# Human Perception and Detection of AI-Generated Phishing Emails: A Red-Teaming and Multi-Layered Detection Approach

Rohan Mehra*
rohanvmehra@outlook.com
rmehra26@riverdale.edu
Riverdale Country School
Bronx, New York, USA

## Abstract

Large Language Models (LLMs) now generate phishing emails indistinguishable from legitimate correspondence, challenging traditional detection systems. While baseline classifiers achieve 97–99% accuracy on established corpora, they remain untested against adversarially-generated AI variants. We introduce a detectability gap metric ($\Delta$) quantifying confidence differences between AI and human-authored phishing, enabling continuous monitoring of generative model evolution. Evaluating 155,659 emails (72,941 legitimate; 81,519 human-phishing; 1,199 AI-phishing) from three primary sources plus eleven datasets from the Champa et al. curated collection, our Random Forest classifier achieved 98.13% test accuracy with $\Delta = -0.016 \pm 0.003$ ($p < 0.01$, Cohen's $d = 0.18$). A GAN→GPT adversarial pipeline generating 280 variants revealed 84.3% detection rate, with misclassifications concentrated in formal, non-urgent examples. Feature analysis showed reliance on traditional phishing markers rather than AI-specific patterns, suggesting vulnerability as generative attacks evolve. This framework provides early-warning capability for tracking LLM advancement while maintaining computational efficiency (47MB model, 0.8ms inference) suitable for resource-constrained organizations.

## Keywords

Phishing Detection, Large Language Models, Machine Learning Security, Adversarial Machine Learning, Email Security, Generative AI, Random Forest, Cybersecurity

**ACM Reference Format:**
Rohan Mehra. 2026. Human Perception and Detection of AI-Generated Phishing Emails: A Red-Teaming and Multi-Layered Detection Approach. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR Vol. 3, Issue 1).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.5281/zenodo.18404590

*Mentor: Saranya Vijayakumar, Carnegie Mellon University

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
*IJSCAR Vol. 3, Issue 1, February 2026,*
© 2026 The Author(s). Published by IJSCAR.
https://doi.org/10.5281/zenodo.18404590

## 1 Introduction

### 1.1 The Escalating Threat of AI-Generated Phishing

Large Language Models (LLMs) such as GPT-4 and Claude 3.5 have fundamentally altered the phishing threat landscape. These models enable attackers to automate production of highly convincing, contextually relevant phishing emails that eliminate traditional detection cues—grammatical errors, awkward phrasing, and generic messaging [10, 11, 29]. The scope of this threat is substantial: the FBI's 2023 Internet Crime Report documented 298,878 phishing complaints in the United States, with approximately 91% of phishing attacks originating via email [8, 13]. Business email compromise (BEC) alone generated $2.9 billion in losses in 2023 [8], with three-year (2022–2024) total BEC losses approaching $8.5 billion [18].

Vulnerable populations face disproportionate risks. Small businesses experience 350% more phishing attacks than enterprises, with organizations of 1 to 250 employees receiving malicious email at rates as high as 1 in 323 messages [25]. Educational institutions report 60% exposure to cloud-targeting phishing attacks [20], and U.S. school districts sustained 855 cyberattacks between 2016–2019, many phishing-initiated [6]. These organizations typically operate with limited security resources while attackers leverage LLMs to generate high-quality phishing at minimal cost [10].

### 1.2 The Detection Gap

Traditional machine learning approaches achieve 95–99% accuracy on benchmarks like SpamAssassin and PhishTank [15, 17], but were designed for human-authored phishing containing linguistic anomalies and structural inconsistencies. These assumptions fail when LLMs eliminate errors and adapt seamlessly to context. Recent work addresses AI-generated phishing through stylometry and semantic embeddings [7, 19], achieving F1≈0.98 (T5-based) and 96% accuracy (XGBoost on GPT-4o emails). However, critical gaps remain: static datasets cannot track evolving detection difficulty across LLM generations, adversarial robustness testing remains sparse, quantitative metrics for measuring relative AI versus human phishing detectability are lacking, and human perception remains unintegrated despite being the ultimate defense layer. The challenge now spans both human perception and computational detection, since LLM fluency erodes the surface cues users once depended on and forces a reassessment of how phishing should be evaluated.

### 1.3 Our Contribution

We introduce four primary contributions addressing these gaps. **First**, we define the detectability gap (Δ) quantifying classifier confidence differences between AI and human-authored phishing. Unlike binary metrics (accuracy, F1), Δ enables continuous monitoring of generative model evolution and early detection of emerging evasion capabilities. This framework extends naturally to human detection: $\Delta_{\text{human}}$ will quantify whether humans find AI-phishing harder to detect than traditional phishing, enabling measurement of the perception gap alongside the computational gap. **Second**, we demonstrate classical Random Forest maintains competitive performance (98.03–98.13% accuracy) on realistically imbalanced datasets while offering computational efficiency suitable for resource-constrained settings: 47MB storage, 0.8ms inference on standard CPU, macro-F1=0.99 with perfect AI-phishing precision. **Third**, we develop a GAN→GPT adversarial pipeline generating stylistically diverse phishing across seven conditions, revealing that detectors rely on traditional markers (urgency verbs, action requests) rather than AI-specific stylometry—achieving 84.3% detection but failing on formal, non-urgent examples. **Fourth**, we provide comprehensive, reproducible methodology with detailed hyperparameters, addressing transparency gaps in phishing detection literature.

### 1.4 Paper Organization

Section 2 reviews pre-LLM and post-LLM detection approaches while contextualizing our contributions. Section 3 details dataset construction, model training, and adversarial evaluation methodology. Section 4 presents performance analysis, vectorization comparisons, and adversarial robustness findings. Section 5 discusses temporal bias, sample size constraints, and detector mechanism vulnerabilities. Section 6 outlines longitudinal evaluation, human perception integration, and hybrid architectures. Section 7 synthesizes implications for adaptive defenses in the evolving arms race between synthetic deception and computational detection.

## 2 Literature Review

### 2.1 Problem Setup and Audience Impact

This research addresses a fundamental question: How reliably can automated systems and users distinguish well-written, natural-sounding AI-generated phishing attempts from legitimate messages? We hypothesize that distinguishing AI-generated phishing from legitimate emails will become increasingly difficult as LLMs advance, driven by three convergent factors: lowered attack costs enabling mass customization, enhanced linguistic quality eliminating traditional error-based detection cues, and scalable personalization capabilities exploiting recipient-specific context.

