JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

IMPACT OF GENERATIVE AI ON CYBERSECURITY RISKS, PHISHING, AND DIGITAL DEFENSE IN THE US

Muhammad Imran[1]

Affiliations

[1] Department of Information Technology, Washington University of Science and Technology, Alexandria, Virginia, USA

Corresponding Author's Email

[1] imran.student@wust.edu

License:

CC BY NC SA

ABSTRACT

The rapid advancement of generative artificial intelligence (GenAI) has reduced the expenses of creating natural, personalized and context-relevant content, thus challenging defenses based on the linguistic weaknesses of traditional phishing. This paper explores the effects of GenAI on cybersecurity risk and phishing and digital defense in the United States with a structured synthesis of 23 studies and a controlled detection experiment. The synthesis is structured around offensive and defensive applications of GenAI. The experiment evaluates six different detectors (keyword rules, word and character n-gram classifiers, header/URL metadata model, metadata-plus-intent model, and fused model) against traditional phishing, GenAI-style phishing, and GenAI phishing from compromised accounts. Pre-GenAI mail was trained on the detectors, which were then evaluated at a 0.5% false-positive rate, using five different seeds. GenAI-style messages did not include any misspellings or classic lure words and were lexically nearer to legitimate messages (median cosine similarity 0.22 compared to 0.13 for traditional phishing). None of them were detected by keyword rules and the word-level classifiers dropped from 1.0 on traditional phishing to 0.508. Metadata was effective in detecting 97% of GenAI phishing messages originating from attacker infrastructure, but just 3.6% of messages originating from compromised accounts. Adding 1,500 defender-generated examples of phishing using GenAI improved the results of fused-model recall from 0.149 to 0.790 on the compromised-account case with no corresponding increase in false positives. The results suggest that GenAI has transformed the phishing threat from content to identity and context; thus, U.S. organizations need to implement email authentication, defender-side generative augmentation and continuous retraining, and account-takeover controls.

Keywords: Generative AI, Large Language Models, Phishing, Business Email Compromise, Cybersecurity Risk, Email Security, Digital Defense, United States.

I. INTRODUCTION

Within a few years, systems that generate text, code, images and voice on demand have shifted from research labs to everyday tools, known as generative AI. The same ability that can write business correspondence or summarize threat reports can write convincing fraudulent emails, adapt malware or impersonate a trusted voice [1], [2], [3]. GenAI is therefore a double-edged technology, as it extends attackers' reach and quality, but is also equipping defenders with new tools to detect, simulate and respond [4], [5], [6].

This tension is most pressing in the case of phishing. Email-based social engineering is still one of the top ways fraudsters gain access to the U.S. market and the same applies to business email compromise (BEC), which attacks payment processes in all types of organizations [7], [8]. Many defenses — secure email

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

gateways, user-awareness training and analyst triage — have been based on tell-tale indicators like misspellings or awkward grammar, generic greetings or overt urgency for two decades. Large language models (LLMs) delete these signs, almost for free, can tailor each message with public information, and can create an infinite number of variants [2], [8], [9]. It's not a question of if GenAI will impact phishing, but which defenses will become redundant and which will stay strong – and how defenders can leverage GenAI.

Much of the previous research on this issue is theoretical. The threats and opportunities of catalogue GenAI have been reviewed [1], [10], [11], [3] and studies show that chatbots can be coaxed into generating phishing content or attack code [2], [12], [13]. Under controlled settings, few studies quantify the effectiveness of existing detection techniques as the language characteristics of phishing evolve, and which mitigations bring performance back. This paper addresses that gap with the following contributions:

- Structured synthesis of 23 studies, organizing the impact of GenAI on cybersecurity into offensive and defensive uses, and linking to digital-defense outcomes in the United States.
- A controlled email corpus and evaluation protocol that differentiate three threat conditions: traditional phishing, GenAI-style phishing from attacker infrastructure and GenAI phishing from compromised accounts, and profile their linguistic, intent and infrastructure characteristics.
- A comparison of 6 detectors, with 5-seed repetition, at fixed operational false-positive rate, highlighting the detector signals that degrade and the ones that are robust.
- An assessment of two defensive strategies: augmenting training with defender-generated GenAI phishing, and selectively introducing observed attack samples, and a series of recommendations to align with U.S. security frameworks.

The literature is reviewed in Section II. This section provides a description of the methodology, followed by results (Section IV), implications for U.S. digital defense (Section V), and a conclusion (Section VI).

