# A Defensive Mirroring Framework for Explainable Detection of AI-Enabled Spear Phishing

Jack Stavrakas
*Molinaroli College of Engineering and Computing*
*University of South Carolina*
Columbia, US
stavrakj@email.sc.edu

Jorge Crichigno
*Molinaroli College of Engineering and Computing*
*University of South Carolina*
Columbia, US
jcrichigno@cec.sc.edu

Neset Hikmet
*Molinaroli College of Engineering and Computing*
*University of South Carolina*
Columbia, US
nhikmet@cec.sc.edu

***Abstract*—Spear phishing driven by large language models (LLMs) and open-source intelligence (OSINT) increasingly evades traditional signature- and reputation-based email defenses. This paper presents a phishing detection system that defensively mirrors adversarial tradecraft by repurposing the same LLM-enabled reconnaissance and OSINT aggregation techniques used by attackers, integrating automated indicator extraction, OSINT enrichment, a novel social engineering (SE) extractor, and AI synthesis over raw email artifacts (.EML). The system was evaluated through a controlled user study using realistic emails from a large R1 university environment, comparing unaided detection, OSINT-only exposure, and full AI-synthesized decision support. Results show that OSINT alone yields limited gains, while AI-synthesized OSINT produces non-linear improvements in discrimination ability exceeding an order of magnitude without inducing response bias, demonstrating that contextual synthesis is critical for scalable, explainable phishing defense.**

***Index Terms*—Phishing, social engineering, OSINT, LLMs, AI, cybersecurity decision support, email security**

## I. Introduction

Phishing remains one of the most persistent and consequential cybersecurity threats [1], [2]. Cybercrime has evolved over time to exploit a key factor in the cyber kill chain that resists complete technological control: the user [3], [4]. Recently, spear phishing attacks, among the most dangerous types of phishing, have surged in volume and reduced in cost per phish [5], [6].

Existing methods to mitigate phishing attacks are primarily backend interventions that incorporate signature-based filtering, statistical learning, heuristic analysis, and email authentication protocols [2], [7]. As implemented, these filters are vulnerable to adversarial degradation [4]. Dynamic rotation of sender addresses, email infrastructure, and malicious domains reduces the effective lifetime of signatures (and corresponding reputation scores) while challenging statistical learning [7], [8]. Similarly, high-fidelity attack infrastructure automation increasingly incorporates transient certificates and configurations to evade authentication-based detection [4], [5]. Additionally, advanced social engineering (SE) tactics and OSINT increase email personalization, further reducing detection efficacy [5], [6].

As phishing attacks shift toward low-volume, high-fidelity delivery, the limiting factor in detection is no longer individual indicators but the ability to apply context to observed features [9], [10]. Existing defenses largely evaluate emails as collections of independent features such as headers, URLs, domains, or lexical cues without reconstructing the social, semantic, or operational context the message is intended to be interpreted in [7]. As a result, attacks that employ individually benign components but express malicious intent often evade detection, especially in spear phishing scenarios where context, timing, and persuasion dominate [6], [11].

This paper proposes a phishing detection system built on a defensive mirroring architecture that reproduces attacker OSINT reconnaissance and AI-driven synthesis for defensive analysis [1], [6]. The system was implemented as a Python-based pipeline comprising an EML parser, social engineering extractor, OSINT enrichment engine, and AI synthesizer, evaluated against a corpus of sanitized legitimate and phishing emails stratified by SE tactics, IoC type and density, and phishing goal [1], [6]. Implementation details are provided for replication; deployment-ready code is withheld due to misuse risk [10].

The contributions of this paper are summarized as follows:

- Designing and implementing a social engineering extractor that operationalizes a standardized threat taxonomy to classify persuasion tactics within incoming mail
- Integrating IoC extraction and multi-source OSINT enrichment through an "adversarial mirroring" design that repurposes attacker reconnaissance tradecraft for defensive assessment
- Empirically evaluating unaided detection, OSINT-only exposure, and AI-synthesized decision support using ecologically valid, raw email artifacts drawn from a large R1 university environment