The stakeholders benefiting from improved phishing protection extend beyond enterprise security teams. Small and midsize businesses without dedicated IT security teams face increased risks of wire fraud, credential theft, and business email compromise. Educational institutions risk credential theft, data breaches, and service disruptions from ransomware. Local and family businesses face exposure to fake invoices and supplier fraud. Everyday users relying on familiar branding and professional tone face eroding intuitive defenses as AI convincingly reproduces these signals.

### 2.2 Selected Data Points

Drawing from recent threat intelligence and incident reports, the scale and impact of phishing threats is substantial and well-documented. The FBI Internet Crime Complaint Center (IC3) received 298,878 phishing reports in the United States in 2023, with approximately 91% of phishing attacks originating via email [13]. Educational institutions face particularly severe exposure, with 60% of educational organizations reporting phishing attacks targeting cloud data, and many remaining unaware of breaches for days [20]. Phishing frequency in the education sector exceeded a 40% cross-vertical average [20], and between 2016 and 2019, 855 recorded cyberattacks occurred on U.S. school districts, many of which involved phishing or ransomware as the initial attack vector [6].

Small businesses face disproportionate risks compared to larger organizations. Businesses with fewer than 100 employees suffer an estimated 350% more phishing and social-engineering attacks than larger enterprises, and for organizations with 1 to 250 employees, approximately 1 in 323 emails is malicious [25]. These statistics underscore the urgent need for accessible, resource-efficient detection systems that can protect vulnerable populations operating without enterprise-grade security infrastructure.

### 2.3 Pre-LLM Detection Approaches

Phishing detection in the pre-LLM era relied predominantly on classical machine learning and deep learning methods operating over text content, URL features, and email metadata. Comprehensive reviews spanning over 80 studies reported that CNN and LSTM models trained on Word2Vec or TF-IDF representations achieved accuracy ranging from 95.0% to 99.8% with F1 scores up to 0.998 on established benchmark datasets including SpamAssassin, PhishTank, and UCI [17, 26]. Lightweight 1D-CNN architectures combined with recurrent layers (LSTM, Bi-GRU) achieved comparable performance, with some studies reporting accuracy or F1 scores up to 99.68%.

Ensemble approaches sought to combine multiple feature types and detection strategies. The DARTH framework integrated NLP clustering with multi-feature analysis, reporting F1 scores near 99.98% on UCI and PhishTank datasets [17]. These ensemble methods demonstrated that combining lexical features (word frequencies, character n-grams), structural features (email headers, MIME structure), HTML-based features (embedded scripts, form elements, hidden content), and URL-based features (domain reputation, URL length, suspicious TLDs, subdomain depth) could yield incremental performance improvements over single-feature approaches.

Despite strong benchmark performance, pre-LLM detection research exhibited several persistent gaps. Dataset diversity remained limited, with most studies evaluating on a small set of canonical benchmarks that may not capture the full heterogeneity of real-world phishing. Real-time deployment considerations received insufficient attention, with few studies reporting inference latency, throughput constraints, or computational requirements for production systems. Multilingual support was sparse, with the vast majority of research focusing exclusively on English-language emails. Most critically, these approaches assumed phishing emails would contain detectable linguistic or structural errors—an assumption that LLM-generated phishing fundamentally invalidates.

## 2.4 Post-LLM Detection Approaches

With LLMs enabling generation of fluent, grammatically correct, and contextually appropriate phishing emails, detection research has shifted toward stylometric analysis and semantic representations that can identify subtle patterns distinguishing AI-generated from human-authored text [28]. Recent work demonstrates that LLMs enable attackers to generate highly convincing, tailored phishing content at scale [10], challenging traditional detection approaches that relied on linguistic errors. End-to-end analyses of LLM-generated textual phishing have proposed semantic machine learning pipelines combining T5 embeddings with multilayer perceptron classifiers, achieving F1 scores approximately 0.98 on Nazario and Enron corpora [19]. Stylometric approaches using XGBoost classifiers have demonstrated 96.0% accuracy and AUC 0.99 when tested against GPT-4o-generated phishing emails [7].

Comparative studies across multiple architectures (SVM, Random Forest, CNN, BiLSTM) on AI-generated datasets report F1 scores as high as 100.0% in controlled settings, though such perfect performance likely reflects dataset-specific characteristics rather than generalizable robustness [12]. Evaluations comparing quantized LLMs with classical models show Bi-GRU architectures achieving 98.77% accuracy, while smaller LLM-based classifiers lag behind (approximately 81.0% accuracy) but demonstrate greater resilience to adversarial rephrasing attacks [12]. This suggests a potential trade-off between raw classification accuracy and robustness to input perturbations. GPT-based detection tools demonstrate the potential for LLMs to identify phishing emails and generate explanatory warnings for users [4], while watermarking techniques combined with machine learning show promise for detecting LLM-generated phishing emails [2], offering additional defensive layers against AI-crafted attacks.

Despite these advances, post-LLM detection research exhibits shortcomings that our work addresses. Corpora remain fragmented, with different studies using non-overlapping datasets that prevent direct performance comparisons. Multilingual handling remains inconsistent, with most studies focusing on English despite LLMs' multilingual generation capabilities. Integration of human perception remains limited, even though human detection capabilities represent a critical component of real-world defense. Most significantly, adversarial robustness analyses are sparse: few studies systematically evaluate detectors against adversarially optimized AI-generated phishing designed specifically to evade detection.

## 2.5 Human Detection of Phishing

Classical studies of human phishing detection reveal systematic reliance on heuristic cues rather than systematic verification. Users tend to evaluate emails based on familiar branding, professional tone, polite language, and visual layout consistency rather than carefully examining sender addresses, hovering over links, or verifying unusual requests through independent channels [1, 5, 14, 23, 27]. Research has identified individual differences in phishing vulnerability based on cognitive processing styles, with some users more prone to heuristic-based decision making that favors attackers [27]. Understanding the interaction among different human factors remains crucial for measuring phishing susceptibility [9].