II. LITERATURE REVIEW

A. Generative AI and the Threat Landscape

Prior to the advent of LLMs, Bauer and Bindschaedler surveyed generative models in security and demonstrated that the same models can be used for attacks (synthetic identities, evasion) and defenses (data augmentation, anomaly detection) [4]. Kucharavy et al. provided an overview of the basics of generative LLMs and predicted that they would be leveraged in cyber-defense and cyber-offense [6]. Barrett et al. gave a systematic overview of the GenAI security issues, distinguishing between attacks against the models, including prompt injection and data extraction, and misuse of models by attackers, and described mitigations [3]. Neupane et al. evaluated the impact and risk of GenAI for cyber defence at every stage of the cyber attack life-cycle [1] and Oh and Shon outlined cybersecurity concerns for GenAI services [11]. Iqbal et al. investigated the potential of using conversational chatbots for malicious purposes such as phishing and malware generation and disinformation [2]. Teichmann demonstrated experimentally how GenAI can help in planning ransomware attacks even though they are protected [12] and Elashmawi et al. assessed GenAI-generated scripts in penetration testing [13]. On the national level, Xu et al. discussed the threat of AI systems to national security [9] while Areghan wrote a review about the evolution of threats from data breaches to deepfakes in online risks [7].

B. AI-Driven Phishing and Social Engineering

Shrestha et al. summarized phishing techniques using AI, such as automated reconnaissance, personalized phishing lures, voice and video impersonation, and defensive measures against them [8]. These reports all share a common theme: that GenAI enhances the quality of phishing messages by eliminating the mistakes that filters and users use to spot them, and that it makes them more believable by using names, roles, and projects to target specific individuals, and that it enables attackers to try out numerous variants at a low cost [2], [8], [5]. Areghan highlighted how deepfake audio and video carry over the same dynamics to vishing and executive impersonation [7] and Badmus and Adebayo discussed governance controls required to mitigate deepfake and data-privacy risks for synthetic-media deployment [14].

C. Generative AI for Digital Defense

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

Krishnamurthy [15] and Dhoni and Kumar [10] on the defensive side, explained the potential of GenAI in enhancing cyber defense and the need for the collaboration of technology firms, agencies, and government. Adeyinka and Adeyinka suggested GenAI-driven approaches for automated threat simulation and response [16], while Batel et al. used GenAI for cyber-threat-intelligence analysis and automated incident response [17]. Kumar and Bhushan reviewed the recent developments of AI in cybersecurity and fraud detection [18] while Muthusamy presented GenAI-based architectures for privacy and risk management in banking and trade systems [19]. Svendsen and Garvey evaluated the effectiveness of prompt-engineered LLMs in assisting decision-making [20]. Sarker offered an overview of multi-aspect adversarial learning and modelling in cybersecurity robustness using AI [21]. Truong et al. [22] and Bonfanti [23] discussed the off/defensive balance changes due to AI, and Aramide described automation as a double-edged sword [5].

D. Synthesis and Research Gap

The literature is summarized by theme in Table I. There are three main gaps. First, there is very little empirical research that has been conducted on the effectiveness of existing phishing defenses when it comes to GenAI-quality messages; most assertions are based on demonstrations or expert judgment [1], [2], [8]. Second, the literature often fails to distinguish between the content advantage GenAI provides for attackers and the infrastructure and identity circumstances in which messages are sent, which conditions dictate which controls remain effective. Third, the proposals to use GenAI defensively are rarely tested in the face of content generated by an attacker, rather than content generated defensively by the defender with GenAI [16], [17]. The following experiment is set up to answer all three.

Table I: THEMATIC SYNTHESIS OF THE REVIEWED LITERATURE

Theme	Principal findings and concerns	References
GenAI threat landscape	Misuse of LLMs for fraud, malware and disinformation; attacks on GenAI systems (prompt injection, data leakage); national-security implications	[1], [2], [11], [3], [4], [6], [9]
Phishing & social engineering	Fluent, personalized and scalable lures; voice/video impersonation; erosion of user-facing red flags	[2], [7], [8], [5]
Malware & offensive automation	GenAI-assisted ransomware planning and attack-script generation despite safeguards	[12], [13]
Defensive applications	GenAI for detection, threat simulation, CTI analysis, incident response, fraud detection and decision support	[10], [15], [16], [18], [17], [19], [20]
Robustness & offense–defense balance	Adversarial learning, automation as a double-edged sword, shifting advantage between attackers and defenders	[5], [23], [21], [22]
Governance & privacy	Model governance, data controls, deepfake and privacy risk, roles of government and industry	[10], [3], [14], [19]

Fig. 1: CONCEPTUAL FRAMEWORK DERIVED FROM THE LITERATURE: OFFENSIVE AND DEFENSIVE USES OF GENERATIVE AI AND THEIR COMBINED EFFECT ON U.S. DIGITAL-DEFENSE OUTCOMES.



JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

III. METHODOLOGY

A. Research Design

The study is a mixture of the literature synthesis (Section II, Fig. 1) and a controlled experiment. There are no publicly available real phishing corpora containing labelled, verified LLM-generated attacks and identical organization's legitimate mail, and it would be unethical to generate live attacks against real users. A parametric corpus generator was therefore created to be able to reproduce the properties that the literature attributes to each threat condition [2], [8] so that the effect of each threat condition on detection could be isolated. The generator changes the wording, structure, personalization and metadata, and the experiment is about detecting, not memorizing set messages.

B. Email Corpus

Four classes were created (Table II). Legitimate mail includes meetings, invoices, IT notices, HR announcements, document sharing and vendor notifications, and conveys urgency, payment references, MFA notices and links—just like real business mail. Traditional phishing utilizes generic greetings, brand spoofing, explicit urgency and threats, misspellings and grammatical mistakes at 4–15% of words. GenAI phishing mirrors the characteristics of the lures created by LLMs [2], [8]: error-free copy, personalized greetings, mentions of colleagues, vendors and internal projects, soft deadlines rather than threats, and business-context pretexts (vendor bank-detail change, MFA re-enrollment, shared document, HR or payroll update, and executive request). GenAI + compromised account is the same content delivered through a previously known account that is legitimately authenticated, often within an existing thread, and is account takeover and thread hijacking. 12 header/url attributes are included in each email: SPF/DKIM/DMARC result, first time sender, sender-domain age, display-name mismatch, reply-to mismatch, link-domain mismatch, newly registered link domains, URL shorteners, existing thread membership, HTML attachments, off-hours sending, and number of links. In the different classes, there are different distributions of infrastructure. Compared to traditional phishers, GenAI attackers use more sophisticated look-alike domains and compromised accounts pass authentication and look like known senders.

A second generator creates 1,500 synthetic GenAI phishing samples with different wording and some different pretexts (vendor bank change, MFA, shared document, payroll) for the defender to test. Test time GenAI attacks are solely from the attacker's generator and augmentation is tested on content that the defender hasn't seen.

Table II: EMAIL CORPUS DESIGN (PER SEED)

Class	Train / val. / test	Content characteristics	Infrastructure profile
Legitimate	6,000 / 2,000 / 4,000	Business correspondence; some urgency, links, payment and MFA references	Authenticated, established domains, known senders
Traditional phishing	3,000 / 1,000 / 1,000	Generic greetings, overt urgency and threats, brand lures, 4–15% misspelled words	Often unauthenticated, young domains, shorteners, mismatched links
GenAI phishing	— / — / 1,000	Error-free, personalized, business pretexts, soft deadlines, credential or payment intent	Better-configured look-alike domains; mostly first-time senders
GenAI + compromised account	— / — / 600	As GenAI phishing; frequently a reply in an existing thread	Fully authenticated, known sender, old domain
Defender synthetic GenAI	1,500 (augmentation only)	Different phrasing and partially different pretexts from attacker generator	As GenAI phishing

Note: An additional pool of 1,400 attacker GenAI messages provides the observed samples in Section IV-D. Five seeds regenerate all sets.

C. Detectors

A comparison was made between six detectors that represent a variety of email-security solutions. (i) Keyword rules score emails based on the number of classic lures (e.g., "dear customer", "verify your account",

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

"suspended", "click here") and the misspelled-token rate, similar to legacy gateway rules. (ii) Word TF-IDF + LR is the logistic regression using word unigrams and bigrams. (iii) Char n-gram + LR: character 3–5-grams that model misspellings and obfuscation. (iv) Metadata: A gradient-boosted tree model (XGBoost) trained on the 12 header and URL features. (v) Metadata + intent provides eight general, or content independent, features: Indicators of credential requests, payment-change requests, secrecy, document lures, overt and soft urgency, misspelled-token rate, and greeting personalization. (vi) Gradient-boosted the fused stacks out-of-fold probabilities from the two lexical models with the metadata and intent features. The phishing score s(x) of each detector is thresholded at τ, the 99.5th percentile of scores on legitimate validation mail:

$$\tau = Q0.995(\{ s(x) : x \in Vlegit \}), \quad Recallc = | \{ x \in Tc : s(x) > \tau \} | / | Tc | \qquad (1)$$

To ensure that each detector has a 0.5% false positive rate (FPR), which is in line with enterprise mail volumes. The lexical proximity to the legitimate mail was calculated as the cosine similarity between the TF-IDF vector of each message and the centroid of all the legitimate test mail.

