Adaptive AI Sentinels Against Phishing Attacks: Democratizing Cybersecurity Through Interactive Learning

Rishabh Pagaria, Jason Xiong, Ruihong Huang and Shreyas Kumar
Department of Computer Science & Engineering, Texas A&M University, USA

rishabh.pagaria@tamu.edu
jxx1008@email.tamu.edu
huangrh@cse.tamu.edu
shreyas.kumar@tamu.edu

Abstract: Phishing attacks have become more convincing as generative AI enables attackers to create polished, context-aware emails that closely resemble legitimate communication. These messages often evade traditional filters that rely on surface features and leave users without a clear understanding of why a message may be harmful. This work introduces an adaptive phishing-detection system that uses natural language processing to model semantic, linguistic, and stylistic signals and produce a risk score indicating how phish-like or benign an email appears. A complementary large language model layer then performs contextual and intent-based reasoning to interpret the deeper meaning of the message and detect subtle social engineering cues. The system incorporates adversarial and prompt-safety checks to strengthen reliability against AI-generated threats and through a web app, it delivers short micro-lessons for each detection, helping users understand the psychological tactics involved and learn to recognize them in future messages. This research contributes to both cybersecurity and NLP by showing how semantic scoring and LLM-based reasoning can be operationalized together to counter AI-enabled social engineering while remaining interpretable for non-expert users. By combining accurate detection with continuous user education, the proposed solution strengthens trust, awareness, and long-term resilience, offering a scalable defense mechanism for modern phishing attacks.

Key words: Social engineering, Generative AI, Large language models, Natural language processing, Cybersecurity

1. Introduction

Phishing remains one of the most damaging and persistent cyber threats, enabling credential theft, financial fraud, unauthorized access, and large-scale breaches across healthcare, education, finance, and government sectors (Salloum et al., 2022; Verizon, 2025). Recent threat intelligence reports show that phishing activity is not stabilizing but accelerating. IBM (2025) notes an 84 percent weekly rise in infostealer distribution through phishing emails, and the IC3 (2023) reports nearly three billion dollars in losses from business email compromise in a single year. These trends show that phishing is a fast-evolving and high-impact threat that affects both individuals and organizations.

This landscape has shifted further with the growth of generative Artificial Intelligence (AI). Attackers now use AI writing tools, chatbots, and website generators to produce messages that are coherent, polished, and tailored to different audiences. Unit42 (2025) finds that a significant portion of AI-assisted phishing leverages automated website generation and writing assistants, while Kumar et al. (2023) report a dramatic rise in AI-generated phishing emails after the release of modern language models. These attacks eliminate earlier signals such as spelling mistakes or awkward phrasing, making them more convincing and harder to detect. They also scale across languages and demographics, expanding attacker reach (Yu et al., 2023; Kumar et al., 2023).

The challenge, however, is not purely technical. Phishing succeeds because it exploits psychological triggers such as authority, trust, fear, and urgency (Hong, 2012). Even with advanced filtering, users often fall for messages that appear contextually relevant or personally meaningful. Major platforms such as Google and Microsoft block large volumes of malicious mail, but these systems rarely explain why a message is harmful or how the manipulation works (Tang and Mahmoud, 2021). Without this insight, users cannot build the judgment needed to resist new or highly personalized attacks. Research further shows that traditional awareness training has limited long-term impact in real inbox environments (Hillman et al., 2023).

To address these challenges, this work proposes an adaptive AI sentinel framework built around a dual Natural Language Processing (NLP) and Large Language Model (LLM) reasoning pipeline and an explainable output layer. The NLP component models semantic, linguistic, and stylistic signals to estimate how phish-like or benign a message appears, while the LLM component performs deeper contextual and intent-based reasoning to identify subtle social engineering strategies such as persuasion, impersonation, urgency framing, or emotional manipulation. The system is designed to produce structured, interpretable JavaScript Object Notation (JSON) based outputs that summarize deceptive cues, confidence levels, contextual indicators, and concise user tips.

These explanations form the foundation of the framework's educational design. A web app consumes this JSON output and converts each detection into a short micro-lesson, helping users understand why the email is risky and how similar manipulative tactics may appear in future messages. This transforms phishing detection from a passive alert into an active learning moment.

By combining semantic scoring, contextual reasoning, interpretable outputs, and real-time micro lessons, the proposed sentinel supports both accurate detection and continuous user empowerment. This approach strengthens transparency, improves user trust, and builds long-term resilience against rapidly evolving phishing and social engineering attacks.