These heuristics worked reasonably well against traditional phishing, which often contained spelling errors, awkward phrasing, and unprofessional formatting. However, LLM-generated emails can perfectly mimic all surface-level legitimacy cues. Recent research suggests these human vulnerabilities extend directly to AI-generated phishing. Studies comparing human click-through rates on expert-crafted human phishing versus LLM-generated phishing find statistically equivalent susceptibility [12, 24], indicating that current AI-generated phishing already matches human-expert quality in terms of deceiving end users. These findings motivated our development of computational detection methods as a necessary complement to user education, since relying solely on human vigilance may prove insufficient as generative capabilities advance.

## 2.6 Detectability Gap: Formal Definition

We introduce the detectability gap ($\Delta$) as a novel metric quantifying how differently a classifier responds to AI-generated versus human-developed phishing content. We define $\Delta$ as the difference in average classifier confidence between AI and human-crafted phishing:

$$\Delta = \mathbb{E}_{x \sim P_{\text{AI}}}[f(x)] - \mathbb{E}_{x \sim P_{\text{human}}}[f(x)], \tag{1}$$

where $f(x) \in [0, 1]$ represents the detector's phishing confidence score (higher values indicate greater confidence that the email is phishing), $x$ denotes a single email (subject and body concatenated), and $P_{\text{AI}}$, $P_{\text{human}}$ represent distributions over phishing emails authored by LLMs versus humans respectively. The expectation $\mathbb{E}_{x \sim P}[\cdot]$ denotes the average over emails sampled from distribution $P$.

The sign of $\Delta$ indicates detection difficulty: $\Delta < 0$ means AI-phishing receives lower detection confidence than human phishing (harder to detect), while $\Delta > 0$ indicates the opposite. Unlike standard binary classification metrics (accuracy, AUC, F1) that collapse performance into single aggregate scores, $\Delta$ provides continuous monitoring capability by tracking confidence distributions over time. As new LLM generations are released, measuring $\Delta$ for each generation enables quantitative tracking of whether detection difficulty increases, decreases, or remains stable.

We measured $\Delta = -0.016 \pm 0.003$ (95% CI) via bootstrap resampling with 10,000 iterations from per-class probability distributions. Statistical significance was assessed using the Mann-Whitney U test ($p < 0.01$) comparing distributions of $f(x)$ for AI-phishing versus human-phishing samples. Effect size quantified via Cohen's $d = 0.18$ indicates a small but measurable difference in classifier confidence. While the absolute magnitude is modest, this consistent detectability gap provides a quantifiable baseline for longitudinal tracking of LLM evolution and enables early detection of emerging evasion capabilities.

# 3 Methods

## 3.1 Dataset Construction

We assembled a dataset of 155,659 emails from three primary sources plus eleven constituent datasets from the Champa et al. curated collection, intentionally imbalanced to emulate real-world email ecosystems. The final corpus contains 72,941 legitimate emails,

81,519 human-authored phishing, and 1,199 AI-generated phishing (0.7% of total), mirroring operational environments where phishing typically represents fewer than 3% of messages.

Our three primary sources provide AI-generated and modern phishing examples. Nahmias et al.'s prompted contextual vectors for spear-phishing detection, produced by a proprietary ensemble of LLMs, contribute sophisticated AI-generated phishing with contextual targeting [19]. Eze & Shamir's AI-generated phishing email corpus (GPT-3.5/4) provides synthetically generated phishing with controlled linguistic patterns [7]. The University of Twente phishing emails dataset supplies additional labeled classification examples [16].

The Champa et al. curated collection [3, 22] aggregates eleven established benchmark datasets spanning 1995–2022, providing comprehensive coverage of phishing evolution. This curated collection includes: CEAS_08 (early corporate spam with 2008-era patterns), Enron (>500,000 authentic corporate emails from 1999–2002), Ling-Spam (academic spam patterns), two Nazario collections (spear-phishing with emotional appeals and narrative deception tactics), two Nigerian Fraud archives (advance-fee linguistic structures), SpamAssassin (canonical ham/spam with clean legitimate corpora), and three TREC spam track datasets (TREC_05, TREC_06, TREC_07 providing standardized NIST benchmarks). Note that original Nazario URLs (monkey.org/~jose/) are unstable; we accessed these via the Champa curated repository.

This mixture balances authenticity, diversity, and realism across emotional, formal, and transactional linguistic forms. The temporal mismatch (1999–2002 vs. 2024) represents a limitation but enables assessment of cross-era generalization. Legitimate and human-phishing subsets ground classification in established norms, while AI-phishing introduces novel generative patterns.

## 3.2 Text Preprocessing

We concatenated subject and body fields into unified text representations, then applied tokenization using NLTK's `word_tokenize` function. Lemmatization via `WordNetLemmatizer` reduced morphological variants to their base forms, improving feature consistency while preserving semantic content. We filtered English stopwords using NLTK's standard stopword list to reduce dimensionality and focus on content-bearing terms.

## 3.3 Vectorization Approaches

We compared three complementary text representation schemes. Bag-of-Words (BoW) captures raw token frequencies without weighting adjustments, providing a baseline that emphasizes repeated keywords common in phishing. Word-level TF-IDF reweights tokens by their discriminative power, down-weighting terms appearing frequently across all email types while emphasizing distinctive vocabulary. Character-level TF-IDF with 3–5 character $n$-grams captures sub-word patterns including punctuation, capitalization, and character sequences that may signal synthetic generation. All vectorization schemes used `max_features=5000` to balance representational richness with computational tractability.

## 3.4 Model Training

We employed a two-stage hyperparameter optimization process using stratified 60/20/20 (train/validation/test) splits with 5-fold cross-validation and out-of-bag (OOB) validation:

**Stage 1: Grid Search.** We explored $n_{\text{estimators}} \in \{100, 200\}$, max_depth $\in \{\text{None}, 10, 20\}$, and min_samples_split $\in \{2, 5\}$, yielding 12 candidate configurations ($2 \times 3 \times 2$) evaluated with 5-fold cross-validation on the training set (60 total fits). Grid search identified $n_{\text{estimators}} = 200$, max_depth = None, min_samples_split = 2 as optimal within the tested range.

**Stage 2: Extended Optimization.** We further tested $n_{\text{estimators}} = 300$ with the optimal hyperparameters from Stage 1 to assess potential performance gains from additional trees. OOB error analysis (Figure 1) showed continued improvement up to 300 trees, establishing this as the final configuration.