D. Experimental Protocol

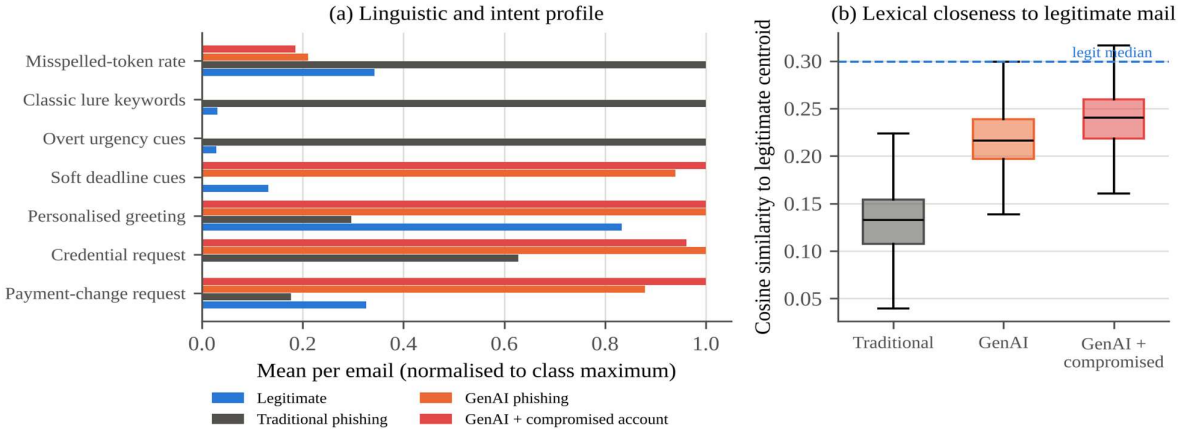
Three regimes were evaluated. The detectors were trained in the pre-GenAI era, against legitimate and traditional phishing emails, which were models built prior to the widespread use of LLM-generated lures. For synthetic augmentation, the 1,500 defender-generated GenAI examples are added to training. In the observed-sample regime, k ∈ {25, 100, 400, 800} attacker GenAI messages drawn from only two pretexts (vendor bank change and MFA re-enrollment) are added, and recall is measured on the three pretexts not seen in training. Each regime was replicated five times across the five different seeds which regenerate the corpus, and the results are presented as mean ± SD. Fused-model attributions were derived using TreeSHAP and categorized as lexical, metadata and intent/style evidence. For implementation, Python 3 was used, along with scikit-learn 1.9, XGBoost 3.4 and SHAP 0.54.

IV. RESULTS

A. How GenAI Changes the Profile of Phishing

The four classes are compared in Fig. 2. The traditional phishing signals that defenders are familiar with include a misspelled-token rate of 7.5%, an average of 1.9 classic lure phrases and 1.4 overt urgency phrases per message and personalized greetings in only 30% of messages. GenAI-style phishing turns this profile on its head. It contains no lures, no sense of urgency, a misspelled-token rate less than that of legitimate mail (1.6% vs. 2.6%) and a personalized salutation for each message. But its risk is borne by intent: credential-entry requests (0.57 per message) and payment-change requests (0.77) that are cloaked in soft, business-like deadlines (0.76). Lexically, GenAI phishing is much closer to legitimate mail (median cosine similarity 0.216) than traditional phishing (0.133), and the compromised-account variant (which includes quoted thread history) is even closer (0.240; legitimate median 0.299).

Fig. 2: PROFILES OF THE FOUR EMAIL CLASSES ON THE TEST SET: (A) MEAN LINGUISTIC AND INTENT INDICATORS, NORMALIZED TO THE MAXIMUM ACROSS CLASSES; (B) COSINE SIMILARITY OF EACH MESSAGE TO THE CENTROID OF LEGITIMATE MAIL.



JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

B. Detector Performance Against GenAI Phishing

The results from pre-GenAI training are presented in Table III and Fig. 3a. All learned detectors were able to accurately detect traditional phishing, while keyword rules detected 78%. The situation with GenAI phishing took a dramatic turn. None were detected by keyword rules (AUC 0.356, below chance level since the GenAI lures do not contain the exact wording that the rules look for). Recall fell to 0.508 ± 0.042 for the word model and 0.352 ± 0.050 for the character model. The character model, which most benefitted from misspellings, degraded most. Metadata based detection was only marginally affected (0.971 ± 0.010), as the attacker infrastructure (new domains, first-time senders, mismatching links) still differs from the legitimate correspondents.

The compromised-account condition was successful against nearly all detectors. Metadata recall drops to 3.6%, word-model recall to 35.4% and the fused model to just 14.9% when the GenAI content comes from an authenticated known account in an existing thread. The most significant outcome of the experiment. GenAI eliminates the content signals, while ATK eliminates the infrastructure signals and combined they leave conventional detectors with little to work with. The secondary finding is that a strong signal can be weakened through naive fusion: the pre-GenAI fused model performed slightly worse than the metadata alone model (0.913 versus 0.971) on the GenAI phishing task. Fig. 4 explains why. Its lexical inputs score GenAI messages as legitimate, which pulls part of the score distribution toward zero, whereas traditional phishing remains cleanly separated.