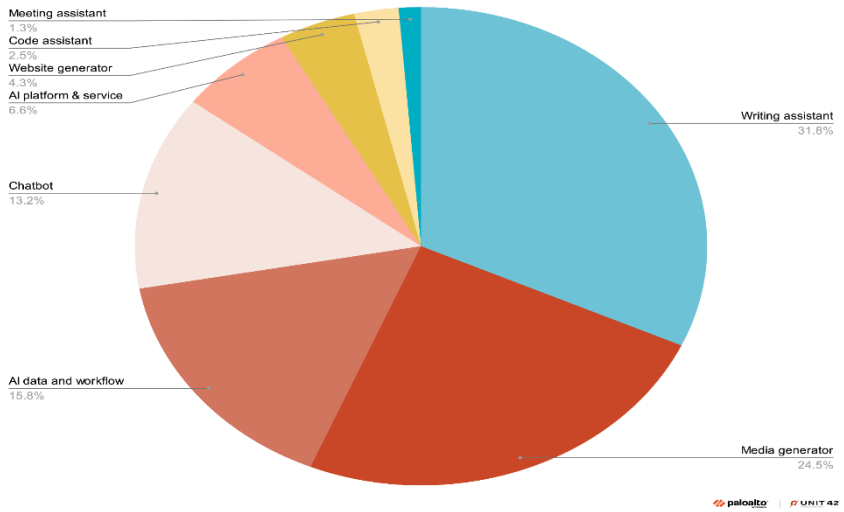

Figure 1: Distribution of categories of AI services misused for phishing attacks. Source: (Unit42, 2025)

2. Background

Phishing has long been a major security challenge, with early defenses relying on simple rules and heuristic checks. These approaches were fast and easy to deploy but were easily bypassed when attackers modified wording or created more convincing replicas. Machine Learning (ML) methods such as decision trees, logistic regression, and Support Vector Machines (SVMs) improved detection by learning richer linguistic and structural patterns (Basnet et al., 2012; Zareapoor & Seeja, 2015), yet still depended on manually crafted features that struggled to generalize across evolving writing styles, emerging social-engineering tricks, and multilingual content (Salloum et al., 2022; Almomani et al., 2013).

To address the limitations of handcrafted features, Natural Language Processing (NLP) approaches incorporated intent analysis, contextual cues, and semantic patterns. PhishNet-NLP demonstrated that phishing emails exhibit "action-eliciting" behavior across headers, links, and body text (Verma et al., 2012). While these methods captured deeper cues such as manipulative tone, early NLP systems used n-grams and bag-of-words representations that remained brittle under paraphrasing and subtle linguistic edits (Chawla & Chouhan, 2014). Dimensionality-reduction techniques like PCA and LSA improved stability but still fell short of modeling deeper semantic structure (Zareapoor & Seeja, 2015).

Transformer-based models such as BERT and RoBERTa advanced phishing detection by providing strong semantic and contextual understanding (Devlin et al., 2019; Liu et al., 2019), and demonstrated cross-lingual robustness aligned with broader NLP adoption (Salihovic et al., 2019; Wolf et al., 2020). However, these models remain vulnerable to adversarial perturbations, where small edits like word substitution or paraphrasing significantly reduce accuracy (Ebrahimi et al., 2018). Their internal reasoning is also difficult to interpret, limiting transparency for end users.

Modern phishing has further escalated with the rise of generative AI. Attackers now use AI writing tools and automated website generators to craft fluent, personalized, and multilingual phishing messages that evade older filters (Unit42, 2025; Kumar et al., 2023; Yu et al., 2024). At the same time, Large Language Models (LLMs) can themselves be manipulated through prompt injection, data poisoning, and adversarial perturbations, raising

concerns about safe deployment in security systems (Rossi et al., 2024; Koley et al., 2024). Together, these developments underscore the need for defenses that unify semantic detection, contextual reasoning, adversarial robustness, and user education. Our proposed dual NLP–LLM framework addresses these gaps by identifying phishing tactics, generating structured JSON explanations, and delivering micro-lessons that support long-term user resilience beyond enterprise environments.

3. Related Work

Research on phishing detection has progressed through several stages, yet significant gaps remain between robustness, interpretability, and user empowerment. Early rule-based and machine learning approaches flagged suspicious keywords, misspellings, or blacklisted URLs and achieved detection accuracies of around 60–70%, but were easily bypassed by minor wording changes (Basnet et al., 2012). Classical machine learning models such as SVMs and decision trees improved performance to roughly 75–80% on benchmark datasets (Fette et al., 2007), though they required frequent retraining as attackers shifted writing styles and tactics.