**Stage 3: Vectorization Comparison.** To ensure fair comparison across Bag-of-Words, word TF–IDF, and character TF–IDF, we standardized all models at $n_{\text{estimators}} = 200$ (optimal from initial grid search) during vectorizer evaluation, isolating the effect of text representation from ensemble size.

**Final Deployment Model.** The optimal configuration ($n_{\text{estimators}} = 300$, max_depth = None, min_samples_split = 2) was retrained on the combined train+validation sets using character TF–IDF representation, achieving 98.13% test accuracy with OOB error of 0.0197.

## 3.5 GAN→GPT Adversarial Pipeline

To probe robustness against strategic evasion, we developed a two-stage adversarial synthesis pipeline. The GAN generator comprises three fully connected layers (latent dimension $100 \to 256 \to 512 \to 5000$ TF-IDF features) with LayerNorm for training stability. The discriminator employs spectral normalization and minibatch discrimination to encourage sample diversity. We trained for 300 epochs using Adam optimization (learning rate $10^{-4}$ for generator, $4 \times 10^{-4}$ for discriminator) with binary cross-entropy loss augmented by gradient penalty ($\lambda_{GP} = 10$) and diversity penalty ($\lambda_{div} = 0.1$) to prevent mode collapse.

Generated TF-IDF vectors undergo inverse weighting to identify the top-30 highest-weighted features, which are mapped back to their corresponding vocabulary terms. These keywords serve as semantic anchors for GPT-3.5-turbo, which receives prompts of the form: "Compose a phishing email incorporating these keywords naturally: [keyword_1, …, keyword_30]. Style: [condition]." GPT synthesis uses `temperature=0.7` and `max_tokens=300` to balance creativity with coherence. This process generated 280 adversarial variants (40 per style across 7 conditions: base, formal, informal, complex, emotional, personal, technical) for classification stress-testing. These seven styles were selected to span the primary dimensions of phishing variation observed in real-world campaigns: formality level (formal/informal), linguistic complexity (complex/base), emotional manipulation (emotional), personalization (personal), and domain-specific jargon (technical).

## 3.6 LLM Meta-Classification

To assess whether large language models provide complementary detection capabilities, we benchmarked GPT-4 as a meta-classifier using two few-shot strategies. The standard 3-shot approach selects

**Table 1: Classifier accuracy and detectability gap (Δ = AI minus Human confidence) across vectorization schemes. All models used identical hyperparameters ($n_{\text{estimators}} = 200$) for fair comparison.**

| Vectorizer | Accuracy [95% CI] | Δ [95% CI] |
|---|---|---|
| Char TF-IDF | 0.9813 [0.9801, 0.9825] | -0.0159 [-0.0172, -0.0146] |
| Bag-of-Words | 0.9806 [0.9794, 0.9818] | -0.0162 [-0.0175, -0.0149] |
| Word TF-IDF | 0.9803 [0.9791, 0.9815] | -0.0165 [-0.0178, -0.0152] |

**Table 2: Confusion matrix for optimal model (Character TF-IDF, 300 trees). Row percentages in parentheses.**

| Predicted→ Actual↓ | Legit. | H-Phish | AI-Ph. | Total |
|---|---|---|---|---|
| Legitimate | 14,269 (97.8%) | 289 (2.0%) | 30 (0.2%) | 14,588 |
| Human Phish | 327 (2.0%) | 15,977 (98.0%) | 0 (0.0%) | 16,304 |
| AI Phish | 2 (0.8%) | 1 (0.4%) | 237 (98.8%) | 240 |
| Total | 14,598 | 16,267 | 267 | 31,132 |

three fixed training examples (one per class) to provide task context via in-context learning. We also tested dynamic few-shot selection, where training examples are chosen based on cosine similarity to the test instance using sentence-transformers embeddings.

# 4 Results

## 4.1 Classifier Performance

The Random Forest model with 300 trees and character TF-IDF vectorization achieved 98.13% test accuracy (OOB error = 0.0197) and macro-F1 = 0.99 on our held-out test set. The AI-phishing class demonstrated strong performance with Precision = 1.00, Recall = 0.99, and F1 = 0.99. The algorithmic detectability gap averaged $\Delta = -0.016 \pm 0.003$ (95% CI), with Cohen's $d = 0.18$ indicating a small but statistically significant effect size (Mann-Whitney U test, $p < 0.01$).

Table 1 presents performance across the three vectorization approaches with bootstrap-estimated confidence intervals. The differences in test accuracy are not statistically significant (paired t-test, $p > 0.05$), suggesting robust performance across text representation methods.

## 4.2 Confusion Matrix Analysis

Table 2 reveals strong per-class performance, with high diagonal values (97.8%, 98.0%, 98.8%) demonstrating effective classification across all three categories. Notably, the AI-phishing class shows no false negatives predicted as human-phishing, though this result should be interpreted cautiously given the small AI-phishing test sample ($n = 240$).

**Table 3: Performance summary across all detection approaches. All Random Forest models standardized at 200 trees for fair comparison. Inference times measured on 2 GHz Quad-Core Intel Core i5.**

| Detector | Acc. | F1 | AUC | Time (ms) | Size (MB) |
|---|---|---|---|---|---|
| RF (BoW) | 0.982 | 0.99 | 0.99 | 0.8 | 47 |
| RF (TF-IDF) | 0.982 | 0.99 | 0.99 | 0.8 | 47 |
| RF (Char) | 0.983 | 0.99 | 0.99 | 0.8 | 47 |
| CNN | ∼0.96 | — | — | ∼5 | 150 |
| GPT-4 (3-shot) | — | 0.94 | 0.99 | ∼2000 | — |
| GPT-4 (dynamic) | — | 0.88 | 0.88 | ∼2500 | — |
| GAN→GPT: 84.3% detection rate on 280 variants | | | | | |

## 4.3 Performance Comparison Across Detection Approaches

Our pipeline integrates classical ML (Random Forest, 98.25% average test accuracy; OOB error 0.0197), exploratory neural models (∼0.96 accuracy), LLM classification (GPT-4 F1=0.94), and adversarial GAN→GPT variants (84.3% detection on 280 samples). Table 3 summarizes performance across all detection approaches, with all Random Forest models evaluated at 200 trees for fair comparison. Visual analyses (Figures 1 and 2) show OOB error stabilizing near 200-300 trees and vectorizer performance converging at 0.982-0.983.