Table III: DETECTION PERFORMANCE AT 0.5% FALSE-POSITIVE RATE (MEAN ± SD, FIVE SEEDS)

Detector	Recall: traditional	Recall: GenAI	Recall: GenAI + CA	AUC: GenAI	AUC: GenAI + CA	Test FPR (%)
Trained on pre-GenAI mail						
Keyword rules	0.784 ± 0.007	0.000 ± 0.000	0.000 ± 0.000	0.356 ± 0.006	0.342 ± 0.004	0.35
Word TF-IDF + LR	1.000 ± 0.000	0.508 ± 0.042	0.354 ± 0.023	0.913 ± 0.003	0.870 ± 0.010	0.55
Char n-gram + LR	1.000 ± 0.000	0.352 ± 0.050	0.120 ± 0.012	0.914 ± 0.004	0.811 ± 0.009	0.52
Header/URL metadata	1.000 ± 0.000	0.971 ± 0.010	0.036 ± 0.025	0.999 ± 0.000	0.650 ± 0.010	0.70
Metadata + intent	1.000 ± 0.000	0.949 ± 0.007	0.009 ± 0.010	0.996 ± 0.001	0.546 ± 0.060	0.63
Fused	1.000 ± 0.000	0.913 ± 0.037	0.149 ± 0.077	0.991 ± 0.006	0.645 ± 0.026	0.46
+ 1,500 defender-synthesized GenAI phishing						
Keyword rules	0.784 ± 0.007	0.000 ± 0.000	0.000 ± 0.000	0.356 ± 0.006	0.342 ± 0.004	0.35
Word TF-IDF + LR	1.000 ± 0.000	0.877 ± 0.015	0.849 ± 0.010	0.989 ± 0.001	0.989 ± 0.001	0.65
Char n-gram + LR	1.000 ± 0.000	0.908 ± 0.026	0.780 ± 0.032	0.998 ± 0.001	0.992 ± 0.001	0.52
Header/URL metadata	1.000 ± 0.000	0.982 ± 0.004	0.025 ± 0.016	0.999 ± 0.000	0.570 ± 0.025	0.60
Metadata + intent	1.000 ± 0.000	0.981 ± 0.006	0.257 ± 0.079	0.999 ± 0.000	0.756 ± 0.030	0.56
Fused	1.000 ± 0.000	0.985 ± 0.002	0.790 ± 0.020	0.999 ± 0.000	0.943 ± 0.026	0.73

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

Note: CA = compromised account. AUC is computed against legitimate test mail. Bold marks the best value per column within each training regime.

Fig. 3: RECALL AT A 0.5% FALSE-POSITIVE RATE BY DETECTOR AND THREAT CLASS (MEAN ± SD OVER FIVE SEEDS): (A) DETECTORS TRAINED ON PRE-GENAI MAIL; (B) TRAINING AUGMENTED WITH 1,500 DEFENDER-SYNTHESIZED GENAI PHISHING MESSAGES.

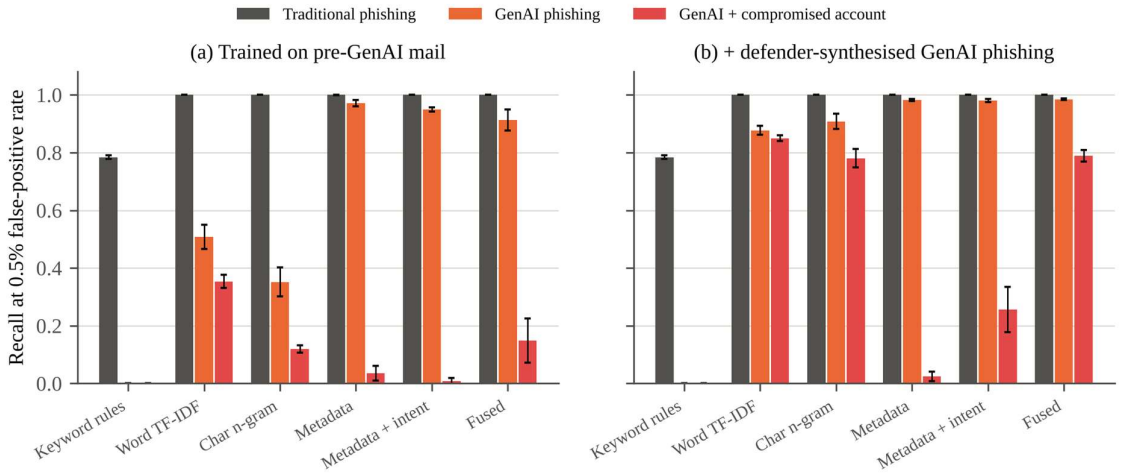


Fig. 4: DISTRIBUTION OF PHISHING SCORES ON THE TEST SET UNDER PRE-GENAI TRAINING (SEED 0, LOG SCALE): (A) WORD TF-IDF MODEL; (B) FUSED MODEL.

