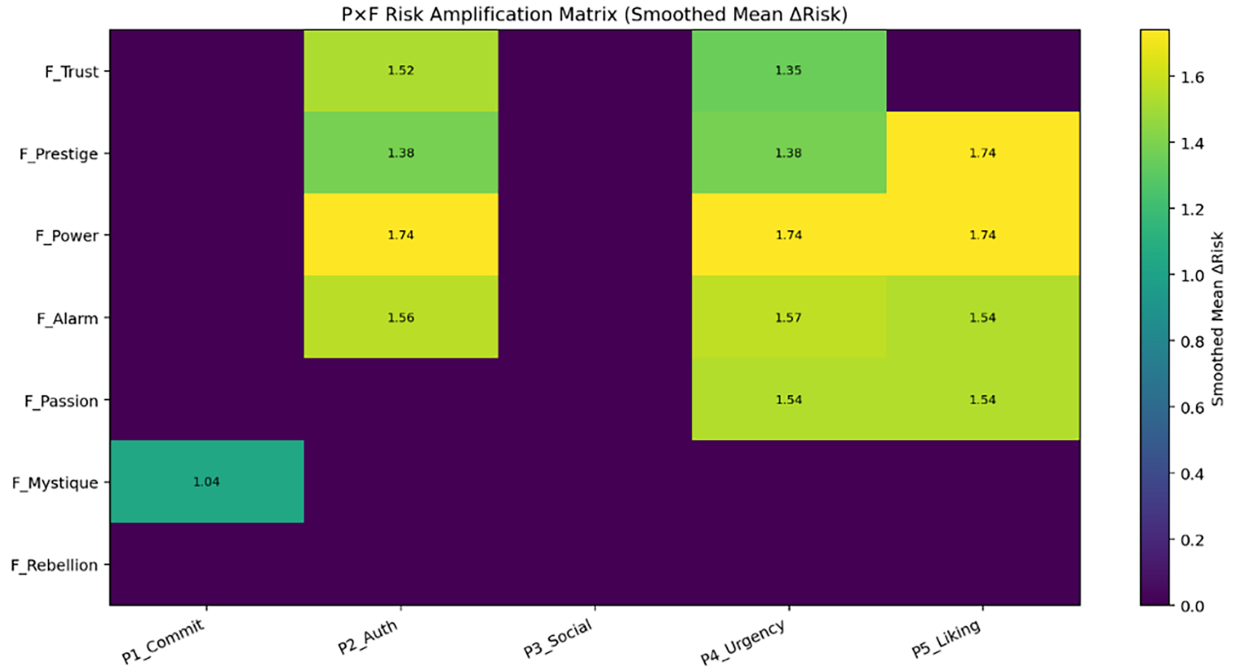

Figure 3: Smoothed P×F human-factor risk amplification matrix (average ΔRisk, shrinkage smoothing).

To examine the robustness of the proposed governance-oriented calibration parameters, a sensitivity analysis was conducted by varying the incremental coefficient (α = 0.05, 0.10, and 0.15) and the amplification cap (0.30, 0.40, and 0.50). For each parameter combination, the ΔRisk values were recomputed and the resulting P×F risk matrix and Top-10 P×F combinations were regenerated. The overlap of the Top-10 combinations relative to the baseline configuration (α = 0.10, cap = 0.40) was used as the primary robustness indicator, and the results are summarized in Table 4.

As shown in Table 4, the proposed P×F prioritization framework remains robust under moderate variations of the calibration parameters. Across all tested parameter settings, the overlap of the Top-10 P×F combinations ranged from 90% to 100%, indicating that the identification of high-risk psychological patterns is largely insensitive to reasonable changes in the calibration coefficient and amplification cap. These findings provide empirical support for selecting α = 0.10 and cap = 0.40 as conservative governance-oriented calibration parameters, as they preserve ranking stability while maintaining human-factor interpretability across the evaluated parameter settings.

Table 5 summarizes the top ten P×F psychological combinations with the highest mean ΔRisk, explicitly mapping each influence strategy (P) to its corresponding persona factor (F) to identify which "persuasion tactic × message framing" patterns most consistently produce significant risk amplification in real-world email scenarios. The results show that high-risk combinations are strongly concentrated around authority- and identity-related persona factors (Power, Prestige, and Trust), with Power most frequently co-occurring

with Urgency, Liking and Authority persuasion tactics. This pattern reflects a stable social engineering mechanism in which perceived legitimacy is first established and then leveraged to compress victims' decision time. By aggregating risk amplification at the P×F level, the proposed framework translates semantic phishing detection into an interpretable and auditable decision layer, directly supporting Security Operations Center (SOC) alert prioritization, analyst workload management, training focus, and National Institute of Standards and Technology (NIST) Cybersecurity Framework (CSF)-aligned risk-informed governance, without reliance on institution-specific rules.

Table 4: Sensitivity analysis of the calibration parameters (α and amplification cap) based on Top-10 P×F ranking stability.

α	Cap	Top-10 Overlap with Baseline (%)
0.05	0.30	90%
0.05	0.40	90%
0.05	0.50	90%
0.10	0.30	100%
0.10	0.40	100%
0.10	0.50	90%
0.15	0.30	100%
0.15	0.40	100%
0.15	0.50	100%

Table 5: Top-10 P×F combinations with highest mean ΔRisk.

F_Factor	P_Tactic	Mean_DeltaRisk
F3-Power	P4-Urgency	2
F3-Power	P5-Liking	2
F2-Prestige	P5-Liking	2
F3-Power	P2-Auth	2
F4-Alarm	P2-Auth	1.6
F-Passion	P4-Urgency	1.6
F-Passion	P5-Liking	1.6
F4-Alarm	P5-Liking	1.6
F4-Alarm	P4-Urgency	1.6
F1-Trust	P2-Auth	1.528571429

The proposed framework assumes that phishing-related psychological influence cues can be represented through the manually curated P and F indicators and that SOC operational records provide a reasonable approximation of organizational exposure and task vulnerability. The framework is intended for governance-oriented risk prioritization rather than forensic attribution or definitive phishing classification. In addition, Hospital A proof-of-concept relies on weak-label operational data and may reflect analyst or workflow-related bias. Finally, the calibration parameters used in the P×F amplification mechanism is intended to support interpretability and ranking stability for governance-oriented deployment and should not be interpreted as universal constants across all organizations.

4.4 Integration with Email Semantic Detection Model

After constructing the P×F human-factor threat matrix, this study establishes a risk representation framework centered on social-engineering psychology, transforming linguistic pressure tactics and message framing into quantifiable, comparable, and auditable human-factor risk components. Prior studies show that phishing effectiveness relies heavily on semantic narratives, contextual framing, and psychological pressure rather than purely technical indicators [8,12,13]. Accordingly, this study adopts persuasion tactics (P) and framing/susceptibility factors (F) as dual encoding axes to form a P×F threat modeling matrix. Rather than classifying individual emails, the matrix highlights recurrent psychological combinations with high-risk amplification potential in organizational contexts. Positioned as a governance-oriented mediation layer, the framework supports alert prioritization, training design, and security control planning, enabling human-factor risks being systematically integrated into organizational risk management.

5 Experimental Design

As illustrated in Fig. 4, the experimental workflow of this study is structured into four sequential stages: dataset construction, data preprocessing, feature engineering, and model training and evaluation. The figure provides a consolidated overview of how raw email corpora are transformed into structured semantic and human-factor representations and how these representations are systematically evaluated under controlled and external testing conditions to assess both detection performance and governance applicability.

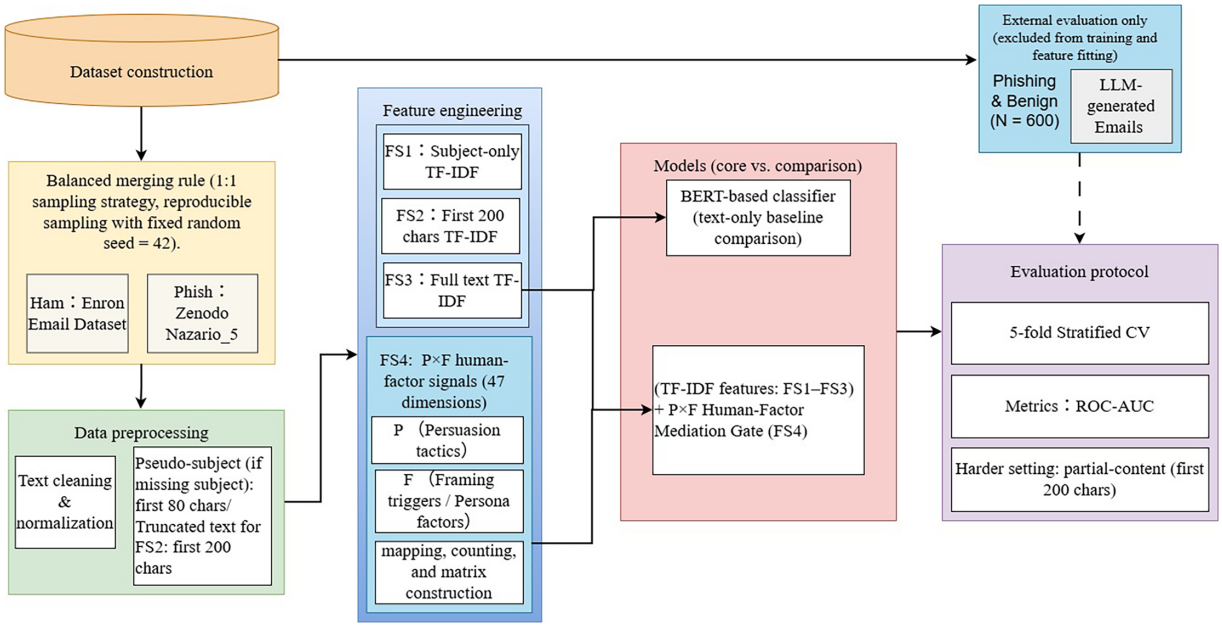

Figure 4: Data set construction and email phishing detection process based on logistic regression model.

On the left side of Fig. 4, the dataset construction stage depicts the balanced sampling process applied to publicly available real-world corpora. Following a 1:1 sampling strategy with a fixed random seed to ensure reproducibility, 3039 legitimate emails (ham) were randomly selected from the Enron Email Dataset [1], and 3039 phishing emails (phish) were sampled from the Nazario_5 dataset curated on Zenodo [2]. These datasets consist entirely of human-authored emails collected prior to the emergence of generative artificial intelligence (AI), making them suitable as baseline corpora for analyzing human-centered persuasion and social-engineering language structures.

The middle section of Fig. 4 corresponds to data preprocessing and feature engineering. Emails were normalized and processed under different text-availability settings to reflect practical detection constraints, including subject-only input, truncated content limited to the first 200 characters, and full-text content. Based on these settings, three Term Frequency-Inverse Document Frequency (TF-IDF) feature representations, denoted as Feature Set 1 to Feature Set 3 (FS1–FS3), were constructed to examine the effect of textual completeness on detection performance [6,8].