Transformer-based NLP models such as BERT, RoBERTa, and XLM-R introduced stronger semantic and contextual understanding, reaching accuracies above 88–93% on several phishing and email classification tasks (Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020). However, these models remain fragile under adversarial text perturbations: small edits such as synonym substitutions or paraphrasing can sharply degrade detection accuracy (Ebrahimi et al., 2018). In parallel, Large Language Models offer zero-shot and few-shot adaptability and can reason about manipulative tactics such as urgency framing or impersonation (Kumar et al., 2023; Jabir et al., 2025), but their susceptibility to prompt injection and data poisoning limits their reliability in security-sensitive deployments (Rossi et al., 2024).

Complementary research in explainable AI introduced tools like SHAP and LIME to improve transparency by highlighting influential features (Ribeiro et al., 2016; Lim et al., 2025). While effective for analysts, these explanations are rarely accessible to non-technical users. Cybersecurity education tools—such as phishing simulations and awareness training offer short-term improvements but lack integration with real-time defenses, resulting in limited lasting impact (Hillman et al., 2023).

Industry evidence underscores the urgency of addressing these limitations. Verizon (2025) attributes phishing to 36% of global breaches, IBM (2025) estimates $2.9 billion in annual losses from business email compromise, and Proofpoint (2024) reports that more than 80% of organizations experienced at least one successful phishing attack in the last year.

Overall, the literature reveals clear trade-offs across existing approaches: rule-based and ML systems offer limited robustness and no explanation, transformer-based NLP improves accuracy but remains adversarially fragile, LLM-based detection provides reasoning but introduces safety risks, and XAI methods improve transparency but remain analyst-focused rather than user-facing. Cybersecurity education tools enhance awareness but lack real-time integration. These gaps motivate a framework that jointly integrates NLP and LLM detection with continuous micro-lesson–based user education, transforming each decision point into a moment of resilience-building and aligning technical robustness with human-centered security.

4. Strategy

This research is anchored in such a way that to understand the threat landscape we analyzed a few of the case studies first and then built out layered system architecture driven by real-world phishing scenarios.

4.1 Case Studies

To contextualize the threat landscape, we reviewed a few recent phishing incidents to come up with a few recurring patterns. Phishing continues to drive serious operational and regulatory consequences across sectors. In healthcare, PIH Health reached a $600,000 HIPAA settlement after a phishing attack exposed electronic protected health information, with investigators citing inadequate safeguards and insufficient employee awareness as root contributors (HIPAA Training, 2025). Similar risks were seen in the UC San Diego Health breach, where credential phishing gave attackers access to patient data stored in compromised email accounts, prompting legal action, federal notifications, and follow-up incidents in later years (Businesswire, 2021; ClassAction, 2021; HIPAA Journal, 2024). These cases highlight how a single deceptive email can escalate into financial penalties, loss of trust, and long-term compliance obligations.

At the same time, attacker capabilities have expanded through generative AI. Unit42 (2025) reports that nearly 40% of AI-assisted phishing now uses automated website generators, while chatbot-driven writing tools produce

coherent, multilingual content that removes traditional warning signs. Kumar et al. (2023) identified a more than 1,200% rise in AI-generated phishing emails, and Yu et al. (2024) showed that LLM-produced messages are significantly harder to distinguish from legitimate communication. Together, these developments demonstrate the need for next-generation defenses that combine advanced detection with educational feedback, helping users recognize manipulative strategies even as phishing techniques rapidly evolve.

4.2 System Architecture

The proposed framework follows a layered architecture composed of three major components. Database Layer, Model Layer, and Application Layer incorporating supporting modules for pre-processing, adversarial defense, and explainability. This modular design ensures scalability, resilience, and adaptability to evolving phishing threats while keeping user education at its core.

4.2.1 Database layer

The Database Layer is responsible for ingesting, storing, and retrieving data in a secure and flexible manner. It integrates multiple sources, including publicly available repositories such as PhishTank (2023) and the Enron-Spam corpus, industry threat intelligence feeds, and synthetic phishing samples generated using large language models in controlled environments. Data is stored in a secure, encrypted Google Cloud Platform (GCP) cloud repository that can be centralized for research environments or deployed locally on-device for privacy-sensitive domains such as healthcare or government agencies. Each email record is supplemented with metadata, including semantic coherence scores, entity recognition tags, and sentiment polarity annotations, which serve both training and interpretability purposes. Privacy is a central design principle for this given layer and personally identifiable information (PII) is minimized or removed through anonymization pipelines to ensure compliance with frameworks such as General Data Protection Regulation (GDPR) and HIPAA (NIST, 2024).