## II. Background and Related Work

Modern phishing attacks have transitioned from bulk, template-based campaigns into low-volume, high-fidelity spear phishing via AI-enabled automation and OSINT reconnaissance [5], [6]. Recent work demonstrates that LLMs can autonomously perform reconnaissance, synthesize context, and generate persuasive spear phishing content that is difficult for users and automated systems to distinguish from legitimate emails [5], [6], [11]. AI-generated spear phishing emails achieve success rates comparable to those authored by human experts while simultaneously reducing the cost and effort required to produce personalized attacks [6], [10]. LLM-assisted phishing also reduces the marginal cost per attack to fractions of a cent, enabling scalable personalization previously limited by human labor [5]. While existing work demonstrates the full AI-enabled offensive cycle, including automated reconnaissance, targeting, and generation, no prior work directly mirrors this tradecraft defensively; this paper addresses that gap.

This shift undermines many assumptions underpinning existing email security systems which are optimized for high-volume, repetitive threats. As adversaries increasingly rotate infrastructure and tailor content per target, signature-based detections, reputation scores, and coarse anomaly thresholds degrade rapidly, motivating new semantic reasoning approaches applied across heterogeneous and transient indicators [6], [7].

Existing phishing detection systems rely primarily on backend analysis of email artifacts including header metadata, URLs, domains, and attachments. Common approaches include signature-based detection, reputation scoring systems, statistical learning, and hybrid methods that combine these signals. Surveys of intelligent spam and phishing detection systems show that these techniques are effective against known and high-prevalence threats but struggle with zero-day attacks, infrastructure churn, and low-volume spear phishing [1], [7].

To improve detection fidelity, security tools increasingly incorporate external threat intelligence and OSINT. Threat intelligence engines aggregate data from domain registries, sandboxing platforms, abuse databases, and network telemetry to enrich IoCs with contextual metadata. Systems such as SecBuzzer demonstrate how multi-source OSINT pipelines can transform unstructured external data into actionable intelligence for cybersecurity analysts [12]. However, most existing pipelines either expose these signals only to backend detection classifiers or require manual correlation by human analysts, limiting both scalability and timeliness [13].

In operational settings, this manual correlation incorporates technical indicators, infrastructure context, and inferred attacker intent. This process is time-sensitive and difficult to scale, motivating research into automated triage and AI-assisted analysis [6], [13]. Prior systems leverage machine learning models to classify emails or highlight salient features, but these approaches typically produce opaque scores or binary verdicts that obscure the underlying reasoning [1], [14].

Recent work explores interpretable and explainable AI-assisted security platforms that emphasize synthesis over mere classification. By aggregating IoCs and providing structured explanations, these systems aim to support decision making while reducing analyst workload [1], [13], [14]. Explainability is increasingly treated as a functional requirement rather than a usability feature, especially in security contexts where incorrect automation erodes trust and adoption [1], [14]. These findings motivate architectures that preserve raw email fidelity, enrich extracted indicators with OSINT, and synthesize results into coherent, human-readable assessments rather than relying solely on invisible, backend filtering [1], [6], [12], [14]. Unlike prior systems that stop at enrichment or produce opaque classifications, the proposed system applies full contextual synthesis to heterogeneous signals, enabling explainable, analyst-grade assessment.

## III. System Architecture

The proposed system implements a defensive mirroring analysis pipeline designed to perform reconnaissance, detection, and synthesis on incoming messages [1], [6]. The architecture is designed to automate the triage process of security analysts while preserving contextual richness that is lost in traditional backend filtering systems [6], [7], [13].

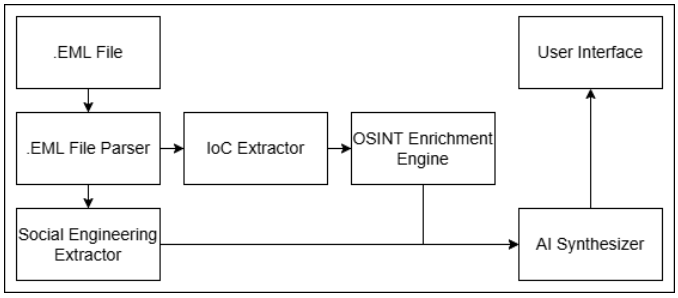


**Fig. 1:** Proposed defensive mirroring system architecture.

### A. Email Ingestion

Email ingestion is performed using raw EML files to preserve transport-level metadata, MIME structure, and authentication results. Upon ingestion, each EML file is parsed into message headers, plaintext body, HTML body, inline elements, and attachments. This decomposition enables targeted analysis of each element while maintaining a consistent representation across all emails [7]. Python-based parsing libraries are used to extract and normalize fields, producing structured JSON that serves as the input to downstream modules.