Specifically, FS1 represents subject-only TF-IDF features, FS2 represents truncated-content TF-IDF features using the first 200 characters of each email, and FS3 represents full-text TF-IDF features derived from the complete email body. In parallel, the proposed P×F human-factor feature layer, denoted as Feature Set 4 (FS4), was derived by mapping persuasion tactics (P) and persona-based framing factors (F) into a structured 47-dimensional representation. These features were further aggregated into a P×F matrix to support interpretable risk analysis and governance-oriented interpretation [10,12].

The right side of Fig. 4 presents the model training and evaluation stage. A Logistic Regression classifier combining TF-IDF and P×F features were adopted as the primary baseline model due to its interpretability, reproducibility, and suitability for operational deployment in security contexts [6,8]. In addition, a Bidirectional Encoder Representations from Transformers (BERT)-based semantic classification model was included as a performance-oriented comparison baseline, using raw text input only and not participating in human-factor feature construction, thereby serving as an upper-bound reference for semantic modeling capability rather than a governance-focused solution [7,13].

All models were evaluated using fixed five-fold stratified cross-validation, with Receiver Operating Characteristic Area under the Curve (ROC-AUC) as the primary evaluation metric. To further assess robustness under distribution shift, an external evaluation set consisting of Large Language Model (LLM)-generated emails (N = 600) was constructed and used exclusively for out-of-distribution testing, without involvement in model training or feature engineering [1,2].

Through this unified experimental pipeline, Fig. 4 clarifies the methodological linkage between semantic detection, human-factor modeling, and governance-oriented evaluation. Rather than pursuing maximum classification accuracy alone, the design emphasizes controlled comparability, robustness under partial observability, and the interpretability required to translate phishing detection outputs into actionable security operations and risk-informed decision-support.

5.1 Dataset Construction

Before constructing the experimental dataset, a structured knowledge engineering process was conducted to establish the P×F human-factor taxonomy used throughout this study. Rather than defining human-factor indicators subjectively, the proposed taxonomy was developed by integrating multiple complementary knowledge sources relevant to phishing and social engineering. These sources included (1) peer-reviewed phishing and social engineering literature, (2) publicly available fraud cases and anti-fraud guidance published by the Taiwan National Police Agency, (3) persuasion principles and psychological manipulation strategies discussed in red-team psychology references, and (4) publicly documented healthcare phishing cases, de-identified hospital cybersecurity awareness materials, and public healthcare threat intelligence reports. The objective was to identify psychological influence cues that could be directly observed from email content while maintaining reproducibility and minimizing subjective interpretation. Accordingly, only observable textual cues were retained for subsequent feature encoding.

Candidate human-factor indicators were extracted from these knowledge sources and subsequently reviewed for semantic overlap and conceptual consistency. Similar concepts describing equivalent attacker

behaviors or psychological manipulation strategies were merged into unified categories, whereas indicators that could not be directly inferred from observable email content were excluded. This process ensured that every retained indicator could be encoded from textual evidence without requiring behavioral observation, user profiling, or subjective expert judgment.

The resulting taxonomy was organized into two complementary dimensions. The P dimension represents persuasion tactics that describe how attackers attempt to influence user decisions, whereas the F dimension represents persona-based framing factors that capture the psychological personas embedded in phishing messages, including Trust, Prestige, Power, Alarm, Passion, Mystique, and Rebellion.

Table 6 summarizes the structured knowledge engineering workflow used to construct the proposed P×F human-factor taxonomy. The resulting taxonomy provides the conceptual foundation for feature encoding, and the construction of the human-factor feature set (FS4) used throughout the experimental evaluation.

Table 6: Knowledge engineering workflow for constructing the proposed P×F taxonomy.

Stage	Knowledge Source	Objective	Output
Literature review	Phishing/SE literature	Extract persuasion concepts	Candidate indicators
Anti-fraud cases	Taiwan NPA	Identify scam patterns	Candidate indicators
Theory & psychology	Persuasion theory/Red-team psychology	Extract psychological strategies	Candidate indicators
Healthcare sources	Public healthcare phishing cases	Identify healthcare-specific cues	Candidate indicators
Refinement	Candidate indicators	Merge concepts; retain observable cues	Final taxonomy

The dataset used in this study consists of two categories of email corpora—legitimate emails (Ham) and phishing emails (Phish)—to support binary phishing detection modeling. Legitimate emails were sourced from the Enron Email Dataset [1], which contains real-world corporate emails and has been extensively used in email detection and text classification research, serving as a representative baseline for business communication patterns. Phishing emails were obtained from the publicly available Nazario_5 dataset on Zenodo [2], which includes real-world phishing samples exhibiting common social engineering semantics such as urgency, authority, and threat cues, aligning with the human-factor focus of this study [2,5].

To mitigate class imbalance effects, a 1:1 balanced sampling strategy was applied, as summarized in Table 7. Random sampling with a fixed seed was used when necessary to ensure reproducibility [5]. After sampling, both corpora were merged into a unified dataset for subsequent preprocessing, feature extraction, and model training.

5.2 Unified Data Preprocessing and Feature Engineering

For notation clarity, F (F_Trust–F_Rebellion) denotes persona-related factors in the P×F framework, while FS1–FS4 refer to the four feature sets used in the experiments. A unified preprocessing pipeline was applied to standardize heterogeneous email corpora and reduce noise. Missing subject or body fields

were padded and merged when necessary, formatting artifacts were removed and forwarded or quoted content was truncated to prevent feature contamination. When subject fields were absent, a pseudo-subject was constructed from the first 80 characters of the email text. Five-fold stratified cross-validation with a fixed seed was adopted to ensure reproducibility [30]. To reflect practical detection constraints, three text availability settings—subject-only, truncated text (first 200 characters), and full text—were defined to evaluate performance under partial and complete information conditions [6,8].

Table 7: Dataset summary.

Class_Name	Label	Number_of_Emails
Ham	0	3039
Phish	1	3039

The proposed P×F framework does not operate as a conventional phishing detection engine. Instead, it encodes observable psychological manipulation strategies derived from persuasion theory and human-factor literature. The semantic interpretation process captures persuasion tactics and persona framing reflected by the overall communication context rather than isolated textual expressions. Email content is encoded according to predefined semantic interpretation rules derived from the manually curated P×F taxonomy. The encoding process identifies observable psychological manipulation cues that correspond to persuasion tactics (P) and persona framing factors (F). Rather than representing lexical occurrences, the resulting P/F indicators encode interpretable semantic human-factor evidence for subsequent risk prioritization. When multiple semantic cues belonging to the same persuasion or persona category are identified within an email, their semantic evidence is accumulated into category-specific scores. Semantic cues corresponding to multiple predefined P/F categories contribute independently to each applicable category.

To balance detection accuracy and interpretability, a hybrid feature architecture combining statistical text representations and human-factor threat lexicon features was employed [8,12]. FS1–FS3 use TF-IDF representations under the three text availability settings to assess the impact of textual completeness on classification performance [6,8]. Since TF-IDF captures lexical salience but cannot directly model psychological manipulation strategies or cognitive bias triggers, FS4 introduces a P×F lexicon feature layer that translates observable social engineering semantics into structured human-factor representations [15]. The P layer encodes persuasion tactics grounded in influence principles [10,12], while the F layer captures message framing styles and susceptibility triggers beyond persuasion-only modeling [11,13]. The resulting P–F mapping is summarized in Table 8, with complete persona definitions and risk rankings provided in Appendix A to support governance-oriented analysis and operational deployment [15,16].

Table 8: Mapping of persona factors (F) to common trap patterns and dominant persuasion levers (P).

Persona Factor (F)	Attacker Message Framing	Common Trap Pattern	Dominant Cialdini Persuasion Levers (P)
F_Trust (Trust)	"I am an internal staff member/official contact/trusted entity"	Credential handover trap: disclosing One-Time Passwords (OTPs) or passwords, clicking login pages, granting remote access	P2 Authority, P1 Commitment & Consistency

(Continued)

Table 8 (continued)

Persona Factor (F)	Attacker Message Framing	Common Trap Pattern	Dominant Cialdini Persuasion Levers (P)
F_Prestige (Prestige)	"Expert/attending physician/hospital authority endorsement"	Authority compliance trap: executing instructions without verification	P2 Authority
F_Power (Power)	"You must comply/I have disciplinary authority"	Coercive command trap: privilege escalation, shortcuts, bypassing approval processes	P2 Authority, P4 Scarcity/Urgency
F_Alarm (Alarm)	"Account anomaly/system compromise/imminent suspension"	Fear-urgency trap: clicking links or resetting credentials under time pressure	P4 Scarcity/Urgency
F_Passion (Passion)	"I urgently need your help/I am a patient's family member"	Empathy exploitation trap: emotional pressure leading to procedural bypass	P5 Liking
F_Mystique (Mystique)	"Internal information/exclusive channel/special vulnerability"	Information asymmetry trap: Induction Use informal channels (e.g., messaging apps, external links)	P3 Social Proof, P4 Scarcity/Urgency
F_Rebellion (Rebellion)	"Ignore the rules-this is faster and more efficient"	Policy evasion trap: bypassing controls, Single Source of Truth (SSOT), or ticketing workflows	P1 Commitment & Consistency, P4 Scarcity/Urgency

As summarized in Table 8, the proposed indicator construction maps common social engineering scenarios into structured persona, framing, trap, and persuasion layers, grounded in clinician-validated healthcare fraud rankings and public anti-fraud statistics to ensure domain consistency and auditability, thereby providing a governance-oriented foundation for feature extraction and interpretation [10,12].

5.3 Models and Settings

Logistic Regression (LR) was adopted as the primary baseline classifier to systematically evaluate the effectiveness of different feature sets (FS1–FS4) for phishing detection. Compared with deep learning approaches, LR provides stable training behavior, strong interpretability, and low computational overhead, making it well suited for feature-level comparison and reproducible evaluation under high-dimensional sparse representations [5,6,30]. All text vectorization settings, classifier parameters, and cross-validation configurations were fixed to ensure consistency and are summarized in Table 9 [30,31].

Table 9: Summary of experimental configurations used in FS1–FS4 and 5-fold cross-validation.

Component	Parameter	Value
TF-IDF	max_features	50,000
TF-IDF	ngram_range	(1, 2)
TF-IDF	min_df	2
TF-IDF	lowercase	TRUE
TF-IDF	token_pattern	(?u)\b\w+\b
Logistic Regression	solver	liblinear
Logistic Regression	max_iter	2000
Logistic Regression	random_state	42
Cross Validation	folds	5
Cross Validation	split	fixed fold_id in dataset

5.4 Evaluation Protocol and Reproducible Implementation

To ensure robustness and reduce bias from single data splits, both the TF-IDF + Logistic Regression (LR) baseline and the BERT-based semantic classifier are evaluated using five-fold stratified cross-validation. Each fold uses approximately 80% of the data for training and 20% for validation, while preserving class balance between ham and phishing emails [30,31].