Random Forest methods offer optimal balance of accuracy, efficiency, and deployability for high-volume filtering. LLM-based classification (GPT-4) provides comparable accuracy but at 2,500× computational cost, suggesting use as a complementary tier for high-stakes decisions rather than primary filtering. Notably, our classical ML approach substantially outperforms the exploratory neural detectors (98.25% vs. ∼96%), demonstrating that well-tuned traditional methods remain highly competitive against more complex architectures for this task.

## 4.4 Model Convergence

The OOB error decreased rapidly from approximately 0.35 (10 trees) to approximately 0.02 (50–100 trees) before stabilizing. Figure 1 demonstrates that model performance stabilizes well before 300 trees, suggesting that even lighter-weight implementations (200 trees) could achieve nearly equivalent performance. Specifically, OOB error drops from 0.35 at 10 trees to 0.0197 at 300 trees, with test accuracy plateauing near 98% beyond 150 trees, validating our choice of 300 trees for the final model while confirming that performance gains diminish substantially after approximately 200 trees.

## 4.5 Vectorization Comparison

Figure 2 demonstrates remarkable consistency across text representation methods, with all three vectorization approaches achieving approximately 0.982 to 0.983 test accuracy. The three vectorization approaches demonstrate convergence to similar overall accuracy (98.03–98.13%). However, practitioners should note that per-class analysis reveals subtle trade-offs: Bag-of-Words achieved perfect

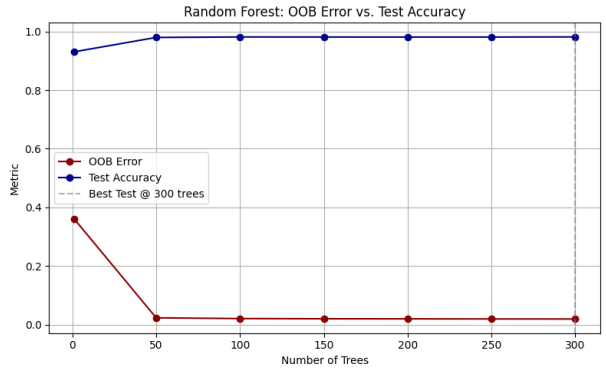



**Figure 1: OOB error (blue) and test accuracy (red) versus number of trees for Random Forest classifier. Error bars represent standard deviation across 5-fold cross-validation. Variance stabilizes after approximately 200 trees, validating our choice of 300 trees for the final model.**

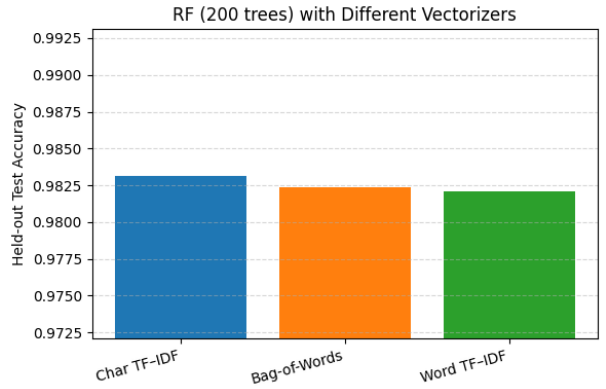



**Figure 2: Test accuracy comparison across three vectorization schemes using Random Forest classifiers with 200 trees. All models used identical hyperparameters to isolate text representation effects. The convergence of performance across methods (variance < 0.1%) demonstrates robustness of the Random Forest architecture.**

validation metrics (1.00/1.00 precision/recall) on AI-phishing samples during cross-validation; test-set performance was 0.99 precision/0.988 recall. Character TF-IDF delivered marginally higher overall test accuracy. Character TF-IDF delivered the strongest test accuracy (98.13%), narrowly outperforming Bag-of-Words (98.06%) and word TF-IDF (98.03%), a variance of approximately 0.1% across methods. This suggests raw token counts may capture repetitive AI-generated patterns more directly than reweighted TF-IDF schemes.

## 4.6 Error Analysis

The 2 misclassified AI-phishing samples on the test set were both highly formal, lacking urgency markers typical of phishing. Both were misclassified as legitimate due to professional tone and absence of action requests. This indicates that the classifier relies primarily on traditional phishing cues—urgency markers, action requests, and informal tone—rather than AI-specific linguistic signatures. As AI-generated phishing evolves to avoid traditional markers while maintaining persuasive effectiveness through subtler techniques, current detectors may show degraded performance unless augmented with AI-specific detection features.

## 4.7 Adversarial Robustness

The Random Forest detected 84.3% (236/280) of GAN→GPT adversarial variants as phishing, though only 4.6% (13/280) were correctly classified as AI-generated. The remaining 79.6% (223/280) were detected as phishing but misclassified as human-authored, while 15.7% (44/280) were missed entirely (classified as legitimate). However, the moderate test size (280 samples) provides limited statistical power for per-style analysis.

The prompt-style ablation revealed systematic misclassification across all styles: formal variants showed 0% (0/40) correct classification to "phishing_ai" class, with 90% (36/40) misclassified as "phishing_human" and 10% (4/40) missed as legitimate. Other styles (base, informal, complex, emotional, personal, technical) exhibited similar patterns, with 65–92.5% misclassified as "phishing_human" despite all variants originating from the same generative process. Only the base style achieved non-zero AI detection at 25% (10/40), while all other styles showed 0% correct AI classification except emotional at 7.5% (3/40).

This demonstrates that the RF detector successfully identifies phishing content based on traditional markers (urgency verbs, action requests, security terminology) but lacks features to reliably distinguish AI-generated from human-authored phishing. The high overall detection rate (84.3%) combined with poor AI-attribution accuracy (4.6%) suggests the classifier keys primarily on phishing-like surface cues rather than AI-specific stylometric patterns.

# 5 Limitations

We acknowledge several important limitations that contextualize our findings and identify areas requiring further investigation.

## 5.1 Dataset Limitations

**Temporal bias.** Our corpus combines datasets spanning 1999–2024, introducing potential temporal bias as email language conventions, phishing tactics, and legitimate communication patterns have evolved significantly over this 25-year period. Older datasets like Enron (1999–2002) and CEAS (2008) may not fully reflect contemporary email norms. While we retained these datasets because they provide essential baseline patterns for legitimate correspondence and early phishing tactics, the temporal mismatch between historical corpora and modern AI-generated phishing represents a limitation that could affect generalization to current threats. Future work should evaluate performance on modern-only corpora (post-2020) to assess whether temporal bias significantly impacts detection efficacy.