### B. Social Engineering Extractor

The SE extractor module identifies social engineering tactics embedded in email content that are commonly used in spear phishing attacks [5], [6]. This module uses a large language model (LLM) to identify persuasion strategies such as urgency induction, authority exploitation, impersonation, familiarity abuse, and incentive framing – all commonly used to elicit rapid and uncritical user responses [5], [6], [11]. Unlike prior work that treats persuasive language as unstructured text

features, this module operationalizes a standardized social engineering taxonomy, enabling tactic-level classification rather than reliance on surface-level linguistic cues.

| SE Tactic | Definition |
|---|---|
| Creating urgency | Pressuring the recipient to act immediately to avoid negative consequences. |
| Inducing fear | Threatening harm, loss, or penalty to coerce rapid compliance. |
| Engineering incentive | Offering rewards or benefits to exploit curiosity or greed. |
| Exploiting context | Leveraging situationally relevant details to increase plausibility and legitimacy. |
| Exploiting authority | Impersonating executives or supervisors to compel compliance through hierarchy. |
| Manipulating trust | Leveraging perceived legitimacy of brands, institutions, or systems. |
| Abusing familiarity | Mimicking known individuals or organizations within the recipient's social or professional sphere. |
| Impersonation | Masquerading as a specific entity via spoofed identities at the social or technical level. |

**TABLE I:** Common social engineering tactics and their operational definitions.

For each tactic detected, the LLM returns the tactic alongside representative excerpts from the email and a brief justification [1], [6]. Multiple tactics may be identified within a single email. By deriving structured, tactic-level SE patterns that are difficult to capture with surface lexical features, the SE extractor augments traditional IoCs and provides additional semantic signal for analysis [6], [7].

### C. Indicator-of-Compromise Extractor

The IoC extraction module performs deterministic pattern matching over parsed email components to identify technical indicators associated with phishing. Extracted IoCs include URLs, domains, IP addresses, sender anomalies, authentication results, attachments, and routing inconsistencies [7]. Header analysis focuses on discrepancies between user-visible sender fields and SMTP envelope data, relay path irregularities, and authentication failures. Body analysis identifies embedded URLs, anchor text mismatches, and suspicious formatting [7], [8]. Attachment analysis extracts filenames, file types, and cryptographic hashes when applicable [7]. This module transforms unstructured email artifacts into a structured set of indicators suitable for automated enrichment and correlation [12].

### D. Open-Source Intelligence Enrichment Engine

The OSINT enrichment engine augments extracted IoCs by querying multiple external OSINT sources via API interfaces. No single intelligence source is enough to determine overarching intent; thus, the system performs many lookups in parallel across domain registration services, URL scanners, reputation databases, fingerprinting services, and breach databases [7], [12]. Returned enrichments include features like domain age, registration history, DNS hygiene, redirect chains, hosting characteristics, sandbox results, and prior abuse signals [7], [8]. This enrichment process mirrors the reconnaissance phase of modern spear phishing campaigns, in which attackers similarly aggregate distributed OSINT to construct credible narratives.

### E. AI Synthesizer

The AI synthesizer module integrates all previously extracted and enriched signals into a unified assessment [1], [7]. Unlike traditional classifiers that operate on fixed feature vectors, this module receives the original email content, extracted IoCs, OSINT enrichment results, and SE indicators as structured input [5], [6]. This input is extensible and accounts for all indicators observed in each email. An LLM is used to reason across these heterogeneous inputs and produce three outputs:

- A binary classification (phishing or legitimate)
- An inferred phishing goal, if applicable
- A concise, natural-language explanation describing the rationale behind the classification

The model is constrained via prompt engineering to produce structured output suitable for presentation and evaluation. This synthesis automates the correlation and reasoning process typically performed during manual email triage, enabling scalable analysis without reducing assessment technique to opaque scoring [13], [14]. This synthesis represents the core of the adversarial mirroring design: the system reconstructs the type of contextual reasoning attackers use when crafting high-fidelity phishing content.