4.2.2 Model layer

The Model Layer functions as a dual-pipeline system integrating transformer-based NLP models with LLM-driven reasoning. Incoming emails undergo preprocessing that removes HTML artifacts, normalizes text, tokenizes content, and extracts linguistic cues such as discourse markers, sentiment shifts, and entity anomalies. To improve robustness, the system applies adversarial augmentation by generating perturbed variants through synonym substitution and paraphrasing, helping the models withstand common evasion strategies (Ebrahimi et al., 2018; Koley et al., 2024). The NLP pipeline fine-tunes transformer classifiers such as BERT, RoBERTa, and XLM-R to detect semantic incoherence and pragmatic inconsistencies with high precision (Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020). In parallel, the LLM pipeline uses generative models such as Gemini or Claude (Jabir et al., 2025; Kumar et al., 2023) to infer higher-level manipulative tactics including urgency framing, impersonation of authority, and contextual mismatch, even in zero-shot settings.

Outputs from both pipelines converge in an aggregation and defense layer. When the two models agree, the system issues a high-confidence classification; when they diverge, the framework flags uncertainty and applies conservative blocking or escalation. This layer also incorporates protections against prompt injection (Rossi et al., 2024), adversarial text perturbations (Ebrahimi et al., 2018), and data poisoning (Fendley et al., 2025). All final decisions are converted into a structured JSON explanation object containing fields such as risk_score, linguistic_features, social_engineering_signals, model_confidence, and plain_language_reasoning. This unified JSON output serves as the interpretability backbone for both user-facing micro-lessons and analyst-level review.

4.2.3 Application layer

The Application Layer translates these model decisions into actionable insights for end users and analysts. Each classification generates a structured JSON explanation object, which the companion web interface and Gmail side-panel plugin directly consume. The plugin parses fields such as risk_score, detected_strategy, semantic_anomalies, and LLM_reasoning, transforming them into concise micro-lessons such as “This message attempts to create urgency by threatening service suspension.” Unlike conventional filters, the system provides not only a verdict but a short, context-aware explanation that reinforces user learning.

For organizational deployments, an analyst dashboard displays these same JSON outputs at scale, aggregating confidence levels, uncertainty flags, semantic cues, and cross-model disagreements. This supports policy refinement, targeted training interventions, and rapid threat triage. A built-in feedback mechanism allows users to confirm or contest detections, and each response is anonymized and written back into the database as a labeled JSON interaction record. This continuous loop ensures that the system not only protects users but also adapts based on real-world behavior, sustaining both its educational mission and long-term robustness.

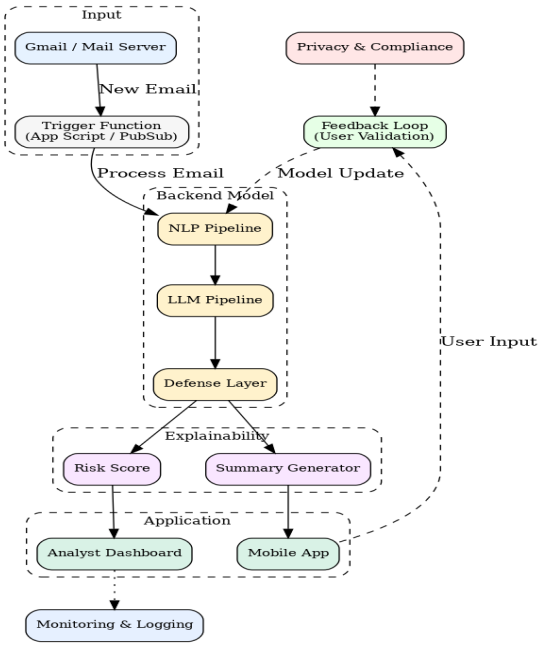

Figure 2: System Architecture

5. Methodology

The methodology was designed to balance technical reproducibility with human-centered evaluation, in alignment with ICAIR's emphasis on trustworthy and resilient AI systems. It unfolds across four stages: data collection, system development, evaluation, and continuous learning.