All experiments are implemented in Python using standardized machine learning toolchains. Feature sets FS1–FS3 are evaluated using TF-IDF with Logistic Regression implemented in scikit-learn, whereas FS4 is subsequently applied as a P×F human-factor mediation layer to the full-text classifier output under the same evaluation protocol. The BERT baseline is trained and tested under the same data splits and evaluation protocol to ensure methodological consistency [5]. To enhance reproducibility, random seeds are fixed during sampling, data splitting, and training, and fold assignments are preserved so that all feature sets and models are compared under identical conditions [30]. ROC-AUC is adopted as the primary evaluation metric to measure overall discriminative capability across decision thresholds, with mean and standard deviation reported over five folds [31]. Precision, Recall, and F1-score (F1) are additionally reported as secondary metrics to characterize trade-offs between false positives and false negatives in deployment-oriented scenarios [5,8]. Finally, harder tests using limited inputs (Subject-only, FS1; First 200 characters, FS2) are conducted to assess model robustness under realistic SOC operational constraints [14].

6 Results and Analysis

This section systematically evaluates the proposed phishing email detection framework under different feature settings and data conditions, focusing on the stability and consistency of the baseline model under a reproducible experimental protocol. Using publicly available real-world email corpora, a TF-IDF-based Logistic Regression baseline is first evaluated across varying text availability and human-factor configurations. As recent phishing attacks exhibit increasingly coherent and context-aware semantics, evaluations relying solely on historical datasets may underestimate model robustness in emerging social engineering scenarios [3,4]. Accordingly, deep semantic models and synthetically generated emails are introduced as external test sets to examine model generalization under distribution shift, serving as a complementary assessment rather than a replacement for real-world data-driven evaluation [5,20].

6.1 Baseline Performance on Real-World Corpora

This section reports phishing email detection results obtained from real-world email corpora and compares the classification performance of different feature settings (FS1–FS4) under a unified evaluation protocol. All experiments adopt fixed 5-fold stratified cross-validation using Logistic Regression with Term Frequency–Inverse Document Frequency (TF-IDF) vectorization as the baseline classifier. All feature settings are evaluated under identical data splits, and the aggregated baseline results for FS1–FS3 are summarized in Table 10 [5]. As shown in Table 10, the three text availability settings exhibit consistently strong discriminative performance in terms of ROC-AUC. Notably, FS1, which relies solely on email subject lines, achieves a mean ROC-AUC of 0.9829, indicating that even under limited information conditions, subject lines alone capture highly discriminative social engineering cues. Expanding the input to the first 200 characters (FS2) or full email content (FS3) results in only marginal changes, with ROC-AUC values of 0.9803 and 0.9871, respectively. These findings suggest that key psychological manipulation signals are often concentrated in the subject line or early content of phishing emails [14].

Table 10: ROC-AUC summary for FS1–FS3.

FeatureSet	Description	AUC_Mean	AUC_Std
FS1	Subject-only (subject_ps)	0.982949971	0.002965773
FS2	First 200 characters	0.980560453	0.003136883
FS3	Full text	0.987103703	0.002028166

The standard deviation of ROC-AUC across folds remains low (0.002–0.003), demonstrating strong model stability and reproducibility across data partitions. Such consistency is important for operational Security Operations Center (SOC) environments where detection quality must remain robust across varying email streams [8]. Fig. 5 presents the ROC curves for FS1–FS4 under the same evaluation protocol. In addition to the TF-IDF baseline settings (FS1–FS3), FS4 introduces the proposed P×F human-factor mediation gate applied to the full-text classifier output. The FS4 configuration achieves an AUC of 0.9857, remaining comparable to the FS3 baseline (AUC = 0.9871). This result indicates that incorporating human-factor mediation does not materially degrade classification performance while enabling governance-oriented risk interpretation for SOC analysis [5]. The primary contribution of FS4 therefore lies in risk prioritization and human-factor interpretability rather than in improving conventional classification metrics.

FS4 is evaluated as a post-processing configuration on top of FS3; therefore, Table 10 reports FS1–FS3 baselines, while Fig. 5 includes FS4 for comparison. Following the ROC comparison in Fig. 5, Table 11 provides a focused quantitative comparison between the full-text baseline (FS3) and the proposed FS4 configuration. Unlike FS1–FS3, which rely purely on textual TF-IDF features, FS4 introduces the proposed P×F human-factor mediation gate applied to the classifier output. As shown in Table 11, the FS4 configuration achieves an AUC of 0.9857, which remains highly comparable to the FS3 baseline (AUC = 0.9871). The small difference indicates that incorporating human-factor signals preserves the discriminative behavior of the baseline model without materially affecting its predictive capability. Instead, the mediation layer enables the integration of persuasion and framing cues for governance-oriented risk interpretation, which is particularly valuable for prioritization and analyst decision-support in operational SOC environments.

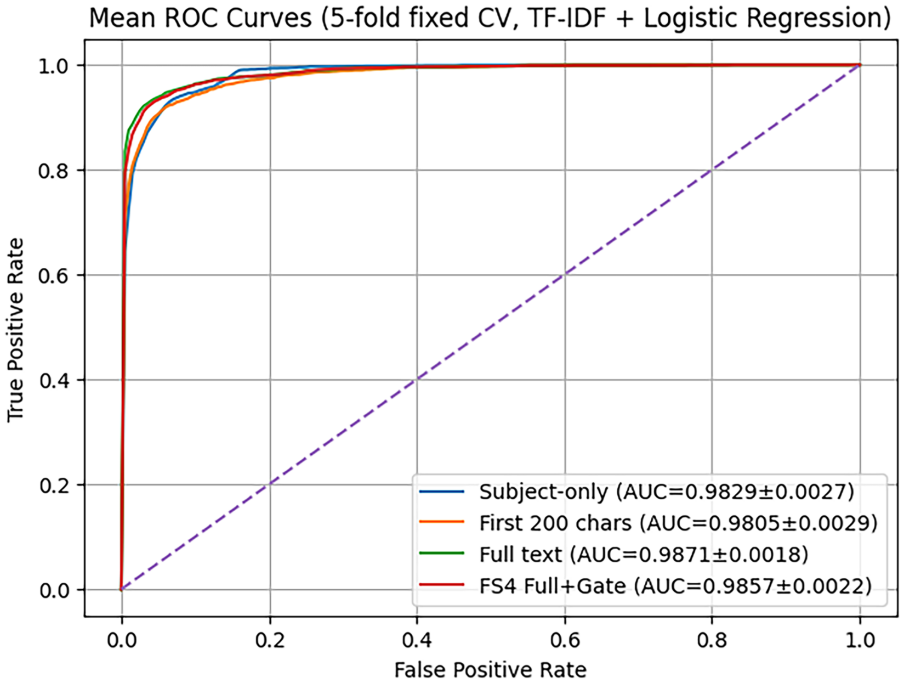

Figure 5: ROC curves (TF-IDF + Logistic Regression) for FS1–FS4 feature settings.

Table 11: Performance comparison between baseline and human-factor representation settings.

Feature	AUC
FS3 baseline	0.9871
FS4 (Full + P×F gate)	0.9857

6.2 Comparison across Feature Settings (FS1–FS4)

This section compares classification performance across feature settings (FS1–FS4) under an identical evaluation protocol to examine how input information affects detection capability. As shown in Table 10 and Fig. 5, ROC-AUC generally increases as available content expands from subject-only (FS1) to full text (FS3), indicating that additional semantic context can improve classification performance. Nevertheless, FS1 alone still achieves strong results, suggesting that phishing emails often contain highly discriminative social engineering cues at the subject level, which can support early-stage alert triage under information constraints. In contrast, FS3 yields the highest baseline performance (AUC = 0.9871), reflecting the benefit of full contextual information in reducing misclassification and improving model stability. Fig. 5 further compares all feature settings (FS1–FS4) under the same five-fold stratified cross-validation protocol. While FS3 achieves the highest baseline discriminative capability, the FS4 configuration—where the classifier output is modulated by the proposed P×F human-factor mediation gate—maintains comparable performance (AUC = 0.9857). This result indicates that incorporating human-factor mediation preserves the baseline discriminative capability while enabling governance-oriented risk interpretation for SOC analysis.

6.3 Robustness Evaluation Using LLM-Generated Emails

With the widespread adoption of large language models (LLMs), attackers can now generate highly coherent and contextually realistic phishing emails at scale. To evaluate whether the proposed human-factor representation remains robust under this emerging threat landscape, an external dataset consisting of 600 balanced phishing and benign emails was generated using a scripted Python-based pipeline with GPT-4o-mini. The generation process followed predefined phishing and benign templates designed to simulate realistic social engineering scenarios and legitimate organizational communications. The generated emails were automatically assigned phishing or benign labels according to the predefined generation templates and were subsequently reviewed for consistency before being used exclusively for external robustness evaluation under distribution shift. They were not involved in model training, parameter tuning, or feature engineering, simulating recent social engineering trends characterized by high semantic coherence and contextual realism [3,4]. This test is not intended to achieve state-of-the-art performance but to assess the stability of human-factor-oriented representations under distribution shift [5]. Fig. 6 compares ROC-AUC performance of different feature settings (FS1–FS3) between in-domain five-fold stratified cross-validation on real-world emails and external testing on LLM-generated emails. The results indicate that the TF-IDF + Logistic Regression baseline preserves robust discriminative capability under the distribution shift introduced by LLM-generated emails, without evidence of structural performance failure. This suggests that the proposed baseline remains suitable for practical SOC deployment under evolving language styles [5,8].

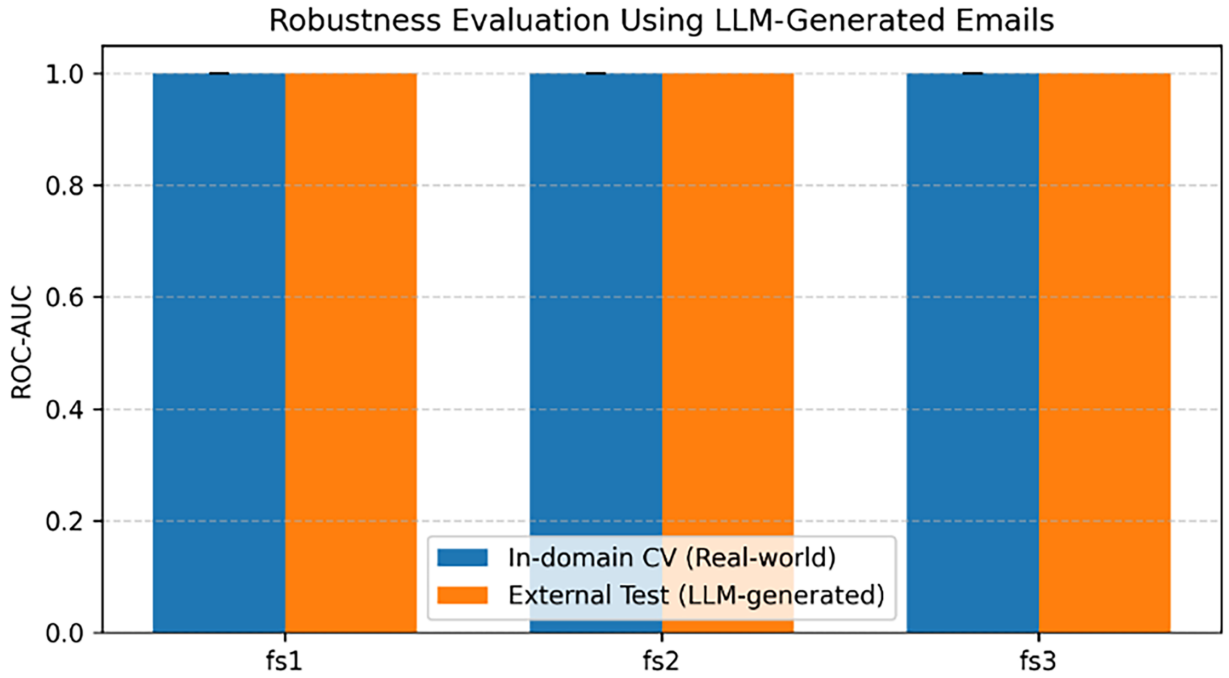

Figure 6: Robustness evaluation on LLM-generated phishing and benign emails.