### F. Output

The system outputs the binary verdict recommendation accompanied by a short explanatory summary [1], [6]. Rather than blocking or silently filtering messages, the system presents its assessment alongside the email under review [14]. This explanation highlights indicators and contextual factors contributing to the decision, enabling downstream consumers to understand and evaluate the result [1], [14]. Providing actionable output is treated as a system requirement rather than a user-interface enhancement [12]. Prior work shows that opaque automation degrades trust and adoption, especially in security contexts [1], [14]. By exposing synthesized reasoning, the system supports informed decision making while mitigating over-reliance on automation [9], [14]. This user-facing mode represents one of several intended deployment configurations, ranging from fully automated backend filtering to analyst-facing synthesis, selected here to evaluate human decision support directly.

## IV. Implementation and Evaluation

This system was evaluated through a controlled, IRB-approved study titled "AI-Enhanced OSINT to Reduce Phishing Susceptibility".

### A. Study Setup

The system was evaluated using a controlled email classification study designed to measure phishing detection performance with and without system assistance [6], [9]. Emails were preserved in full .EML file format and manually sanitized to remove personally identifiable information while retaining structural, linguistic, and technical characteristics. Participants were recruited via a publicly available directory of nonprofit

organizations in South Carolina; education level was collected as a demographic covariate and used as a proxy for technical background, with counts per condition reported in Table III. Participants were presented with emails in a simulated inbox environment and asked to classify each message as either legitimate or phishing [9]. All participants completed the study asynchronously using the same interface and task framing.

### B. Conditions

The study employed a between-subjects design to prevent learning and carryover effect [6], [9]. Participants were assigned to one of the following conditions:

- Control condition, in which participants evaluated emails without any assistance
- Raw OSINT condition, in which participants were provided with the system's enrichment findings without AI synthesis
- AI-synthesized OSINT condition, in which participants were provided with the system's classification output and explanatory summary alongside each email.

Each participant evaluated the same number of emails, drawn from a balanced set of legitimate and phishing messages. Email order was randomized per participants to mitigate ordering effects [9].

### C. Metrics

System effectiveness was evaluated using Signal Detection Theory (SDT) metrics, which separate discrimination ability from response bias in binary decision tasks. Sensitivity quantified participants' ability to distinguish phishing emails from legitimate emails, while the response criterion captured conservative or liberal decision tendencies, enabling assessment of whether performance gains reflected improved discrimination rather than bias shifts.

| Metric | Definition |
|---|---|
| Hit Rate | Proportion of phishing emails correctly identified as phishing. Captures successful detection of malicious signals. |
| False Alarm Rate | Proportion of legitimate emails incorrectly classified as phishing. Reflects costs of overly conservative responses and disruption to workflow. |
| Sensitivity ($d'$) | SDT measure of accuracy. Describes discrimination ability independent of response bias. |
| Decision Criterion ($c$) | SDT measure of bias. Describes tendency to favor phishing or legitimate classifications under uncertainty. |

**TABLE II:** Signal Detection Metrics Used in Analysis

SDT metrics were selected over raw accuracy to account for base-rate effects and asymmetric error costs common in phishing detection, where malicious emails are relatively rare and false positives carry operational consequences [9]. For completeness, overall accuracy, false positive rates, and false negative rates were also computed. Odds ratios (ORs) were used to quantify the magnitude of performance differences between experimental conditions.

### D. Ecological Realism

To maximize ecological validity, the study employed realistic email stimuli rather than simplified, synthetic, or summarized representations [9]. The evaluation corpus consisted of legitimate and phishing emails collected directly from a production enterprise email environment at a large R1 university (>50,000 users), capturing authentic attacker behavior, organizational context, and defensive conditions not reproducible in synthetic datasets. Emails preserved original formatting, sender information, embedded links, and message tone, reflecting the cues users encounter in operational inboxes [7], [10]. The simulated inbox interface allowed participants to inspect emails naturally, approximating real-world workflows in which phishing detection is a secondary task rather than primary focus. This design aligns with prior findings that simplified stimuli and artificially elevated threat prevalence can overestimate user detection performance [9].