5.1 Data Collection

To ground the framework in real-world data, the design leverages multiple sources. Public repositories such as PhishTank (2023) and the Enron-Spam corpus provide well-labeled phishing and benign samples suitable for supervised learning. To reflect the evolution of attacker tactics, additional datasets are synthetically generated in controlled environments using large language models (LLMs) such as (Generative Pre-trained Transformer) GPT, LLaMA, etc. These synthetic samples incorporate adversarially crafted variants, including paraphrasing, synonym substitution, and disguised links, ensuring resilience against common evasion strategies (Ebrahimi et al., 2018; Koley et al., 2024). Metadata such as semantic coherence, sentiment polarity, and named entity annotations are added to support both classification and interpretability. Although the present study emphasizes English-language data, the framework is extensible to multilingual environments such as Spanish and Mandarin, which addresses cross-lingual phishing risks highlighted in prior research (Sahingoz et al., 2019).

5.2 System Development

The system uses a dual-pipeline analytical design that combines discriminative NLP classifiers with generative LLM reasoning. Incoming emails are normalized, stripped of HTML artifacts, tokenized, and enriched with linguistic features such as discourse markers, sentiment shifts, and entity anomalies, while adversarial augmentation prepares the model to withstand common evasion strategies (Ebrahimi et al., 2018). The NLP stream fine-tunes transformer models such as BERT, RoBERTa, and XLM-R to detect semantic inconsistencies and structural anomalies (Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020). In parallel, the LLM stream uses generative models such as GPT and LLaMA to reason about higher-level manipulative cues including urgency framing, impersonation, and contextual mismatches (Kumar et al., 2023; Jabir et al., 2025). Both pipelines feed into an aggregation and defense layer that reconciles outputs, assigns confidence levels, and enforces safeguards against prompt injection, adversarial perturbations, and data poisoning (Rossi et al., 2024; Fendley et al., 2025). To ensure transparency and accessibility, the system produces JSON-based outputs that combine SHAP or LIME attributions with LLM-generated natural-language explanations (Ribeiro et al., 2016; Lundberg & Lee, 2017).

5.3 Testing

Evaluation of the system is structured along two complementary dimensions. On the technical side, the framework is benchmarked against multiple baselines, including keyword-based filters, support vector machine classifiers, transformer-only detection, and LLM-only detection. Metrics such as accuracy, precision, recall, F1-score, and Receiver Operating Characteristic - Area Under Curve (ROC-AUC) provide a comprehensive assessment of performance. Preliminary results indicate that the hybrid dual-pipeline approach outperforms both individual pipelines, particularly under adversarially perturbed phishing scenarios.

Beyond technical evaluation, human-centered validation is also planned. A small-scale user study will examine the educational impact of the system by inviting participants to interact with flagged phishing emails and evaluate the clarity and usefulness of the generated explanations. This step ensures that the framework contributes not only to accurate detection but also to the long-term development of phishing awareness. Cross-lingual robustness will also be evaluated using synthetically generated phishing samples in languages such as Spanish and Mandarin to assess adaptability to diverse communication environments.

5.4 Continuous Learning

The final stage of the methodology embeds a continuous learning mechanism into the system. A feedback loop allows users to confirm or contest detections through the companion application. Each response is anonymized, filtered for quality, and reintegrated into the training dataset, ensuring incremental improvement over time. To support large-scale organizational deployment, the architecture is designed with privacy-preserving mechanisms such as federated learning, allowing multiple entities to contribute to collective model improvement without exposing sensitive raw data. This approach ensures compliance with standards such as GDPR and HIPAA (NIST, 2024) while reinforcing the system's adaptability against emerging zero-day phishing attacks.

6. Discussion

The proposed framework demonstrates how technical innovation in phishing detection can be aligned with principles of trustworthy and resilient AI. By merging advanced natural language processing with user-centered education, the system contributes not only to improved detection performance but also to broader organizational resilience and AI literacy. This section reflects on the benefits, challenges, and limitations of the approach, situating it within ongoing debates on governance and responsible deployment.

6.1 Benefits

Earlier phishing detection methods such as rule-based filters and classical machine learning improved accuracy but relied on engineered features that did not generalize well to evolving writing styles, multilingual attacks, or modern social-engineering strategies (Basnet et al., 2012; Zareapoor & Seeja, 2015). Transformer models improved semantic understanding and cross-domain performance, yet studies show they remain vulnerable to subtle adversarial text manipulations such as synonym substitutions and paraphrasing (Ebrahimi et al., 2018; Koley et al., 2024). At the same time, effective enterprise-grade filtering offered by Google and Microsoft remains largely inaccessible to individuals and SMEs who lack paid security tiers (Microsoft, 2024).