The corresponding numerical results are summarized in Table 12, which reports the baseline performance obtained within the independent robustness-evaluation pipeline together with the corresponding external evaluation on LLM-generated emails. The in-domain baseline is included solely for ΔAUC calculation under the robustness-testing framework. As shown in the table, FS1 (Subject-only) and FS2 (First 200 characters) exhibit only marginal ROC-AUC changes under external testing, with ΔAUC values of −0.0144 and +0.0028, respectively, indicating highly consistent discriminative performance despite shifts in language

style and narrative structure. These results suggest that even when relying solely on email subjects or limited textual content, the model can still capture social engineering-related psychological manipulation cues that remain stable across language generations, demonstrating strong linguistic-level robustness.

Table 12: Robustness evaluation on LLM-generated emails.

Feature Set	Description	Baseline AUC (Robustness Pipeline)	External AUC (LLM-Generated)	ΔAUC
FS1	Subject-only	0.9946	0.9805	−0.0144
FS2	First 200 characters	0.9958	0.9986	+0.0028
FS3	Full text	0.9978	0.9417	−0.0561

Note: Robustness evaluation was conducted using the TF-IDF + Logistic Regression baseline with fixed model architecture and hyperparameters across all experiments. ΔAUC denotes the ROC-AUC difference between the external test on LLM-generated emails and the corresponding in-domain five-fold cross-validation results.

In contrast, FS3 (Full text) shows a more noticeable ROC-AUC decrease under external testing (ΔAUC = −0.0561), indicating that models heavily dependent on full contextual information are more sensitive to changes in language generation mechanisms and narrative styles. This effect can be reasonably attributed to distributional differences between LLM-generated emails and early real-world corpora in terms of coherence and syntactic naturalness. Nevertheless, even under this more challenging condition, FS3 maintains a ROC-AUC above 0.94, suggesting a predictable performance degradation rather than a structural failure. The external ROC-AUC values reported in Table 12 indicate that the TF-IDF + Logistic Regression baseline maintains robust discriminative capability under external testing despite the distribution shift introduced by LLM-generated emails. The reported ROC-AUC values support the robustness of the proposed framework under evolving language styles; however, they do not by themselves provide evidence regarding the balance between false positives and false negatives at a specific operating threshold. It is important to emphasize that the LLM-based evaluation serves as a complementary analysis to assess long-term effectiveness under distribution shift and evolving attack content, rather than replacing the primary experiments based on real-world datasets [20].

It should be noted that the in-domain AUC values reported in Table 12 were recomputed within an independent robustness-evaluation pipeline developed specifically for the LLM-generated email experiments. These values serve only as the reference baseline for ΔAUC calculation within the robustness analysis and are therefore intentionally independent of the primary five-fold cross-validation results reported in Table 10. Table 10 reports the primary experiments conducted on the real-world email corpora, whereas Table 12 reports the baseline and external evaluation results obtained within the independent robustness-testing framework for the LLM-generated email experiments.

6.4 Methodological Choice and Design Rationale for the BERT Benchmark Classifier

The BERT benchmark was implemented using the pretrained BERT-base-uncased checkpoint together with the corresponding Hugging Face tokenizer. Standard supervised fine-tuning was performed under the same five-fold stratified cross-validation protocol as the TF-IDF baseline to ensure methodological consistency. The maximum sequence lengths were set to 64, 256, and 512 tokens for FS1, FS2, and FS3, respectively. Training was conducted for two epochs with a learning rate of 2×10^{-5}, using a batch size of 16 for training and 32 for evaluation under a fixed random seed (42). No additional threshold calibration or domain-specific adaptation beyond the standard supervised fine-tuning procedure was applied. BERT was included as a strong semantic benchmark to contextualize the performance of the proposed framework, given

its widespread use as a high-performing baseline in phishing detection research [7,13,32]. Table 13 compares the external evaluation performance of the BERT benchmark under different text availability settings. For reference, the TF-IDF baseline AUC values reproduced from Table 12 are included in the first column solely to facilitate comparison with the corresponding BERT external evaluation results.

Table 13: BERT Comparison of classification performance under different language availability ranges (including external testing).

Mode	TF-IDF Baseline AUC (Table 12)	External_Auc	External_Precision	External_Recall	External_F1
FS1	0.9946	0.9801	1	0.03	0.0582
FS2	0.9958	0.9985	0.9676	0.9966	0.9819
FS3	0.9978	0.9416	0.9950	0.6633	0.796

Note: The TF-IDF baseline AUC values shown in the first column are reproduced from Table 12 for reference only. The primary purpose of Table 13 is to compare the external evaluation performance (ROC-AUC, Precision, Recall, and F1-score) of the BERT benchmark under different text availability settings with the corresponding TF-IDF baseline.

Under the subject-only setting (FS1), BERT attains an in-domain AUC of 0.9946, yet its external recall collapses to 0.03 despite perfect precision, indicating severe false-negative risk that is operationally unacceptable for early-stage SOC triage. This behavior demonstrates that high aggregate accuracy does not necessarily translate into stable or controllable detection behavior under distribution shift and constrained semantic input [5,8]. Fig. 7 further illustrates this phenomenon by contrasting in-domain and external ROC-AUC across different text availability settings, showing that performance degradation under distribution shift is most pronounced when semantic input is highly compressed. While richer textual context in FS2 and FS3 partially restores the precision–recall balance, the FS1 result highlights a fundamental limitation: model sophistication alone does not guarantee operational usefulness in high-risk environments such as healthcare Security Operations Centers (SOCs) [6,30].

Accordingly, BERT is treated in this study as an upper-bound semantic benchmark rather than a deployable primary detector, reinforcing the methodological choice to prioritize stable, interpretable baselines that better support auditability, traceability, and governance-oriented deployment [15,16,31]. Because the BERT benchmark was evaluated under a standardized implementation without additional threshold calibration or domain-specific adaptation beyond standard supervised fine-tuning, the reported results should be interpreted as comparative benchmark outcomes rather than fully optimized transformer performance.

Fig. 7 Comparison of in-domain and external ROC-AUC for the BERT-based classifier under different text availability settings. While in-domain performance is saturated, external testing with LLM-generated emails reveals sensitivity to distribution shift, particularly under low-information input.

6.5 PoC Assessment of Hospital A's SOC Alarm Prioritization (Top-K)

This section presents a proof-of-concept (PoC) assessment of the proposed framework using Hospital A as a representative deployment scenario. Rather than focusing on classification accuracy, the following subsections evaluate the operational feasibility, analyst workflow, alert prioritization, and practical deployment characteristics of the proposed SOC sidecar architecture.

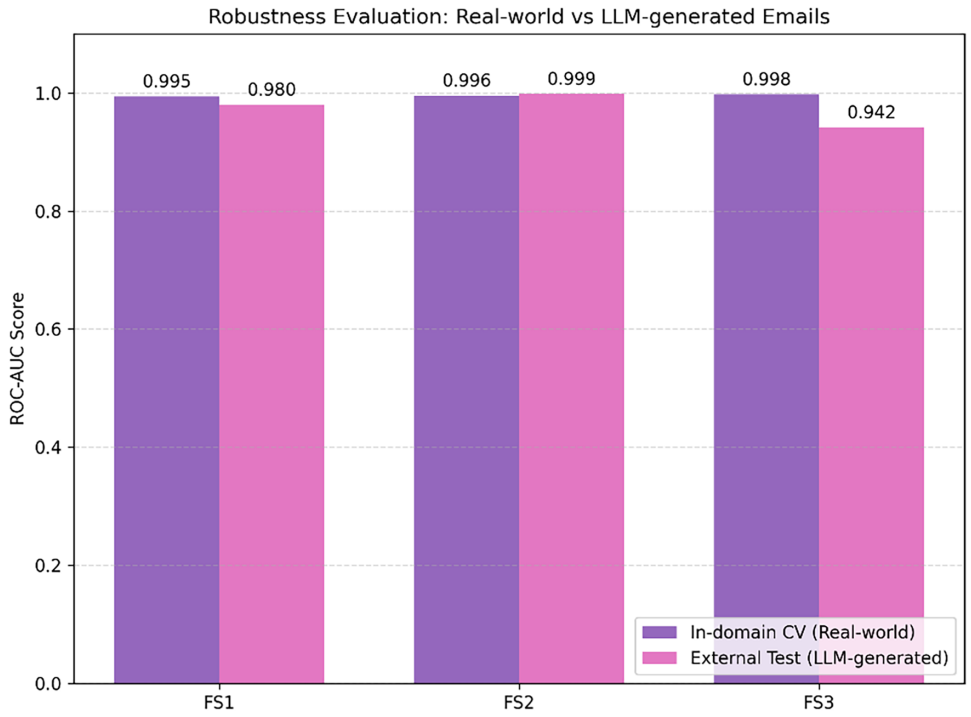

Figure 7: Comparison of BERT performance between real-world emails and LLM-generated emails.

It should be noted that the Hospital A labels are operational proxy labels rather than verified phishing ground truth. Actions such as "separate" and "rename" reflect analyst-driven handling decisions within routine SOC workflows and may therefore contain analyst or workflow-related bias. Emails with no recorded action are retained as unlabeled rather than assumed benign. Accordingly, Hospital A PoC is interpreted as a weak-label ranking evaluation focused on SOC risk prioritization and analyst attention allocation, rather than as a fully supervised phishing-detection benchmark.

6.5.1 Proof of Concept Objectives and Deployment Background

This proof-of-concept (PoC) study evaluates the feasibility and practical value of the proposed human-factor-oriented phishing risk scoring and Top-K prioritization mechanism without interfering with existing email delivery or SOC operations (using Hospital A as an example). This design addresses prior findings that phishing has become a dominant source of human-factor cybersecurity risk, particularly in high-sensitivity domains such as healthcare [3]. As illustrated in Fig. 8, no additional manual labeling or rule-based filtering was introduced during prioritization; all rankings were generated solely from model-derived risk scores. Each SOC action can be traced back to the corresponding ranked email and score, supporting post-hoc verification and auditability.