## V. Results

### A. Classification Performance

Participants' phishing classification performance differed substantially across experimental conditions. In the unaided condition, participants performed near chance, correctly classifying approximately half of all emails ($\approx 51\%$). Providing structured IoC and OSINT enrichment yielded a marked improvement in performance ($\approx 77\%$), while participants receiving AI-synthesized OSINT decision support achieved near-ceiling accuracy ($\approx 98\%$). Mixed-effects logistic regression models with random intercepts for participants and email items confirmed reliable effects of decision support. Relative to the unaided condition, OSINT exposure significantly increased the odds of a correct classification (OR $\approx$ 3.26), while AI-synthesized OSINT produced a larger effect, increasing the odds of correct classification by more than an order of magnitude (OR $\approx 50$).

### B. Signal Detection Outcomes

| Condition | n | Hit Rate | FA Rate | $c$ |
|---|---|---|---|---|
| Control | 70 | 0.503 | 0.497 | 0.000 |
| OSINT | 70 | 0.737 | 0.203 | 0.098 |
| AI | 67 | 0.991 | 0.029 | -0.242 |

**TABLE III:** SDT Response Metrics by Condition

| Condition | $d'$ | 95% CI (L) | 95% CI (U) |
|---|---|---|---|
| Control | 0.020 | -0.187 | 0.227 |
| OSINT | 1.428 | 1.277 | 1.580 |
| AI | 2.475 | 2.426 | 2.524 |

**TABLE IV:** SDT Sensitivity and CI by Condition

SDT analyses indicate that performance gains across conditions were driven primarily by improvements in discrimination ability rather than shifts in response bias. In the unaided condition, hit and false alarm rates were nearly identical,

yielding near-zero sensitivity (d' $\approx 0$), consistent with chance-level discrimination. Participants in the OSINT condition exhibited a clear improvement in response separation, with increased hit rates and reduced false alarms, corresponding to a moderate increase in sensitivity (d' $\approx 1.47$). This pattern was further amplified in the AI-synthesized OSINT condition, which exhibited a large separation between signal and noise distributions (d' $\approx 4.3$), reflecting a substantial enhancement in the ability to distinguish phishing emails from legitimate emails. Across all conditions, decision criterion values differed only modestly, indicating that improved performance was not driven by indiscriminate caution or liberal responding.

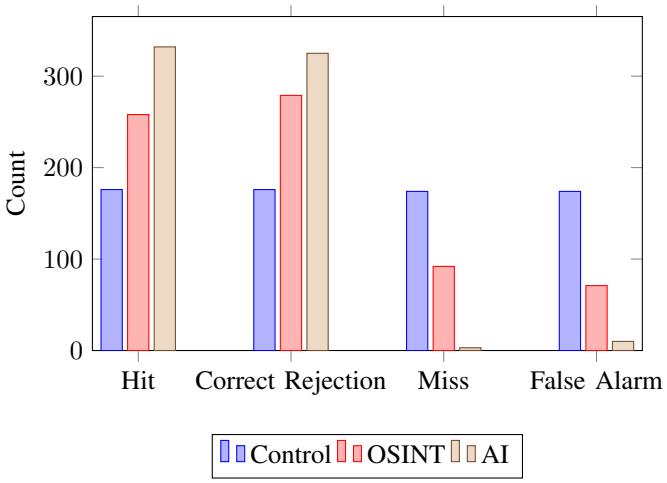


**Fig. 2:** Response outcome counts for Control and AI conditions.

Mixed-effects models decomposing signal and noise trials corroborated these findings. Relative to the control condition, both OSINT and AI-synthesized decision support significantly increased the odds of correctly identifying phishing emails while also increasing the acceptance of legitimate messages, confirming that performance improvements reflected genuine gains in discrimination rather than a trade-off between detection and false alarms.

### C. Confidence and Calibration

Decision support also influenced participants' confidence and calibration. Overall confidence increased monotonically across conditions, with the lowest confidence in the unaided condition, higher confidence with OSINT support, and the highest confidence in the AI-synthesized OSINT condition. Across all conditions, confidence was systematically higher on correct trials than on incorrect trials. Despite near-ceiling accuracy under AI-synthesized support, confidence did not inflate uniformly; confidence remained meaningfully lower on the rare incorrect decisions. This alignment between subjective certainty and objective correctness indicates improved confidence calibration rather than overconfidence and suggests that AI-synthesized OSINT enhances both decision quality and metacognitive reliability.