The proposed dual-pipeline architecture addresses these gaps by combining transformer-based NLP classification with LLM-driven contextual reasoning. This complementary design captures both fine-grained linguistic anomalies and higher-level manipulative strategies such as urgency framing, impersonation, and tone shifts (Devlin et al., 2019; Conneau et al., 2020; Kumar et al., 2023). Because each pipeline compensates for the limitations of the other, the system provides stronger resilience to paraphrasing attacks and adversarial perturbations than either technique alone, making it more suitable for detecting AI-generated phishing content.

A further benefit lies in the system's focus on explainability and user empowerment. Each detection is paired with a brief micro-lesson that explains the deceptive cues in clear language, helping users develop long-term recognition of psychological triggers such as emotional pressure and authority misuse (Gaspar et al., 2024; Hillman et al., 2023). Combined with built-in protections against prompt injection, data poisoning, and adversarial text manipulation (Rossi et al., 2024; Ebrahimi et al., 2018; Fendley et al., 2025), the framework aligns with emerging AI governance guidelines emphasizing transparency and human oversight. Together, these benefits show that the system strengthens detection accuracy while also democratizing cybersecurity for users beyond enterprise environments.

6.2 Challenges

Despite these benefits, several challenges remain. The sophistication of AI-generated phishing attacks evolves rapidly, and adversarial training cannot anticipate every possible manipulation strategy. Computational demands pose another constraint, as transformer-based and generative models are resource intensive, potentially limiting deployment in environments without dedicated infrastructure (Wolf et al., 2020). Additionally, sustaining user engagement without overwhelming individuals is complex. While educational explanations strengthen awareness, repeated alerts risk producing cognitive overload if not carefully designed (Hillman et al., 2023).

Ethical challenges also arise in balancing transparency with privacy. Explanations must remain clear without revealing sensitive user data, and bias within LLMs may produce uneven detection outcomes across cultural or linguistic groups. These risks highlight the importance of ongoing monitoring, fairness audits, and governance mechanisms.

6.3 Limitations

The present research has several limitations. First, the datasets used, such as PhishTank and Enron-Spam, are widely adopted but do not fully capture the diversity of enterprise phishing incidents (PhishTank, 2023). Future work will require collaboration with industry partners to integrate anonymized real-world email streams. Second, the current system focuses primarily on text-based phishing. Multimodal threats including image-based lures, voice phishing (vishing), and deepfake-enabled social engineering remain outside the present scope but represent critical vectors for future extension (Jabir et al., 2025). Third, conservative thresholds that reduce false negatives may increase false positives, creating a risk of alert fatigue. Prior work shows that repeated phishing-related warnings can overwhelm users and reduce engagement over time (Hillman et al., 2023).

6.4 Future Work

Future work will focus on developing the full phishing-detection framework using a combined NLP and LLM pipeline implemented with software engineering best practices, including modular design, reproducible model training, and structured evaluation. To strengthen data-privacy safeguards, the system will also be tested with compact Small Language Models such as Gemma or LLaMA, which offer local or limited-exposure processing while still providing contextual reasoning. Once the core system is implemented, Institutional Review Board (IRB) approval will be pursued to enable controlled user studies. These studies will evaluate how effectively the app based micro-lessons improve users' phishing recognition skills, decision confidence, and long-term awareness, helping validate the framework as both a security tool and an educational intervention.

7. Conclusion

Phishing attacks have evolved from simple deceptive messages into highly convincing, AI-generated campaigns that bypass many traditional defenses. Early heuristic and rule-based filters were easy to deploy but not adaptive, while classical machine learning improved detection yet remained limited by manually engineered features that failed to generalize to new writing styles and multilingual attacks. Even transformer-based NLP systems, though capable of deeper semantic analysis, remain vulnerable to adversarial text manipulation that subtly alters surface form without changing underlying intent (Ebrahimi et al., 2018; Koley et al., 2024). At the same time, enterprise-grade protections offered by major vendors remain accessible primarily to large organizations, leaving individuals and small businesses with fewer options (Microsoft, 2024).