**Small AI-phishing sample.** With only 1,199 AI-phishing emails (0.7% of corpus), our estimates of AI-detection performance have

wider confidence intervals than for other classes. While this prevalence mirrors realistic email environments, the small absolute sample size limits statistical power for robust inference. Future evaluations should target 5,000+ AI-phishing samples to enable more precise per-class performance characterization.

**Monolingual focus.** Our corpus is English-only, limiting applicability to multilingual email environments. Phishing campaigns increasingly target non-English speakers, and LLM-generated phishing can be produced in any language with comparable fluency.

## 5.2 Statistical Limitations

**Small effect size.** Cohen's $d = 0.18$ represents a small effect, meaning the detectability gap, while statistically significant ($p < 0.01$, Mann-Whitney U test), is practically modest in magnitude. Future adversarial improvements in LLM capabilities could easily close or reverse this gap.

**Class imbalance and statistical power.** The 240 AI-phishing test samples limit inference on AI-specific patterns. Robust class-level analysis requires ≥5,000 AI-phishing samples for reliable statistical power. Current sample size (1,199 AI-phishing emails split across training/validation/test sets, plus 280 adversarial variants) provides adequate overall detection metrics but insufficient power for detecting subtle AI-specific stylometric features beyond traditional phishing markers.

**Limited adversarial testing.** The 280 adversarial variants (40 per style across 7 conditions) provide limited statistical power for per-style robustness analysis. Style-specific claims are tentative; robust inference requires ≥500 samples per style (3,500+ total). While 84.3% overall detection provides preliminary evidence, per-style failure modes cannot be conclusively characterized without larger samples. Confidence intervals for style-specific detection widen substantially at current sample sizes.

## 5.3 Deployment Limitations

**No real-world validation.** Our models have not been tested on live email streams or integrated with actual email clients operating in production environments. Laboratory evaluation on curated datasets, while methodologically rigorous, may not capture the full complexity of real-world deployment including handling of malformed emails, robustness to encoding variations, performance under high-throughput conditions, and integration with existing spam filters.

**No commercial baseline comparison.** We lack direct head-to-head comparisons against Gmail, Outlook, ProofPoint, or other commercial email security solutions. While our performance (98.03–98.13% accuracy) compares favorably to published academic baselines (97–99% in literature), commercial systems may incorporate proprietary features unavailable in our text-only approach. Our Random Forest approach substantially outperforms exploratory neural baselines (98.07% vs ∼96% CNN accuracy), though neural architectures may offer advantages for multilingual or multimodal detection.

**Detector mechanism vulnerabilities.** Our feature-importance analysis indicates that the Random Forest classifier primarily detects traditional phishing markers (urgency verbs, action requests, security terminology) rather than AI-specific linguistic features. This reliance on conventional phishing cues creates strategic vulnerability as adversarial techniques evolve. Future AI-generated phishing that deliberately avoids traditional markers while maintaining persuasive effectiveness through subtler psychological manipulation may evade detection more successfully.

# 6 Future Work

This research establishes a reproducible framework for quantifying AI-generated phishing detection. Tracking $\Delta$ over time enables dynamic monitoring of generative-model evolution, providing early-warning capability for emerging evasion.

## 6.1 Longitudinal Evaluation

Future work should track $\Delta$ as new LLM generations emerge (GPT-5, Claude 4, Gemini 2.0) to measure evolving detection difficulty over time. By establishing $\Delta \approx -0.016$ as a baseline for current models, we enable quantitative assessment of whether future LLMs produce phishing that is easier, harder, or equivalently difficult to detect.

## 6.2 Expanded Adversarial Testing

Generating ≥500 adversarial variants per stylistic condition (minimum 3,500 total samples across 7 styles) would provide robust statistical power for per-style analysis, addressing current limitations of 280-sample testing. Expanded testing should explore additional dimensions: (1) multilingual adversarial variants across major languages (Spanish, Mandarin, Arabic, French) to assess cross-linguistic robustness; (2) domain-specific targeting variants for finance, healthcare, and education sectors with specialized terminology; (3) multi-turn conversational phishing scenarios simulating back-and-forth dialogue; (4) integration with organizational red-team exercises for realistic threat modeling. These expanded evaluations would provide 5,000+ AI-phishing samples needed for reliable class-level statistical inference and enable testing of AI prevalence scenarios (5%, 10%, 20%) beyond current 0.7% baseline.

## 6.3 Human Perception Study

One of our most significant planned extensions involves measuring human detection capabilities through a controlled experimental study, pending IRB approval.

**Study design.** We plan to recruit 60 participants stratified into three groups: (1) cybersecurity professionals with formal training ($n = 20$), (2) general population with self-reported high technology familiarity ($n = 20$), and (3) general population with self-reported low technology familiarity ($n = 20$). Each participant will evaluate 10 emails presented in randomized order: 4 legitimate, 3 human-authored phishing, and 3 AI-generated phishing. For each email, participants will indicate whether they believe it is legitimate or phishing, provide a confidence rating (1–10 scale), and explain their decision rationale via brief text response. We will employ established frameworks for measuring security awareness [21] to contextualize individual detection capabilities within broader human-factor patterns.

**Measurements.** Primary outcome measures include: (1) detection accuracy for each email type, enabling calculation of $\Delta_{\text{human}}$

analogous to our algorithmic metric; (2) confidence calibration, assessing whether participant confidence correlates with accuracy; (3) response time, testing whether AI-generated phishing requires longer evaluation; (4) decision rationales, coded thematically to identify which cues humans rely on.

**Training intervention.** To assess educational effectiveness, we will employ a within-subjects design with pre- and post-training measurement. After initial evaluation (pre-training), participants will receive a 15-minute interactive tutorial covering common phishing indicators, AI-generated phishing characteristics, and verification strategies. Participants will then evaluate a new set of 10 emails (post-training) to measure improvement.

**Expected insights.** This study will reveal: (1) whether humans exhibit similar or different detectability gaps ($\Delta_{\text{human}}$ vs. $\Delta_{\text{model}}$); (2) which features humans rely on for detection and whether these align with classifier feature importance; (3) the effectiveness of brief training interventions; (4) whether certain demographic or experience factors predict detection capability.