As illustrated in Fig. 8, the proposed PoC architecture introduces human-factor-oriented risk scoring and prioritization into a healthcare SOC without interfering with existing email delivery or protection workflows. All emails first pass through Hospital A's production mail system and Secure Email Gateway (SEG) for rule-based filtering, signature checks, and known-malware detection [1,5]. A sidecar design is adopted to extract only metadata and subject text, minimizing intrusion while preserving sufficient linguistic cues for preliminary risk assessment [6,8]. Within the sidecar layer, a TF-IDF + Logistic Regression

model produces phishing risk scores, while a P×F module translates embedded psychological manipulation cues into interpretable risk rationales grounded in persuasion theory and phishing-susceptibility research [9,10,13]. Emails are subsequently ranked according to their calculated risk scores to generate daily Top-K review lists under SOC workload constraints, and operational actions recorded by analysts are used as weak labels for PoC evaluation [15,16].

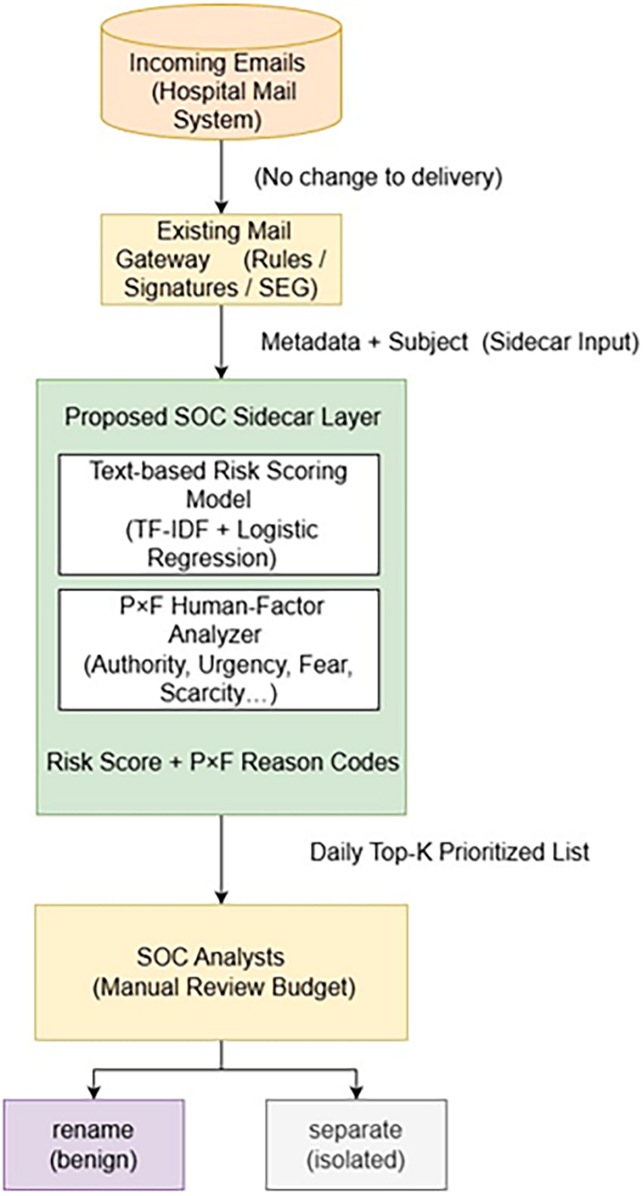

Figure 8: Email risk scoring and Top-K prioritization process in a healthcare SOC context.

In healthcare SOC environments, security analysts often face substantial alert volumes and limited investigation capacity. Under such operational constraints, alert handling is commonly implemented through risk-prioritized investigation queues rather than formal optimization-based task assignment models. Therefore, the proposed Top-K prioritization mechanism serves as an operational alert prioritization approach by determining the order in which analysts investigate phishing alerts. Rather than allocating

personnel or solving resource optimization problems, the framework focuses on investigation-order scheduling, ensuring that the most suspicious and potentially harmful emails are reviewed first. This prioritization strategy supports efficient analyst workload management, reduces the likelihood of high-risk alerts being overlooked, and aligns with governance-oriented risk management practices in healthcare cybersecurity operations [15,16,32].

6.5.2 SOC Data and Weak Label Definition for Hospital A

This section describes the structure of real-world email handling data collected from Hospital A's SOC and formalizes its methodological positioning as weak labels. As shown in Fig. 9 and Table 14 labels are derived from analysts' operational actions in routine workflows, serving as proxy supervision signals rather than exhaustive ground-truth annotations for all emails. Accordingly, emails without explicit actions are retained as unlabeled rather than assumed benign, preventing systematic bias under highly imbalanced and resource-constrained SOC conditions [15,18].

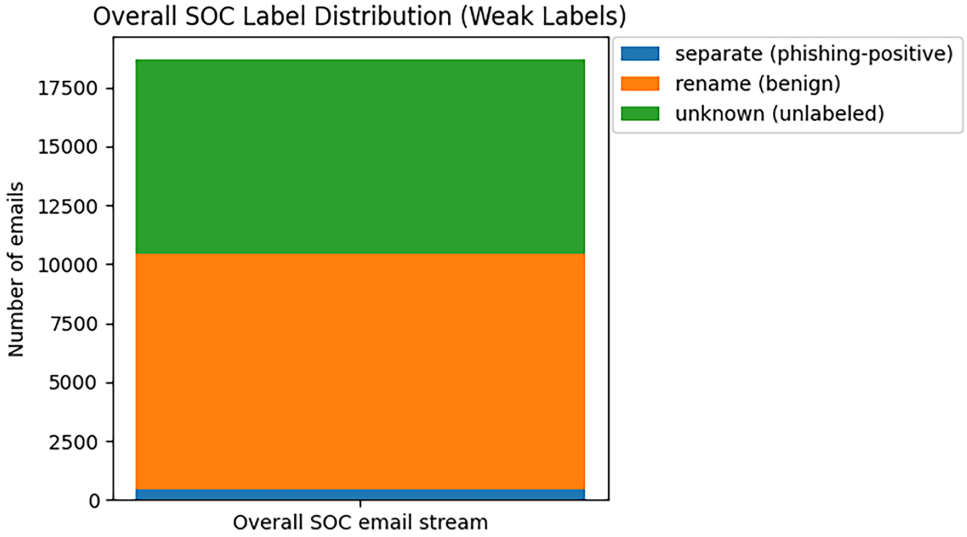

Figure 9: Overall SOC email distribution under weak label definitions.

Table 14: Correspondence between SOC actions and research agent tags.

SOC Action	Research Interpretation
Separate	Phishing-positive (proxy)
Rename	Benign (proxy)
None	Unlabeled

Fig. 9 presents a stacked-bar overview of the SOC email stream under the weak-label setting, illustrating the practical reality of partial, non-exhaustive supervision. Specifically, rename (benign proxy) constitutes the largest share, reflecting that most emails follow low-risk archival workflows, consistent with the typical imbalance where benign emails vastly outnumber phishing attempts [5]. In contrast, separate (phishing-positive proxy) accounts for a much smaller portion, corresponding to emails explicitly isolated due to high risk; although sparse, these cases carry disproportionate operational cost and governance relevance [20]. A substantial unknown/unlabeled portion further reflects SOC resource constraints and

risk-driven workflows, where only a subset of emails receives explicit analyst actions rather than confirmed benign status [15]. Such partially labeled data are not a data-quality limitation but a structural characteristic of real SOC operations, widely assumed in threat intelligence and incident response frameworks [18,20]. Consequently, subsequent evaluations in this study do not target fully supervised classification accuracy, but instead assess whether the proposed risk scores can effectively support Top-K alert prioritization and analyst decision-making under limited labeling and workload constraints [15].

Table 14 summarizes the correspondence between SOC practical actions and research labels to ensure that weak label definitions are methodologically verifiable, reproducible, and auditable. The operation is as follows:

Table 14 summarizes the correspondence between operational SOC actions and the proxy labels adopted in this study, ensuring that the weak-label definition remains verifiable, reproducible, and auditable. Emails explicitly isolated or rerouted (separate) indicate analyst judgment of elevated phishing risk and are therefore treated as phishing-positive proxies, following a deliberately conservative positive-labeling strategy that assigns positives only when SOC intervention occurs, thereby reducing systematic bias from unprocessed emails [16]. In contrast, emails that are renamed or archived (rename) are regarded as benign proxies, reflecting low-risk judgments in routine workflows, while acknowledging that these remain action-derived proxies rather than exhaustive ground-truth annotations [17]. Emails with no recorded action (none) are retained as unlabeled to avoid conflating "not processed" with "confirmed benign" under highly imbalanced and resource-constrained SOC conditions [18]. This mapping explicitly delineates which operational signals can serve as supervision and which must remain unlabeled, thereby supporting ranking-oriented evaluation with methodological consistency and interpretability [15].

6.5.3 Risk Score Definition and Top-K Priority Ranking Mechanism

In this study, the risk score is defined as the phishing probability predicted by the classifier for each email and is used as a continuous indicator of relative threat severity and ranking priority, a practice widely adopted in phishing detection research [5,6]. Higher risk scores indicate greater urgency for handling. In operational settings, emails are ranked in descending order of risk score, and only the Top-K emails are selected for manual review by SOC analysts. The parameter K reflects the daily alert-handling capacity, constrained by staffing, incident response workflows, and alert fatigue, and is explicitly considered in security governance frameworks [15–17]. Such ranking-based alert triage is a common approach to balancing detection effectiveness and operational feasibility under limited resources [5,15,20].

6.5.4 Overall Ranking Performance (Precision@K, Recall@K)

This section evaluates the proposed risk scoring and Top-K prioritization mechanism in the operational Security Operations Center (SOC) environment of Hospital A. Given limited daily analyst capacity, different K values are used to represent varying workload scenarios, and performance is analyzed from both cross-day average behavior and overall risk coverage, thereby assessing operational stability and usability [15,18]. Table 15 reports the mean *Precision@K* and *Recall@K* with standard deviations over 12 operational days, where *Precision@K* denotes the proportion of phishing-positive emails among the Top-K ranked emails reviewed by analysts, and *Recall@K* denotes the proportion of all phishing-positive emails successfully captured within the Top-K ranked emails. The results show that increasing K leads to lower *Precision@K* but higher *Recall@K*, reflecting the expected trade-off whereby broader inspection improves coverage of risky emails at the cost of increased false positives. Importantly, this pattern remains stable across days, indicating robust real-world performance of the ranking mechanism [5].

Table 15: Average daily Precision@K and Recall@K results across the observation period.