To address these gaps, this work introduced an adaptive AI sentinel that combines transformer-based NLP classification with LLM-driven contextual reasoning. This dual-layer architecture is designed to detect semantic cues such as impersonation, urgency framing, tone manipulation, and cross-lingual inconsistencies, while integrated adversarial and prompt-safety checks reduce the risks associated with using LLMs in security workflows (Rossi et al., 2024). By merging linguistic scoring with deeper intent analysis, the system is positioned to recognize modern AI-generated phishing messages that mimic natural communication styles and target users with increasing precision (Kumar et al., 2023; Unit42, 2025).

A key contribution of this work is the emphasis on user education. Research shows that phishing continues to succeed because it exploits human psychology, and traditional one-time training has limited long-term impact (Hong, 2012; Hillman et al., 2023). The proposed framework addresses this gap through micro-lessons delivered via a companion web interface, transforming each detection into a short, accessible explanation of the manipulative tactics used in the message. This approach supports not only technical detection but also continuous awareness-building, making the defense more scalable and empowering for everyday users.

Ethics Declaration: This study did not involve human participants, animals, or sensitive data requiring ethical clearance. Hence, no ethical approval was required.

AI Declaration: AI-assisted tools i.e. ChatGPT was used in the preparation of this paper for language editing, restructuring, and improving readability. All conceptual ideas, analysis, and interpretations were developed by the authors.

References

Almomani, A., Gupta, B.B., Atawneh, S., Meulenberg, A. and Almomani, E. (2013) 'A Survey of Phishing Email Filtering Techniques', in IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2070-2090, Fourth Quarter 2013, doi: 10.1109/SURV.2013.030713.00020

Basnet, R., Sung, A. and Liu, Q. (2012) 'Rule-based phishing attack detection', International Journal of Information Security and Privacy, 6(3), pp.24–42.

Businesswire (2021) 'University of San Diego Health substitute notice of data breach', Businesswire, 27 July.

Chawla, M. and Chouhan, S.S. (2014) 'A Survey of Phishing Attack Techniques', International Journal of Computer Applications, vol. 93 - no. 3

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N. and Najada, H. A. (2014) 'Survey of review spam detection using machine learning technique', Journal of Big Data, doi:10.1186/s40537-015-0029-9

ClassAction (2021) 'Data breach: UC San Diego Health hit with class action over alleged four-month phishing attack', ClassAction.org, 19 August.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2020) 'Unsupervised cross-lingual representation learning at scale', Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.8440–8451. doi:10.48550/arXiv.1911.02116.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', Proceedings of NAACL-HLT 2019, pp.4171–4186. doi:10.18653/v1/N19-1423.

Ebrahimi, J., Rao, A., Lowd, D. and Dou, D. (2018) 'HotFlip: White-box adversarial examples for text classification', Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp.31–36. doi:10.48550/arXiv.1712.06751.

Fendley, N., Staley, E., Carney, J., Redman, W., Chau, M. and Drenkow, N. (2025) 'A systematic review of poisoning attacks against large language models', arXiv preprint. doi:10.48550/arXiv.2506.06518.

Fette, I., Sadeh, N. and Tomasic, A. (2007) 'Learning to detect phishing emails', Proceedings of the 16th International Conference on World Wide Web, pp.649–656. doi:10.1145/1242572.1242660.

Gaspar, D., Rezaei, M. and Zhao, Z. (2024) 'Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron', IEEE Access, 12, pp.3368377–3368389. doi:10.1109/ACCESS.2024.3368377.

Hillman, D., Barlow, J. and Evans, R. (2023) 'Evaluating organizational phishing awareness training on an enterprise scale', Computers & Security, 129, p.103364. doi:10.1016/j.cose.2023.103364.

HIPAA Journal (2024) 'Patient data exposed in phishing attack on UC San Diego Health', HIPAA Journal, 13 March. Available at: https://www.hipaajournal.com/march-13-2023-healthcare-data-breaches/

Hong, J. (2012) 'The state of phishing attacks', Communications of the ACM, 55(1), pp.74–81. doi:10.1145/2063176.2063197.

IBM (2025) 'IBM X-Force 2025 threat intelligence index', IBM Institute for Business Value. Available at: https://www.ibm.com/thought-leadership/institute-business-value/report/2025-threat-intelligence-index

IC3 (2023) 'Federal Bureau of Investigation Internet Crime Report 2023', Internet Crime Complaint Center. Available at: https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf

Jabir, R., Le, J. and Nguyen, C. (2025) 'Phishing attacks in the age of generative artificial intelligence: A systematic review of human factors', AI, 6(8), p.174. doi:10.3390/ai6080174.