C. Defender-Side Generative Augmentation

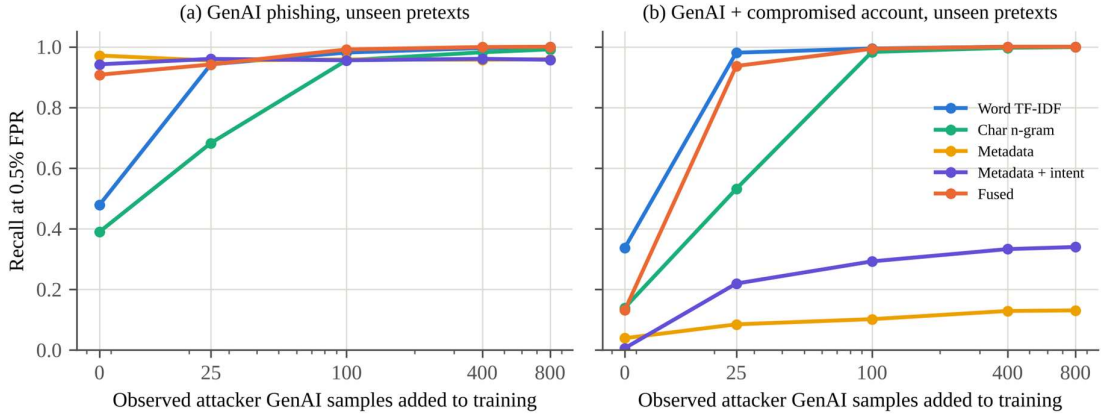
Fig. 3b and the bottom block of Table III demonstrate the impact of training on 1500 defender-generated GenAI phishing messages using another phrasing than the one used by the attacker. For the word model, the recall on attacker GenAI increased from 0.508 to 0.877, and for the character model, the recall increased from 0.352 to 0.908. The fused model reached 0.985 ± 0.002. The same was true for the compromised-account condition; augmentation improved fused recall to 0.790 ± 0.020, and word-model recall to 0.849, as content was now the source of the signal that infrastructure was no longer able to supply. Metadata only could not benefit (0.025). Test false-positive rates remained near the 0.5% goal (0.73% for the augmented fused model, compared to 0.46% without augmentation), which means that the detectors did not become any more suspicious of ordinary business mail as a result of augmentation. Per-pretext results (Fig. 6b) indicate that the defender's generator was able to make the largest gains for MFA re-enrollment and HR pretexts, while the recall was lowest for the shared-document lures (word-model recall 0.540), which are very similar to real collaboration notices.

D. Adaptation From Observed Attacks

Recall on GenAI pretexts not observed among the samples is shown as k attacker messages from two other pretexts are added to training. Lexical models adapted rapidly: Word-model recall on unseen pretexts increased from 0.478 to 0.942 based on 25 observed messages, and on the compromised-account condition increased from 0.337 to 0.980. It took approximately 100 samples to create the character model. Little adaptation was done on content-agnostic detectors on the compromised-account condition (metadata + intent reached 0.339 at k = 800). This rapid transfer should be interpreted as an upper bound. The attacker generator is not uniform but has stylistic regularities which would not be found in real LLM output sampled with different prompts. The qualitative conclusion is still strong, however: quick turnaround of reported or quarantined phishing into the model retraining process is one of the best countermeasures defenders have.

Fig. 5: RECALL ON GENAI PHISHING FROM PRETEXTS NOT PRESENT IN THE OBSERVED SAMPLES, AS ATTACKER MESSAGES FROM TWO OTHER PRETEXTS ARE ADDED TO TRAINING (MEAN OVER FIVE SEEDS): (A) ATTACKER INFRASTRUCTURE; (B) COMPROMISED ACCOUNTS.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/



E. What the Fused Model Relies On

The mean absolute SHAP attribution of the fused model is decomposed in Fig. 6a. Lexical evidence is the predominant type of evidence (82% of attribution) for traditional phishing. In the context of GenAI phishing, the model's weight is increased for the metadata of the headers and URLs (from 29% without pre-GenAI training to 41% with augmentation), with the age of the sender domain being the most influential single metadata feature. When the lexical models are available, almost no credit was attributed to the hand-crafted intent, since the same requests are already contained in the lexical models. However, the intent features did improve the content-agnostic detector on the compromised-account condition when augmented (0.257 vs. 0.025 for metadata alone). This indicates a potential application of semantic intent analysis in cases where content cannot be stored or reviewed completely.