K	Precision@K (mean ± std)	Recall@K (mean ± std)	Days
20	0.646 ± 0.238	0.334 ± 0.078	12
50	0.425 ± 0.148	0.549 ± 0.110	12
100	0.262 ± 0.119	0.651 ± 0.109	12

The second analysis, summarized in Table 16, illustrates the overall coverage of high-risk emails under fixed Top-K review capacity constraints based on the full-period labeling sample. This analysis quantifies how many quarantined phishing emails can actually be captured when the SOC is only capable of reviewing the highest-ranked K emails, thereby reflecting the cumulative protection benefit provided by the prioritization mechanism under realistic resource-constrained conditions. This evaluation approach is consistent with risk-oriented cybersecurity governance practices that prioritize resource allocation efficiency and incident coverage under operational constraints [20].

Table 16: Cumulative Top-K coverage results aggregated over the entire observation period.

K	Precision@K	Recall@K	Positives in Top-K	Total Positives	Labeled Emails
20	0.95	0.0416	19	457	10,463
50	0.94	0.1028	47	457	10,463
100	0.95	0.2079	95	457	10,463

Unlike Table 15, which reports day-level average Precision@K and Recall@K across 12 operational days, Table 16 reports cumulative performance over the entire labeled observation period. Therefore, the Precision@K values in the two tables are calculated using different aggregation schemes and should not be interpreted as equivalent performance estimates or compared directly.

6.5.5 Reduced Daily Analysis and Analyst Workload

This analysis evaluates whether the proposed ranking mechanism can consistently support practical SOC workflows under realistic workload constraints rather than merely achieving favorable aggregate performance metrics [6].

Fig. 10 illustrates the daily-average Precision@K and Recall@K under different Top-K settings (K = 20, 50, 100). Smaller K yields higher Precision@K, enabling analysts to focus on higher-risk emails under limited capacity, whereas larger K substantially improves Recall@K by expanding risk coverage. This pattern reflects the typical trade-off between analyst workload and coverage in SOC operations and supports flexible K selection aligned with staffing and risk tolerance [18,20].

6.5.6 Case Study: High-Risk P×F Patterns Identified in Practice

This section presents representative Top-K high-risk email cases to illustrate the practical outputs of the proposed ranking mechanism in Hospital A's SOC. Table 17 lists emails sorted in descending order according to the model-generated risk scores, including the rank, truncated email subject, risk score, risk level, and identified P×F indicators. The risk scores support alert prioritization under limited review capacity, while the presented cases provide an interpretable view of how high-priority emails are identified in operational environments. The table further provides a traceable linkage between risk scoring, ranking decisions,

and SOC operational responses, thereby demonstrating the practical applicability and auditability of the proposed Top-K prioritization mechanism in real-world SOC workflows. Not all high-risk emails necessarily trigger predefined P×F indicators. Some instances receive high risk scores primarily due to phishing-related linguistic patterns captured by the underlying text classifier, whereas P×F indicators provide additional human-factor interpretability when persuasion-related cues, such as urgency, fear, or scarcity, are present. Therefore, the P×F layer should be interpreted as an explainable governance-oriented enhancement rather than a mandatory condition for phishing identification.

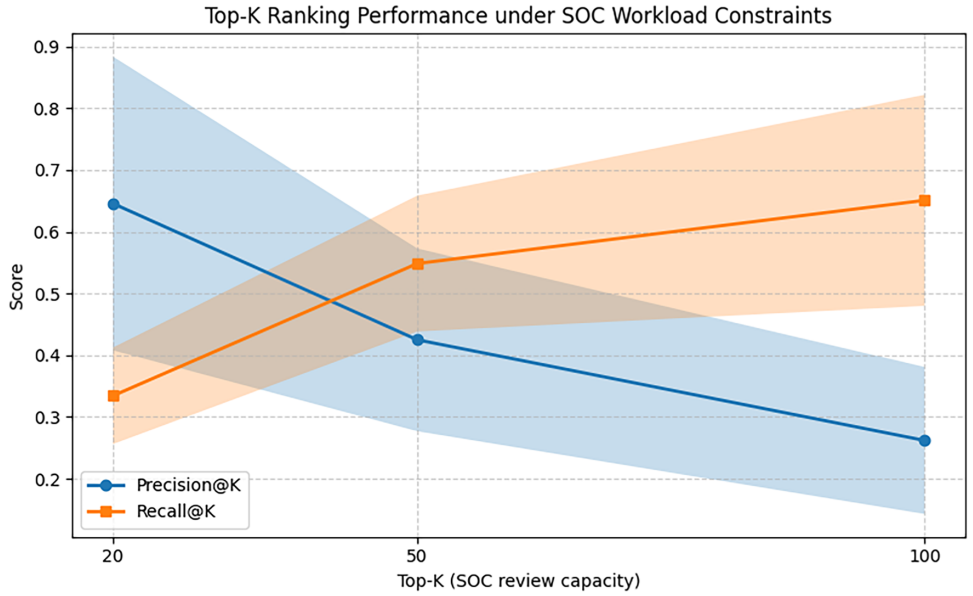

Figure 10: Daily Precision@K and Recall@K performance under SOC workload constraints.

Table 17: Prioritized high-risk email instances identified in the Hospital A proof-of-concept study.

Rank	Email Subject (Truncated)	Risk Score	Risk Level	P×F Indicators
1	Dear [redacted], Your email account is outdated...	0.9969	High	–
2	Please confirm your email address...	0.9964	High	Fear
3	Mailbox quota notification...	0.9962	High	–
4	Email account server congestion...	0.9961	High	–
5	Policy violation—immediate action required...	0.9960	High	–
6	Failed email deliveries...	0.9951	High	–
7	Account scheduled for disconnection...	0.9945	High	Urgency
8	Account has been disabled...	0.9939	High	Urgency
9	Your email account will expire...	0.9936	High	Urgency; Scarcity
10	Mailbox lost connection to server...	0.9936	High	–
11	Account flagged as spam sender...	0.9934	High	Fear
12	Webmail server pending messages...	0.9934	High	Scarcity
13	Password set to expire today...	0.9929	High	–
14	Low security verification detected...	0.9914	High	–
15	Shutdown alert—verify your account...	0.9913	High	Scarcity

(Continued)

Table 17 (continued)

Rank	Email Subject (Truncated)	Risk Score	Risk Level	P×F Indicators
16	Account transmitting viruses...	0.9902	High	Urgency
17	Unusual activities detected...	0.9891	High	–
18	Some mails are not reaching your mailbox...	0.9890	High	Urgency
19	Account will be suspended...	0.9889	High	Fear
20	Account password expires today...	0.9887	High	Urgency; Scarcity

Note: The subjects were anonymized, truncated, and partially redacted to protect sensitive information associated with healthcare and defense-related operational environments. Empty P×F indicators indicate that no predefined persuasion-related cues were triggered, although the emails may still exhibit strong phishing characteristics captured by the underlying classifier.

6.5.7 Experimental Verification Results

Taken together, the empirical analyses and the proof-of-concept (PoC) study conducted in Hospital A's Security Operations Center (SOC) demonstrate that the proposed risk-scoring and Top-K prioritization mechanism can translate text-based phishing detection outputs into actionable SOC decision-support without requiring additional manual labeling or rule-based intervention. Validation on real SOC email streams shows that, under limited analyst capacity, the approach consistently concentrates high-risk emails within a small review set, thereby improving triage efficiency and information density. Importantly, many unlabeled emails reflect routine SOC operations rather than verified benign cases. By combining proxy labels with a ranking-oriented evaluation framework, the design avoids misusing incomplete labels in conventional classification metrics and instead focuses on operational decision-support. These findings indicate that, in weakly labeled and resource-constrained SOC environments, risk-score-driven prioritization provides greater practical value than optimizing classification accuracy alone. Under a fully reproducible experimental protocol, the framework is validated from four complementary perspectives: (1) baseline performance on real-world corpora, (2) the effect of human-factor mediation, (3) robustness under distribution shift and deep-model controllability risks, and (4) Top-K operational feasibility in a real healthcare SOC (Hospital A).

First, on real email corpora with fixed five-fold stratified cross-validation, the TF-IDF + Logistic Regression baseline achieves consistently high ROC-AUC across all text availability settings: FS1 (subject-only) AUC_mean = 0.9829 (AUC_std ≈ 0.0030), FS2 (first 200 characters) = 0.9805 (≈0.0031), and FS3 (full text) = 0.9871 (≈0.0020) Table 10. The performance difference across FS1–FS3 remains small (ΔAUC ≈ 0.0042), indicating that phishing-related linguistic cues can be reliably captured even under highly constrained information conditions, while expanded text scope provides only marginal improvement. Second, the proposed human-factor mediation configuration (FS4) introduces the P×F behavioral gate applied to the full-text classifier output. As shown in Table 11, FS4 achieves an AUC of 0.9857 compared with the FS3 baseline of 0.9871. The small difference indicates that incorporating persuasion and persona signals does not materially degrade classification capability while enabling interpretable human-factor risk representation for governance-oriented SOC analysis. Accordingly, FS4 should be interpreted as an operational decision-support enhancement rather than a classification-performance enhancement. Third, external evaluation using LLM-generated phishing emails demonstrates predictable robustness under distribution shift rather than systemic failure: FS1 external AUC = 0.9805 (ΔAUC = −0.0144), FS2 = 0.9986 (+0.0028), and FS3 = 0.9417 (−0.0561) Table 12. Models relying heavily on full contextual information appear more sensitive to generative language styles, whereas subject-level or early-text features remain comparatively stable. Fourth, although BERT achieves near-saturated in-domain AUC (≈0.9946–0.9978), it exhibits severe

precision–recall imbalance under external testing, particularly for FS1 (precision = 1.00, recall = 0.03, F1 = 0.0583) Table 13. This result highlights that high AUC alone does not guarantee SOC-operable performance, as recall collapse introduces unacceptable miss risk. Consequently, adopting TF-IDF + Logistic Regression as the core model, with BERT used only as a semantic reference benchmark, reflects a deliberate deployment- and governance-oriented design choice rather than metric optimization.

Finally, the PoC study at Hospital A confirms that Top-K ranking can translate model scores into auditable SOC actions without additional labeling or rule-based intervention. Across multiple days, Precision@K–Recall@K trade-offs remain stable (Table 15). Overall coverage results show that Top-20, Top-50, and Top-100 capture 19, 47, and 95 of 457 positives, respectively, with Precision@K ≈ 0.94–0.95 Table 16. These results demonstrate traceable, reviewable, and governance-aligned alert prioritization under realistic SOC resource constraints. FS1–FS4 denote TF-IDF + Logistic Regression feature configurations, while BERT is reported separately as a semantic benchmark.