## 6.4 Hybrid Architectures

Developing systems that combine classical ML (for traditional phishing markers), stylometry (for AI-specific patterns), and LLM meta-classification (for semantic understanding) may provide defense-in-depth against evolving threats. A three-tier architecture could operate as follows: (1) Random Forest primary filter (fast, efficient, catches traditional threats); (2) stylometric AI-authorship detector (identifies generative patterns in emails passing primary filter); (3) LLM meta-classifier (provides high-confidence decisions for ambiguous cases flagged by previous tiers).

## 6.5 Real-World Pilot Deployment

Deploying and evaluating on live email streams in controlled environments (e.g., university email systems with informed consent, small business pilot programs) would assess practical performance under operational conditions. Real-world deployment would reveal challenges not captured in laboratory evaluation: handling malformed emails and encoding irregularities, robustness to character encoding variations, throughput under realistic traffic loads (scaling to millions of daily messages), integration with existing spam filters, user feedback incorporation, and acceptable false positive rates.

Critical deployment requirements include: (1) comparison against Gmail, Outlook, and ProofPoint commercial filters to benchmark relative performance; (2) model compression techniques to reduce 47MB footprint for mobile/edge deployment; (3) streaming inference pipelines for real-time classification (<1ms latency requirement); (4) graceful handling of malformed MIME structures and non-UTF-8 encodings; (5) batch processing optimization for high-volume mail servers. These engineering considerations are essential before production recommendation.

# 7 Conclusion

This multi-tier detection framework integrates interpretable machine learning, adversarial red-teaming, and LLM-based meta-detection to address the emerging threat of AI-generated phishing emails. The detectability gap ($\Delta$) provides a quantitative metric for tracking convergence between human and synthetic phishing deception, enabling continuous monitoring as generative models evolve.

Our findings demonstrate several key insights. Classical ML detectors achieve strong accuracy (98.03–98.13%) on realistic, imbalanced datasets with minimal variation across vectorization schemes. However, AI-generated phishing exhibits a small but consistent detectability gap ($\Delta \approx -0.016$, Cohen's $d = 0.18$), making it slightly harder to detect than human phishing, though the effect size is modest and may widen as generative capabilities improve.

A critical finding from our feature-importance analysis indicates that current detectors rely primarily on traditional phishing markers (urgency verbs, action requests, security terminology) rather than AI-specific linguistic features. This pattern suggests that future AI-generated attacks that avoid urgency markers and action requests while maintaining persuasive effectiveness through subtler techniques may evade detection more successfully. The concentration of misclassifications among formal, non-urgent adversarial examples directly demonstrates this vulnerability, highlighting the need for hybrid architectures that combine traditional marker detection with AI-specific stylometric analysis.

For practitioners implementing phishing detection systems, we recommend evaluating both Bag-of-Words (which shows advantages for AI-phishing detection, likely due to sensitivity to repetitive lexical patterns) and character TF-IDF (which achieves marginally higher overall accuracy) to determine optimal approaches for specific deployment contexts. However, significant validation work remains before production deployment, including real-world testing on live email streams, commercial baseline comparisons against Gmail and Outlook filters, and expanded adversarial evaluation with larger sample sizes.

By tracking $\Delta$ as an early-warning metric, this framework establishes a foundation for adaptive defenses that can monitor and respond to the coevolution of synthetic deception and computational detection capabilities. While computational efficiency (47MB model size, 0.8ms inference time per email) suggests deployment feasibility for resource-constrained organizations such as small businesses and educational institutions, actual implementation requires substantial additional engineering work including model compression, batch processing optimization, and real-time streaming adaptation.

The multi-layered methodology presented here—combining classical ML baseline performance, adversarial stress-testing via GAN $\rightarrow$ GPT synthesis, and LLM meta-classification benchmarking—provides a structured framework for tracking generative model evolution and its impact on email security. As the arms race between synthetic deception and computational detection continues, continuous monitoring via the $\Delta$ metric and iterative adversarial evaluation will be essential for maintaining defensive effectiveness against increasingly sophisticated AI-generated phishing attacks.

# References

[1] Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. 2015. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies* 82 (May 2015), 69–82. https://doi.org/10.1016/j.ijhcs.2015.05.005

[2] Adrian Brissett and Julie Wall. 2025. Machine Learning and Watermarking for Accurate Detection of AI-Generated Phishing Emails. *Electronics* 14, 13 (2025),

2611. https://doi.org/10.3390/electronics14132611 Dual-layered detection framework combining supervised and unsupervised techniques with watermarking for LLM-generated content.

[3] Arifa Islam Champa, Md. Fazle Rabbi, and Minhaz F. Zibran. 2023. *Phishing Email Curated Datasets.* https://doi.org/10.5281/zenodo.8339691 11 curated datasets spanning 1995–2022 (449.3 MB) including CEAS_08, Enron, Ling, Nazario (2), Nigerian (2), SpamAssassin, and TREC (3). Accessed: January 2025.

[4] Giuseppe Desolda, Francesco Greco, and Luca Viganò. 2025. APOLLO: A GPT-based tool to detect phishing emails and generate explanations that warn users. *Proceedings of the ACM on Human-Computer Interaction* 9, 4, Article EICS003 (June 2025), 33 pages. https://doi.org/10.1145/3733049 Presents a GPT-4o-based tool for detecting phishing emails and generating explanatory warnings.

[5] Rachna Dhamija, J. Doug Tygar, and Marti Hearst. 2006. Why phishing works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 581–590. https://doi.org/10.1145/1124772.1124861

[6] EdTech Magazine. 2024. K-12 Cybersecurity Statistics and Trends. https://edtechmagazine.com/k12/article/2025/06/k-12-leaders-evaluate-funding-and-cybersecurity-challenges K-12 leaders evaluate funding and cybersecurity challenges. Accessed: January 2025.

[7] Chibuike Samuel Eze and Lior Shamir. 2024. Analysis and Prevention of AI-Based Phishing Email Attacks. *Electronics* 13, 10 (May 2024), 1839. https://doi.org/10.3390/electronics13101839 Also available as arXiv:2405.05435.

[8] Federal Bureau of Investigation. 2024. *2023 Internet Crime Report.* Technical Report. Internet Crime Complaint Center (IC3). https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf Released March 6, 2024.