Fig. 6: (a) SHARE OF THE FUSED MODEL'S MEAN |SHAP| ATTRIBUTION BY EVIDENCE GROUP AND THREAT CLASS (CA = COMPROMISED ACCOUNT); **(B)** RECALL ON GENAI PHISHING BY PRETEXT FOR THE WORD AND FUSED MODELS, BEFORE AND AFTER SYNTHETIC AUGMENTATION (SEED 0).

V. DISCUSSION

A. Impact on U.S. Cybersecurity Risk

The findings corroborate the literature's assertion that GenAI is altering the economics and nature of phishing instead of introducing a novel threat [1], [2], [8]. It is primarily used to counteract the linguistic heuristics that are the basis of keyword filters, character-level classifiers and much awareness training. Traditional phishing rules were unable to detect any of the GenAI variants in the experiment, which achieved a 78% detection rate. The risk is disproportionately on the small and midsize U.S. organizations that are still using these heuristics, including local governments and individual users [10], [9]. The second effect is the risk is moved towards identity. With the GenAI content and a compromised mailbox, the typical thread hijacking BEC scenario, virtually all technical signals vanish. For U.S. organizations, the most impactful GenAI-era control could therefore be account-takeover prevention (phishing-resistant MFA, anomalous-login detection, mailbox-rule monitoring) rather than any adjustments to content filtering.

B. Implications for Digital Defense

Table IV summarizes the results and provides recommendations that correspond to popular U.S. guidance such as the NIST Cybersecurity Framework 2.0 and the NIST AI Risk Management Framework. Four points need to be emphasized. Email authentication (SPF, DKIM, DMARC enforcement) and domain-reputation signals continue to be effective controls against GenAI phishing from attacker infrastructure and should be treated as a base-line control. Secondly, defenders should employ GenAI as well. In-house synthesized phishing, using intentional variations in both pretext and style, significantly enhanced detection of hidden attacker content, aligning with ideas for using GenAI to create threat simulations [16], [17]. Third, detection pipelines must have short feedback loops: user reports and quarantined messages should feed into the retraining process within days; a few dozen examples of a new family of lures can make a difference in the recall. Fourth, awareness training should change from “spot the typo” to “verify the request”: out-of-band

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

confirmation of payment changes, MFA prompts and credential requests, whatever the quality of the message [8], [18]. The governance of Defender-side GenAI, which includes data controls, model risk management, and safeguarding models against prompt injection, is also necessary [3], [14].

Table IV: IMPLICATIONS AND RECOMMENDED CONTROLS FOR U.S. ORGANIZATIONS

GenAI-era risk	Evidence (this study)	Recommended control	Owner / framework function
Fluent, personalized lures defeat keyword and spelling heuristics	Keyword rules: 0% recall on GenAI phishing; char-model recall 0.35	Retire error-based filtering as a primary control; adopt ML detectors retrained on current lures	Email providers, enterprise security teams (CSF: Detect)
Attacker infrastructure still distinguishable	Metadata recall ≈ 0.97 on GenAI phishing from attacker domains	Enforce SPF/DKIM/DMARC; weight domain age, first-contact and link reputation	Domain owners, mail operators (CSF: Protect)
Compromised accounts remove infrastructure signals	Metadata recall ≈ 0.04; pre-GenAI fused recall ≈ 0.15	Phishing-resistant MFA, impossible-travel and mailbox-rule monitoring, out-of-band payment verification	Identity teams, finance (CSF: Protect, Detect)
Defender-side GenAI improves robustness	Fused recall on GenAI + CA 0.15 → 0.79 with synthetic data	Generate diverse synthetic phishing for training and awareness drills; govern that use under an AI risk framework	SOC, ML teams (NIST AI RMF: Map, Measure)
Rapid adaptation from observed attacks	25 observed lures lift word-model recall on unseen pretexts 0.48 → 0.94	Route user reports and quarantine into weekly retraining; share indicators across sectors	SOC, ISACs, public–private partnerships
Users cannot rely on spotting errors	GenAI phishing has fewer misspellings than legitimate mail	Reframe awareness training around verifying requests (payments, MFA, credentials) rather than spotting typos	HR, security awareness programs (CSF: Protect)

Note: CSF = NIST Cybersecurity Framework 2.0; AI RMF = NIST AI Risk Management Framework; ISAC = Information Sharing and Analysis Center.