### D. Robustness and Generalization

Robustness analyses indicate that the observed effects generalize across participants and email characteristics. Incorporating demographic covariates did not materially alter effect estimates, nor did demographics emerge as consistent predictors of performance. Additional analyses incorporating email threat goals, social engineering (SE) tactics, and IoC complexity revealed strong main effects of decision support but no reliable interactions. Performance gains increased monotonically from unaided inspection to OSINT enrichment and to AI-synthesized decision support and were consistent across heterogeneous phishing types and complexity levels.

## VI. Discussion

### A. Implications

The results demonstrate that a defensive mirroring approach serves as an effective extension of traditional phishing defenses by improving discrimination between malicious and legitimate emails. Unlike backend filters that rely on isolated indicators or opaque scoring, the proposed system integrates heterogeneous technical and contextual signals into a synthesized assessment that is accurate and explainable. While structured OSINT enrichment alone improves performance relative to unaided inspection, the large sensitivity gains observed with AI synthesis indicate that the most substantial performance improvements arise from enhanced signal integration rather than conservative blocking or indiscriminate rejection.

From a systems perspective, these findings demonstrate the value of automated synthesis over the accumulation of raw indicators. Individual IoCs and OSINT signals are often brittle in isolation; when correlated and interpreted, however, they enable reliable classification. The AI synthesis module automates triage and reasoning tasks typically performed by security analysts, enabling scalable assessment of emails while preserving explainability and human-readable insight. This supports a shift from backend-only filtering toward hybrid defenses that expose synthesized intelligence closer to the point of user interaction. The near-ceiling performance achieved under AI-synthesized decision support suggests that contextual synthesis substantially narrows the gap between unaided human judgment and analyst-style detection under controlled conditions.

### B. Limitations

The evaluated system represents a specific instantiation of AI-synthesized OSINT using fixed prompt engineering, selected intelligence sources, and deterministic outputs. Alternative synthesis strategies, explanation formats, or model choices may yield different performance profiles, and structured OSINT alone may be implemented in different ways that affect user outcomes. Additionally, this evaluation abstracts away from operational factors such as alert fatigue, adversarial adaptation, and organizational policy constraints. As such, the findings should be interpreted as evidence of capability rather than as a claim of immediate operational replacement for existing defenses. Despite these constraints, the results establish

AI-synthesized OSINT as a viable user-facing mechanism for improving phishing detection, providing a defensible blueprint for integration alongside existing security infrastructure.

### C. Future Work

Future work will focus on extending and operationalizing the system along several complementary paths. A primary direction is decomposing the synthesis pipeline, evaluated here as a unified component, into constituent elements such as indicator extraction, synthesis strategy, and taxonomy implementation. Different LLMs also vary in reasoning depth, uncertainty expression, and error characteristics, enabling future studies to treat prompt design, model choice, and threat taxonomy as independent experimental variables.

A second direction is expanding evaluation beyond binary classification. While this study focused on correctness, operational actions such as link interaction, attachment handling, reporting, deferral, and deletion carry distinct security consequences. Future work could instrument these behaviors directly, incorporating temporal features (e.g., hesitation or reversibility) and weighted outcome models that distinguish beneficial, neutral, and harmful actions.

A further direction is conducting field studies. Operational deployment would enable assessment under realistic workload, time pressure, and adversary adaptation, while facilitating measurement of click-through rates, reporting behavior, analyst workload, and alert fatigue.

Finally, future work should evaluate the system across alternative deployment modalities: the user-facing mode assessed here, fully automated backend filtering, analyst-facing synthesis, and a logging configuration that supports organizational threat analytics. Because the system exposes synthesized reasoning rather than acting as an invisible filter, repeated exposure may alter discrimination ability, trust calibration, and reliance over time. Thus, the user- and analyst-facing modes also raise the possibility of longitudinal learning effects, wherein this exposure improves unaided detection over time.

Together, these directions move from controlled validation toward optimization, longitudinal assessment, and evaluation of real operational impact.