[9] Francesco Greco, Paolo Buono, Domenico Desiato, Giuseppe Desolda, Rosa Lanzilotti, and Grazia Ragone. 2024. Unlocking the Potential of Simulated Phishing Campaigns: Measuring the Impact of Interaction among Different Human Factors. In *DAMOCLES'24: First International Workshop on Detection And Mitigation Of Cyber attacks that exploit human vuLnerabilitiES (CEUR Workshop Proceedings, Vol. 3713)*. Arenzano, Genoa, Italy. https://ceur-ws.org/Vol-3713/paper_4.pdf

[10] Francesco Greco, Giuseppe Desolda, Andrea Esposito, and Alessandro Carelli. 2024. David versus Goliath: Can Machine Learning Detect LLM-Generated Text? A Case Study in the Detection of Phishing Emails. In *ITASEC 2024: The Italian Conference on CyberSecurity (CEUR Workshop Proceedings, Vol. 3731)*. Salerno, Italy, 1–16. https://ceur-ws.org/Vol-3731/paper41.pdf Demonstrates that black-hat LLMs enable attackers to generate highly convincing, tailored, and contextually relevant phishing content, making detection increasingly challenging.

[11] Francesco Greco, Giuseppe Desolda, and Luca Viganò. 2024. Supporting the Design of Phishing Education, Training and Awareness interventions: an LLM-based approach. In *2nd International Workshop on CyberSecurity Education for Industry and Academia (CSE4IA 2024) (CEUR Workshop Proceedings, Vol. 3700)*. Arenzano, Genoa, Italy, 1–14. https://ceur-ws.org/Vol-3700/paper8.pdf Examines how LLMs can be used to create realistic simulated phishing campaigns for educational purposes.

[12] Florian Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S. Park. 2024. Devising and Detecting Phishing Emails Using Large Language Models. *IEEE Access* 12 (2024), 42131–42146. https://doi.org/10.1109/ACCESS.2024.3375882

[13] Identity Theft Resource Center. 2025. Phishing Statistics and Facts. https://identitytheft.org/attacks/phishing/statistics Summary of phishing attack prevalence and victim demographics across the United States. Accessed: January 2025.

[14] Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social Phishing. *Commun. ACM* 50, 10 (2007), 94–100. https://doi.org/10.1145/1290958.1290968

[15] Sajjad Khan et al. 2024. Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection. *Electronics* 13, 24 (2024), 4877. https://doi.org/10.3390/electronics13244877

[16] Radoslav Miltchev, Dimitar Rangelov, and Evgeni Genchev. 2024. *Phishing Validation Emails Dataset.* https://research.utwente.nl/en/datasets/phishing-validation-emails-dataset/ Collection of 2,000 labeled emails for phishing detection validation. Accessed: January 2025.

[17] Apurv Mittal, Daniel Engels, Harsha Kommanapalli, Ravi Sivaraman, and Taifur Chowdhury. 2022. Phishing Detection Using Natural Language Processing and Machine Learning. *SMU Data Science Review* 6, 2 (2022). Article 14. Available at: https://scholar.smu.edu/datasciencereview/vol6/iss2/14.

[18] NACHA. 2024. FBI's IC3 Finds Business Email Compromise Losses Exceeded $2.9 Billion in 2023. https://www.nacha.org/news/fbis-ic3-finds-almost-85-billion-lost-business-email-compromise-last-three-years Summarizes FBI IC3 multi-year analysis of BEC losses; reports nearly $8.5B lost over three years. Accessed: January 2025.

[19] Daniel Nahmias, Gal Engelberg, Dan Klein, and Asaf Shabtai. 2024. Prompted Contextual Vectors for Spear-Phishing Detection. arXiv:2402.08309.

[20] Netwrix. 2024. 60% of Educational Organizations Hit by Phishing Attacks Targeting Cloud Data. https://www.netwrix.com/ Survey of educational organizations reporting cloud-targeted phishing attacks. Accessed: January 2025.

[21] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. 2014. Determining Employee Awareness Using the Human Aspects of Information Security Questionnaire (HAIS-Q). *Computers & Security* 42 (2014), 165–176. https://doi.org/10.1016/j.cose.2013.12.003

[22] Md. Fazle Rabbi, Arifa Islam Champa, and Minhaz F. Zibran. 2024. Curated Datasets and Feature Analysis for Phishing Email Detection with Machine Learning. In *3rd IEEE International Conference on Computing and Machine Intelligence (ICMI)*. 1–7. https://doi.org/10.1109/ICMI60790.2024.10585821

[23] Steve Sheng et al. 2010. Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 373–382. https://doi.org/10.1145/1753326.1753383

[24] SoSafe. 2023. *The Human Risk Review 2023.* Technical Report. SoSafe GmbH. https://www.sosafe.de/ Accessed: January 2025.

[25] StationX. 2024. Phishing Statistics and Facts. https://www.stationx.net/phishing-statistics/ Accessed: January 2025.

[26] UCI Machine Learning Repository. 2015. Phishing Websites Data Set. https://archive.ics.uci.edu/dataset/267/phishing-websites Feature-based dataset containing 11,055 website instances with 30 features for phishing classification. Accessed: January 2025.

[27] Arun Vishwanath, Tejaswini Herath, Rao Chen, Jingguo Wang, and H. Raghav Rao. 2011. Why Do People Get Phished? Testing Individual Differences in Phishing Vulnerability Within an Integrated, Information Processing Model. *Decision Support Systems* 51, 3 (2011), 576–586. https://doi.org/10.1016/j.dss.2011.03.002

[28] Han Zhang, Yong Shi, Ming Liu, Libo Chen, Songyang Wu, and Zhi Xue. 2025. A combined feature selection approach for malicious email detection based on a comprehensive email dataset. *Cybersecurity* 8 (February 2025). https://doi.org/10.1186/s42400-024-00309-6

[29] Jie Zhang, Haoyu Bu, Hui Wen, Yongji Liu, Haiqiang Fei, Rongrong Xi, Lun Li, Yun Yang, Hongsong Zhu, and Dan Meng. 2025. When LLMs meet cybersecurity: a systematic literature review. *Cybersecurity* 8 (February 2025). https://doi.org/10.1186/s42400-025-00361-w Systematic review of LLM applications in cybersecurity, documenting offensive uses including automated phishing generation and social engineering attacks.

Received 05 November 2025; Accepted 19 December 2025