7 Conclusion

The proposed framework should therefore be viewed as a governance-oriented human-factor representation layer operating after email detection, rather than as a replacement for conventional phishing detection engines. Instead, it complements existing email security controls by providing interpretable human-factor risk prioritization for SOC decision-support. By generating continuous risk scores, auditable ranking logic, and interpretable P×F human-factor rationale codes. This supports alert prioritization and control assessment aligned with ISO/IEC 27001 auditing requirements and the risk-informed detection principles of the NIST Cybersecurity Framework. As a post-detection layer, the framework requires no modification to existing email gateways or detection pipelines, enabling practical deployment while maintaining an auditable link between emails, risk scores, and psychological risk drivers. Methodologically, phishing defense is examined across three dimensions: model behavior, human-factor feature design, and SOC operational constraints. A TF-IDF + Logistic Regression baseline provides a lightweight semantic model with stable ROC-AUC under constrained information settings (FS1–FS3 ≈ 0.98–0.99), supporting early-stage triage when only partial content is available. The P×F module complements this baseline by encoding persuasion tactics and persona-framing factors defined in the proposed P×F taxonomy, addressing sender-trust failures and Business Email Compromise (BEC)-style attacks that often bypass blacklist-based controls. External evaluation with LLM-generated emails shows bounded performance degradation (ΔAUC within −0.06), indicating robustness under distribution shift.

Operational analysis further shows that high AUC alone does not ensure SOC usability. Although BERT achieves near-saturated in-domain AUC, it suffers severe recall collapse under low-context inputs. In contrast, a Top-K proof-of-concept on 10,463 operational emails converts risk scores into inspection budgets, capturing 95 high-risk cases at K = 100 with stable Precision@K. This establishes an auditable pipeline linking model outputs, alert prioritization, and SOC decisions, demonstrating the value of human-factor-aware risk modeling for governance-oriented cybersecurity decision making.

This study contributes a decision-support-oriented framework for SOC risk prioritization with the following contributions:

(A) Deployable post-detection architecture: The framework operates as a lightweight decision-support layer that can be integrated into existing SOC pipelines without modifying upstream detection infrastructure.

(B) Operational Top-K prioritization: The risk score plus Top-K paradigm quantifies analyst workload–coverage trade-offs under limited review capacity.
(C) Human-factor risk amplification modeling: The proposed P×F framework models the interaction between persuasion tactics and persona framing, formalizing how psychological manipulation signals amplify base technical risk and translating them into interpretable governance-oriented risk indicators.
(D) Stable low-information baselines: Experiments demonstrate consistently high ROC-AUC under subject-only and partial-text conditions, supporting early-stage alert triage when message context is limited.
(E) Distribution-shift robustness: External evaluation using LLM-generated emails shows bounded and predictable performance degradation rather than systemic model failure.
(F) Weak-label operational evaluation: SOC action logs provide proxy labels that link model risk scores to real operational prioritization decisions.

Future research may investigate the integration of more advanced machine-learning architectures within the proposed P×F human-factor risk prioritization framework. Although TF-IDF and Logistic Regression were selected in this study because of their interpretability, low computational overhead, auditability, and suitability for real-world SOC deployment, transformer-based models, graph-based learning approaches, and large language models may be explored in future work to further enhance semantic understanding. Importantly, the proposed P×F framework is model-agnostic and can therefore be integrated with future detection architectures while preserving governance-oriented explainability and operational usability.

Limitations include proxy-label dependency, single-site PoC scope, and the bounded realism of synthetic emails. Future work will extend cross-institutional validation, weak or semi-supervised labeling, translation of P×F outputs into governance artifacts, and dynamic inspection budget allocation with multimodal evidence integration.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Chia-Nan Wang and Tsei-Hsuan Chen; methodology: Chia-Nan Wang, Tsei-Hsuan Chen and Syuan-Yun Wang; software development, data collection, formal analysis, and manuscript preparation: Tsei-Hsuan Chen; validation and manuscript review: Syuan-Yun Wang; supervision and manuscript review: Chia-Nan Wang and Chung-Nan Cheng. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data supporting the findings of this study are available from the corresponding author upon reasonable request. Some data cannot be made publicly available due to data availability restrictions.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Formal Risk Amplification and Matrix Construction

Appendix A.1 Persona-Weighted Risk Amplification and Smoothed P×F Matrix Construction

Building upon the base risk score $\boldsymbol{Risk_{base}}$, this study further incorporates Persona factors to capture the amplifying effects of combined *message framing* and *psychological persuasion strategies*, yielding a human-factor-aware risk score, denoted as $\boldsymbol{Risk_{withF}}$. Let Persona_Sum denote the cumulative persona evidence

score aggregated across the seven predefined Persona categories. -F_Trust, F_Prestige, F_Power, F_Alarm, F_Passion, F_Mystique, and F_Rebellion. A semi-quantitative modifier is then applied to adjust the base risk score as follows [32]:

$$Risk_{withF} = Risk_{base} \times (1 + min(0.4, 0.1 \times Persona\_Sum))$$

Note: The modifier and cap are operational hyperparameters introduced to ensure ranking robustness and prevent excessive score inflation from multiple Persona matches. Since the base risk $\boldsymbol{Risk_{base}}$ = PL × E × V captures primary exposure and impact in governance-oriented risk assessment. Persona factors are treated solely as calibration modifiers. A cap of 0.4 (maximum 40% increase) is applied to preserve discriminability without allowing a few high-hit scenarios to dominate rankings, thereby maintaining governance interpretability. The cap is regarded as a tunable calibration parameter and may be refined through sensitivity analysis or expert consensus to accommodate institutional risk preferences.

Here, the operator *min* (·) is used to impose an upper bound (cap) on the modification magnitude, preventing excessive risk score inflation and ranking distortion caused by multiple Persona matches. This design ensures that the model remains base-risk-driven, with Persona factors functioning strictly as calibration modifiers rather than dominant contributors. To quantify the additional risk introduced by Persona factors, this study defines the risk amplification as follows [32]:

$$\Delta Risk = Risk_{withF} - Risk_{base}$$

Note: $\Delta Risk$ represents the incremental risk contribution attributable to Persona factors under identical exposure and task vulnerability conditions. This difference-based formulation is a common approach for isolating marginal risk effects and supports subsequent aggregation, clustering, and ranking analyses.

Under fixed exposure (E) and task vulnerability (V) conditions, this formulation directly captures the additional risk contribution arising from the co-occurrence of psychological persuasion mechanisms and Persona framing styles. To transform $\Delta Risk$ into a visualizable and operational governance output, the incremental risks are aggregated according to their associated persuasion tactic $\boldsymbol{P_j}$ and Persona factor $\boldsymbol{F_i}$. For any given combination, let $\Omega(\boldsymbol{F_i},\boldsymbol{P_j})$, denote the set of scenarios in which both factors are simultaneously activated. The corresponding P×F matrix cell is then computed as [32]:

$$Matrix(F_i, P_j) = \boldsymbol{S_{ij}} = \sum_{k \in \Omega(Fi,Pj)} \Delta Risk_k$$

Note: In this study, scenarios are mapped to discrete cells defined by (F_i, P_j) combinations, and the incremental risks within each cell are aggregated to produce a visualizable risk matrix of P×F human-factor. Such matrix-based risk aggregation is a commonly adopted interface for governance-oriented risk communication.

Furthermore, the cell value can be presented as the average risk increase:

$$\Delta Risk_{ij} = \frac{1}{n_{ij}} \sum_{k \in \Omega(Fi,Pj)} \Delta Risk_k$$

Note: Each P×F cell estimates the mean incremental risk via sample averaging, serving as a representative indicator for each psychological combination and following standard aggregated risk practices. These values support relative risk ranking and governance decisions rather than absolute risk quantification.

Where $\boldsymbol{n_{ij}} = \Omega(\boldsymbol{F_i},\boldsymbol{P_j})$ denotes the occurrence count of the corresponding combination. Since some P×F combinations may appear only a few times (e.g., $\boldsymbol{n_{ij}} = 1$), they are susceptible to domination by single extreme

cases, leading to biased matrix interpretation and unstable governance rankings. To mitigate estimation bias caused by low sample sizes, this study adopts an additive smoothing (shrinkage) estimator. Let $\mathbf{S_{ij}} = \sum_{k\in\Omega(F_i,P_j)} \Delta\boldsymbol{Risk_k}$ μ the global mean incremental risk, and α the shrinkage strength (pseudo-count). The smoothed cell estimate is then given by [34–36]:

$$\hat{\mu}_{i,j} = \frac{S_{ij} + \alpha\mu}{n_{ij} + \alpha}$$

Note: To reduce estimation instability in low-sample cells (e.g., $\boldsymbol{n_{ij}} = 1$), this study applies an additive smoothing-based shrinkage estimator, introducing a pseudo-count α to shrink cell estimates toward the global mean μ, thereby improving matrix robustness. This approach is consistent with the principles of Empirical Bayes and Bayesian shrinkage [34,36], and additive smoothing is a standard technique for handling sparse observations and small-sample bias in machine learning and statistical modeling [35].

Appendix A.2 P×F Summary of High-Risk Scenarios and Persona-Weighted Risk Results

Appendix A summarizes the representative high-risk scenarios identified by the proposed P×F (Persuasion × Persona) human-factor threat modeling framework. To balance readability and governance relevance, only the top 30 P×F scenarios ranked by risk amplification (ΔRisk) are reported. Each scenario jointly maps a persuasion strategy (P1–P5) and a persona factor (F1–F7) and presents the persona-weighted risk score (Risk_withF), the incremental risk relative to the base risk (ΔRisk), and interpretable psychological mechanisms with corresponding persona labels. The table is intended to provide a governance-oriented risk view, enabling semantic-level social engineering patterns to be directly translated into SOC alert prioritization, security awareness training focus, and control design decisions.

Table A1 details the top 30 P×F psychological combinations with the highest ΔRisk, including dominant persuasion tactics (e.g., P2-Authority, P4-Scarcity/Urgency), aggregated mechanism scores, persona hit counts, persona-weighted risk values, and ΔRisk, together with interpretable mechanism and persona annotations to support risk-informed security operations.

Table A1: P×F high-risk scenarios by risk increase (ΔRisk) (top 30).