## VII. Conclusion

This paper presented a phishing detection system built on a defensive mirroring paradigm, repurposing the same LLM-driven reconnaissance and synthesis techniques used by attackers to enable automated, analyst-grade triage. Operating on raw, ecologically valid .EML artifacts from a live R1 university environment, the system preserves transport-level context and message structure often absent from synthetic or text-only datasets, while incorporating a Social Engineering (SE) extractor that applies a structured persuasion taxonomy alongside technical indicators and multi-source OSINT. Empirical results show that raw OSINT exposure yields only modest gains, whereas AI-synthesized decision support produces nonlinear improvements in discrimination sensitivity ($d'$) without inducing conservative response bias. These findings indicate that defense against AI-enabled spear phishing depends not simply on accumulating indicators, but on synthesizing technical evidence with inferred psychological intent into explainable assessments, establishing adversarial mirroring and contextual AI synthesis as a scalable foundation for next-generation phishing defense.

## VIII. Acknowledgment

This work was supported by VirusTotal, Spur, Shodan, and GreyNoise through academic or research licenses that enabled the OSINT enrichment components of this work.

The work was also supported by the National Science Foundation under awards 2346726 and 2403360, and by the University of South Carolina Information Security Office (UISO), which provided access to its enterprise email environment for the collection of realistic phishing and legitimate email artifacts used in the evaluation.

## References

[1] A. Al-Subaiey, M. Al-Thani, N. A. Alam, K. F. Antora, A. Khandakar, and S. A. Uz Zaman, "Novel interpretable and robust web-based AI platform for phishing email detection," *Computers and Electrical Engineering*, vol. 120, 2024, Art. no. 109625.

[2] S. K. Birthriya, P. Ahlawat, and A. K. Jain, "Detection and prevention of spear phishing attacks: A comprehensive survey," *Computers & Security*, vol. 151, 2025, Art. no. 104317.

[3] H. Liang and Y. Xue, "Understanding security behaviors in personal computer usage: A threat avoidance perspective," *J. Assoc. Inf. Syst.*, vol. 11, no. 7, pp. 394–413, 2010.

[4] N. Kaloudi and J. Li, "The AI-based cyber threat landscape," *ACM Comput. Surveys*, vol. 53, no. 1, pp. 1–34, 2020.

[5] J. Hazell, "Spear phishing with large language models," unpublished, arXiv:2305.06972, 2023.

[6] F. Heiding, S. Lermen, A. Kao, B. Schneier, and A. Vishwanath, "Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects," unpublished, arXiv:2412.00586, 2024.

[7] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A comprehensive survey for intelligent spam email detection," *IEEE Access*, vol. 7, pp. 168261–168295, 2019.

[8] E. Blancaflor, L. F. Deldacan, S. Hunat, B. M. Rivera, and E. K. Liberato, "AI-driven phishing detection: Combating cyber threats through homoglyph recognition and user awareness," in *Proc. 6th World Symp. Softw. Eng. (WSSE)*, 2024, pp. 226–231.

[9] C. I. Canfield, B. Fischhoff, and A. Davis, "Quantifying phishing susceptibility for detection and behavior decisions," *Human Factors*, vol. 58, no. 8, pp. 1158–1172, 2016.

[10] F. Heiding, B. Schneier, A. Vishwanath, J. Bernstein, and P. S. Park, "Devising and detecting phishing emails using large language models," *IEEE Access*, vol. 12, pp. 42131–42146, 2024.

[11] A. R. Emanuela, B. A. Cristina, and S. Luminiţa, "AI and prompt engineering: The new weapons of social engineering attacks," in *Proc. 16th Int. Conf. Electron., Comput. Artif. Intell. (ECAI)*, 2024.

[12] S.-Y. Huang, Y.-W. Huang, and C.-H. Mao, "A multi-channel cybersecurity news and threat intelligent engine—SecBuzzer," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2019, pp. 691–695.

[13] K. Althobaiti, A. D. Jenkins, and K. Vaniea, "A case study of phishing incident response in an educational organization," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, 2021, Art. no. 379.

[14] N. C. Benda, L. L. Novak, C. Reale, and J. S. Ancker, "Trust in AI: Why we should be designing for appropriate reliance," *J. Amer. Med. Inform. Assoc.*, vol. 29, no. 1, pp. 207–212, 2021.