Koley, A., Satpati, P., Choudhary, I. and Sen, J. (2024) 'An Investigation on the Efficiency of Some Text Attack Algorithms', 2024 IEEE North Karnataka Subsection Flagship International Conference (NKCon). doi:10.1109/NKCon62728.2024.10774713.

Kumar, S., Menezes, A., Giri, S. and Kotikela, S. (2023) 'What the phish: Effects of AI on phishing attacks and defense'

Lim, B., Huerta, R., Sotelo, A., Quintela, A. and Kumar, P. (2025) 'EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability', arXiv preprint. doi:10.48550/arXiv.2503.20796.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) 'RoBERTa: A robustly optimized BERT pretraining approach', arXiv preprint. doi:10.48550/arXiv.1907.11692.

Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', Advances in Neural Information Processing Systems (NeurIPS), 30, pp.4765–4774.

Mathews, L. (2017) 'Phishing Scams Cost American Businesses Half A Billion Dollars A Year'

Microsoft (2024) 'Anti-phishing protection in cloud organizations'

NIST (2023) 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)'. National Institute of Standards and Technology. Available at: https://www.nist.gov/itl/ai-risk-management-framework

NIST (2024) 'Building a Cybersecurity and Privacy Learning Program: NIST SP 800-50 Rev.1'. National Institute of Standards and Technology.

PhishTank (2025) 'Phishing dataset repository'. Available at: https://phishtank.org/index.php

Proofpoint (2024) '2024 state of the phish – today's cyber threats and phishing protection', Proofpoint Inc. Available at: https://www.proofpoint.com/us/resources/threat-reports/state-of-phish

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) '"Why should I trust you?" Explaining the predictions of any classifier', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1135–1144. doi:10.1145/2939672.2939778

Rossi, S., Michel, A., Mukkamala, R. and Thatcher, J.B. (2024) 'An early categorization of prompt injection attacks on large language models', arXiv preprint. doi:10.48550/arXiv.2402.00898

Sahingoz, O.K., Buber, E., Demir, O. and Diri, B. (2019) 'Machine learning-based phishing detection from URLs', Expert Systems with Applications, 117, pp.345–357. doi:10.1016/j.eswa.2018.09.029.

Salihovic, I., Serdarevic, H. and Kevric, J. (2019) 'The Role of Feature Selection in Machine Learning for Detection of Spam and Phishing Attacks', Intelligent Technologies and Robotics 2018, vol 60. Springer, doi:10.1007/978-3-030-02577-9_47

Salloum, S., Gaber, T., Vadera, T. and Shaalan, K. (2022) 'A systematic literature review on phishing email detection using natural language processing techniques', IEEE Access, 10, pp.65703–65727. doi:10.1109/ACCESS.2022.3183083.

Tang, L. and Mahmoud, Q.H. (2021) 'A Survey of Machine Learning-Based Solutions for Phishing Website Detection', Machine Learning and Knowledge Extraction, 3(3), pp.672–694. doi:10.3390/make3030034.

Unit42 (2025) 'Fashionable phishing bait: GenAI on the hook', Palo Alto Networks. Available at: https://unit42.paloaltonetworks.com/genai-phishing-bait

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) 'Attention is all you need', Advances in Neural Information Processing Systems (NeurIPS), 30, pp.5998–6008. doi:10.48550/arXiv.1706.03762.

Verizon (2025) '2025 data breach investigations report', Verizon Business. Available at: https://www.verizon.com/business/resources/reports/dbir

Verma, R., Shashidhar, N., Hossain, N. (2012) 'Detecting Phishing Emails the Natural Language Way', ESORICS 2012, pp 824–841, doi:10.1007/978-3-642-33167-1_47

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Brew, J. (2020) 'Transformers: State-of-the-art natural language processing', Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.38–45. doi:10.18653/v1/2020.emnlp-demos.6.

Yu, J., Yu, Y., Wang, X., Lin, Y., Yang, M., Qiao, Y. and Wang, F.-Y. (2023) 'The shadow of fraud: The emerging danger of AI-powered social engineering and its possible cure', arXiv preprint. doi:10.48550/arXiv.2407.15912.

Zareapoor, M. and Seeja, K. R. (2015) 'Feature Extraction or Feature Selection for TextClassification: A Case Study on Phishing Email Detection', IJIEEB, vol.7, no.2, pp.60-65, doi:10.5815/ijieeb.2015.02.08