No.	Scam Scenario (Type)	P2	P4	Mech_Sum	Pers_Sum	Risk_WithF	ΔRisk	Dominant P	Dominant F
1	Fake supervisor/hospital director instruction	1	1	3	2	72.0	12.0	P2, P4	F_Prestige, F_Power
2	Fake Hospital A system notice/password reset	1	1	2	1	55.0	5.0	P2, P4	F_Trust
3	Fake HIS/EMR maintenance announcement	1	1	2	1	55.0	5.0	P2, P4	F_Trust

(Continued)

Table A1 (continued)

No.	Scam Scenario (Type)	P2	P4	Mech_Sum	Pers_Sum	Risk_WithF	ΔRisk	Dominant P	Dominant F
4	Fake duty roster/training link	1	1	2	0	40.0	0.0	P2, P4	—
5	Fake hospital billing/overdue payment notice	1	1	2	2	48.0	8.0	P2, P4	F_Trust, F_Prestige
6	Fake IT helpdesk call	1	1	2	2	48.0	8.0	P2, P4	F_Trust, F_Alarm
7	Fake urgent request from patient/family	0	1	2	2	38.4	6.4	P4, P5	F_Alarm, F_Passion
8	Fake official account anomaly alert	1	1	2	2	28.8	4.8	P2, P4	F_Prestige, F_Alarm
9	Fake medical document download	1	1	2	0	24.0	0.0	P2, P4	—
10	Fake consent form update	1	1	2	1	26.4	2.4	P2, P4	F_Prestige
11	Fake training certificate	1	0	2	0	24.0	0.0	P1, P2	—
12	Fake subsidy application	1	1	2	1	19.8	1.8	P2, P4	F_Trust
13	Fake vendor quotation/contract attachment	1	0	1	0	12.0	0.0	P2	—
14	QR code payment/parking fee scam	0	1	1	0	12.0	0.0	P4	—
15	Fake logistics/parcel notification	0	1	1	1	9.9	0.9	P4	F_Trust
16	Fake shopping transaction	0	1	2	0	8.0	0.0	P1, P4	—
17	Fake financial customer service	1	1	2	1	8.8	0.8	P2, P4	F_Trust
18	Fake billing email	1	1	2	0	8.0	0.0	P2, P4	—
19	Fake national health insurance fee notice	1	1	2	2	9.6	1.6	P2, P4	F_Trust, F_Prestige
20	Fake LINE official account	1	1	2	0	8.0	0.0	P2, P4	—
21	Fake internal survey	0	0	1	1	6.6	0.6	P1	F_Mystique

(Continued)

Table A1 (continued)

No.	Scam Scenario (Type)	P2	P4	Mech_Sum	Pers_Sum	Risk_WithF	ΔRisk	Dominant P	Dominant F
22	Fake meeting invitation	1	0	1	0	6.0	0.0	P2	—
23	Fake academic conference	1	0	1	0	6.0	0.0	P2	—
24	Fake prize/campaign notification	0	1	1	1	4.4	0.4	P4	F_Trust
25	Fake high-return investment	1	1	4	0	4.0	0.0	P1, P2	—
26	Cryptocurrency investment scam	0	1	3	0	3.0	0.0	P1, P3	—
27	Influencer-based investment scam	1	0	2	0	2.0	0.0	P2, P3	—
28	Romance scam	0	0	2	0	2.0	0.0	P1, P5	—
29	Fake loan SMS	1	1	2	0	2.0	0.0	P2, P4	—
30	Fake lottery data submission	0	0	2	0	2.0	0.0	P1, P5	—

The coefficient and cap are intended to be institution-specific calibration parameters and may be adjusted according to organizational risk appetite, governance requirements, and operational preferences without changing the overall PF risk modeling framework.

References

1. Cohen WW. The Enron email dataset [Internet]. 2015 [cited 2026 Jan 1]. Available from: https://www.cs.cmu.edu/~enron/.
2. Champa AI, Rabbi MF, Zibran MF. Phishing email curated datasets [Internet]. 2023 [cited 2026 Jan 1]. Available from: https://zenodo.org/records/8339691.
3. Naqvi B, Perova K, Farooq A, Makhdoom I, Oyedeji S, Porras J. Mitigation strategies against the phishing attacks: a systematic literature review. Comput Secur. 2023;132(6):103387. doi:10.1016/j.cose.2023.103387.
4. Thomopoulos GA, Lyras DP, Fidas CA. A systematic review and research challenges on phishing cyberattacks from an electroencephalography and gaze-based perspective. Pers Ubiquitous Comput. 2024;28(3):449–70. doi:10.1007/s00779-024-01794-9.
5. Popescul D, Radu LD. AI in phishing detection: a bibliometric review. Front Artif Intell. 2025;8:1496580. doi:10.3389/frai.2025.1496580.
6. Tamal MA, Islam MK, Bhuiyan T, Sattar A, Prince NU. Unveiling suspicious phishing attacks: enhancing detection with an optimal feature vectorization algorithm and supervised machine learning. Front Comput Sci. 2024;6:1428013. doi:10.3389/fcomp.2024.1428013.
7. Al Tawil A, Almazaydeh L, Qawasmeh D, Qawasmeh B, Alshinwan M, Elleithy K. Comparative analysis of machine learning algorithms for email phishing detection using TF-IDF, Word2Vec, and BERT. Comput Mater Contin. 2024;81(2):3395–412. doi:10.32604/cmc.2024.057279.
8. Innab N, Osman AAF, Ataelfadiel MAM, Abu-Zanona M, Elzaghmouri BM, Zawaideh FH, et al. Phishing attacks detection using ensemble machine learning algorithms. Comput Mater Contin. 2024;80(1):1325–45. doi:10.32604/cmc.2024.051778.

9. Parsons K, Butavicius M, Delfabbro P, Lillie M. Predicting susceptibility to social influence in phishing emails. Int J Hum Comput Stud. 2019;128:17–26. doi:10.1016/j.ijhcs.2019.02.007.
10. Cialdini RB. Influence: the psychology of persuasion (new and expanded). New York, NY, USA: Harper Business, an imprint of HarperCollins Publishers; 2021.
11. Hogshead S. Fascinate: how to make your brand impossible to resist. New York, NY, USA: HarperBusiness; 2016.
12. Sturman D, Auton JC, Morrison BW. Security awareness, decision style, knowledge, and phishing email detection: moderated mediation analyses. Comput Secur. 2025;148:104129. doi:10.2139/ssrn.4883192.
13. Tooher P, Lallie HS. A two-stage deep learning framework for AI-driven phishing email detection based on persuasion principles. Computers. 2025;14(12):523. doi:10.3390/computers14120523.
14. MITRE Corporation. ATT&CK technique: phishing [Internet]. 2024 [cited 2026 Jan 1]. Available from: https://attack.mitre.org/.
15. National Institute of Standards and Technology. The NIST cybersecurity framework (CSF) 2.0 [Internet]. 2024 [cited 2026 Jan 1]. Available from: https://www.nist.gov/cyberframework.
16. ISO/IEC 27001:2022. Information security, cybersecurity and privacy protection/information security management systems/requirements. Geneva, Switzerland: ISO; 2022.
17. ISO/IEC 27002:2022. Information security, cybersecurity and privacy protection/information security controls. Geneva, Switzerland: ISO; 2022.
18. National Institute of Standards and Technology. Guide for conducting risk assessments, NIST special publication 800-30 Rev. 1 [Internet]. 2012 [cited 2026 Jan 1]. Available from: https://csrc.nist.gov/publications/detail/sp/800-30/rev-1/final.
19. ISO/IEC 27005:2022. Information security, cybersecurity and privacy protection/guidance on managing information security risks. Geneva, Switzerland: ISO; 2022.
20. European Union Agency for Cybersecurity (ENISA). ENISA threat landscape 2024 [Internet]. 2024 [cited 2026 Jan 1]. Available from: https://www.enisa.europa.eu/publications.
21. Federal Bureau of Investigation. Internet crime report (IC3) [Internet]. 2024 [cited 2026 Jan 1]. Available from: https://www.ic3.gov/Media/PDF/AnnualReport/2024_IC3Report.pdf.
22. Moura GCM, Daniels T, Bosteels M, Castro S, Müller M, Wabeke T, et al. Characterizing and mitigating phishing attacks at ccTLD scale. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security; 2024 Oct 14–18; Salt Lake City, UT, USA. p. 2147–61. doi:10.1145/3658644.3690192.
23. Health-ISAC. Health-ISAC annual report/threat intelligence (healthcare sector) [Internet]. 2024 [cited 2026 Jan 1]. Available from: https://health-isac.org/.
24. Ministry of the Interior Police Department. 165 national anti-fraud network [Internet]. 2025 [cited 2026 Jan 1]. Available from: https://165.npa.gov.tw/.
25. Al-Subaiey A, Al-Thani M, Abdullah Alam N, Antora KF, Khandakar A, Uz Zaman SA. Novel interpretable and robust web-based AI platform for phishing email detection. Comput Electr Eng. 2024;120(7):109625. doi:10.1016/j.compeleceng.2024.109625.
26. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74. doi:10.48550/arxiv.1705.07874.
27. Egelman S, Peer E. Scaling the security mindset: measuring and predicting susceptibility to phishing attacks. Comput Hum Behav. 2023;145:107754.
28. Gupta BB, Gaurav A, Arya V, Attar RW, Bansal S, Alhomoud A, et al. Advanced BERT and CNN-based computational model for phishing detection in enterprise systems. Comput Model Eng Sci. 2024;141(3):2165–83. doi:10.32604/cmes.2024.056473.
29. Bari N, Saleem T, Shah M, Algarni A, Patel A, Ullah I. A filter-based feature selection framework to detect phishing URLs using stacking ensemble machine learning. Comput Model Eng Sci. 2025;145(1):1167–87. doi:10.32604/cmes.2025.070311.
30. Scikit-learn developers. Scikit-learn: machine learning in python [Internet]. 2022–2025 [cited 2026 Jan 1]. Available from: https://scikit-learn.org/stable/.

31. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74. doi:10.1016/j.patrec.2005.10.010.
32. ISO 31000:2018. Risk management-guidelines. Geneva, Switzerland: ISO; 2018.
33. Šijan A, Viduka D, Ilić L, Predić B, Karabašević D. Modeling cybersecurity risk: the integration of decision theory and pivot pairwise relative criteria importance assessment with scale for cybersecurity threat evaluation. Electronics. 2024;13(21):4209. doi:10.3390/electronics13214209.
34. Efron B. Large-scale inference: empirical bayes methods for estimation, testing, and prediction. Cambridge, UK: Cambridge University Press; 2010. doi:10.1017/cbo9780511761362.
35. Bishop CM. Pattern recognition and machine learning. New York, NY, USA: Springer; 2006.
36. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. Boca Raton, FL, USA: CRC Press; 2013.
37. Nikbakht M, Rouhani S, Mojtahed V. A novel ranking model for information technology security controls through COBIT and MCDM. Rec Manag J. 2025;35(3):251–76. doi:10.1108/rmj-03-2024-0007.
38. Garosi A, Di Nardo M, Antonelli RG. Healthcare failure mode and effects analysis: a systematic review of recent applications and methodological advances. Saf Sci. 2023;163(1):106126. doi:10.1016/j.ssci.2023.106126.
39. Wang D, Zhou M, Lin Q. Human reliability analysis for reducing human errors in healthcare: a systematic literature review. Qual Reliab Eng Int. 2026;42(5):2638–53. doi:10.1002/qre.70177.
40. Hollnagel E, Wears RL, Braithwaite J. From Safety-I to Safety-II: a white paper. Middelfart, Denmark: Resilient Health Care Net; 2015.
41. Jalalvand F, Baruwal Chhetri M, Nepal S, Paris C. Alert prioritisation in security operations centres: a systematic survey on criteria and methods. ACM Comput Surv. 2025;57(2):1–36. doi:10.1145/3695462.
42. Tariq S, Baruwal Chhetri M, Nepal S, Paris C. Alert fatigue in security operations centres: research challenges and opportunities. ACM Comput Surv. 2025;57(9):1–38. doi:10.1145/3723158.