C. Offense–Defense Balance

Here is a debate among analysts on whether AI is an advantage for the attacker or the defender [23], [22]. The experiment implies that the equilibrium is dependent on the speed of feedback and access to data. Attackers benefit in the short term due to the impact of GenAI on detectors trained on past phishing attacks. Defenders with high mail volume, user-reporting programs, and the capacity to create their own synthetic attacks can easily make up for the loss in a short period of time. Organizations that do not have these capabilities can't. This asymmetry is a reason to believe that there is a role for "threat sharing" among industry, managed detection services and public-private initiatives, in which smaller U.S. entities benefit from the size of their larger counterparts [10], [9].

D. Limitations

The experiment did not involve live LLM output or actual organizational mail, but instead relied on a parametric generator. It shares the literature's characteristics of reproducing the properties of GenAI phishing, but its vocabulary is more limited than real models, which likely overestimates the efficacy of lexical adaptation (Fig. 5) and the separability of the classes. The metadata distributions are guesses, not measurements of U.S. mail traffic. The experiment did not explore human susceptibility, voice and video deepfakes, and attacks on the GenAI systems themselves. Some of the referenced studies are reviews or preprints. The quantitative results should then be interpreted as evidence about mechanisms—the signals that GenAI is eroding and the defenses that are compensating—and should be tested on real email data and using LLM-generated attacks, under suitable ethical precautions

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 2, Issue 1, 2024
https://jecir.com/

VI. CONCLUSION

The authors conducted a structured literature synthesis and controlled detection experiment to explore the effects of generative artificial intelligence on cybersecurity risk and phishing and digital defense in the U.S. GenAI phishing bypassed the misspellings, the generic greetings and overt urgency that many defenses rely on: keyword rules caught none of it, and lexical classifiers trained on past phishing caught ~50%. Signals associated with infrastructure and identity were resilient to attacks from attacker-hosted campaigns, but were not resilient to attacks from compromised accounts which most closely resemble modern BEC. Defenders compensated for much of the loss by training their own synthetic GenAI phishing and quickly adding observed attacks. For organizations in the U.S., the top priorities include phishing-resistant authentication and account-takeover detection, enforcing email authentication, using GenAI at the defender's end and continuous retraining, and awareness initiatives focused on verifying email requests instead of detecting mistakes. Future research should be conducted on production email data and live LLM-generated attacks and expand the analysis to voice and video deepfakes, along with measuring human susceptibility under the same conditions.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] S. Neupane, I. A. Fernandez, S. Mittal, and S. Rahimi, "Impacts and risk of generative AI technology on cyber defense," 2023, arXiv:2306.13033.
[2] F. Iqbal, F. Samsom, F. Kamoun, and Á. MacDermott, "When ChatGPT goes rogue: Exploring the potential cybersecurity threats of AI-powered conversational chatbots," Front. Commun. Netw., vol. 4, Art. no. 1220243, 2023.
[3] C. Barrett et al., "Identifying and mitigating the security risks of generative AI," Found. Trends Privacy Secur., vol. 6, no. 1, pp. 1–52, 2023.
[4] L. A. Bauer and V. Bindschaedler, "Generative models for security: Attacks, defenses, and opportunities," 2021, arXiv:2107.10139.
[5] O. O. Aramide, "AI-driven cybersecurity: The double-edged sword of automation and adversarial threats," Int. J. Humanities Inf. Technol., vol. 4, no. 4, pp. 19–38, 2022.
[6] A. Kucharavy et al., "Fundamentals of generative large language models and perspectives in cyber-defense," 2023, arXiv:2303.12132.
[7] E. Areghan, "From data breaches to deepfakes: A comprehensive review of evolving cyber threats and online risk management," Commun. Phys. Sci., vol. 9, no. 4, pp. 538–553, 2023.
[8] L. Shrestha, H. Balogun, and S. Khan, "AI-driven phishing: Techniques, threats, and defence strategies," in Proc. Int. Conf. Global Secur., Safety, Sustainability. Cham, Switzerland: Springer, Oct. 2023, pp. 121–143.
[9] L. Xu, E. Qereshniku, H. Hazari, and M. Edwards, "Understanding national security threats enabled by artificial intelligence: Implications for CSIS," Artif. Intell., vol. 101, no. 9, 2023.
[10] P. S. Dhoni and R. Kumar, "Synergizing generative artificial intelligence and cybersecurity: Roles of generative artificial intelligence entities, companies, agencies and government in enhancing cybersecurity," J. Global Res. Comput. Sci., vol. 14, no. 3, pp. 1–16, 2023.
[11] S. Oh and T. Shon, "Cybersecurity issues in generative AI," in Proc. Int. Conf. Platform Technol. Service (PlatCon), Aug. 2023, pp. 97–100.
[12] F. Teichmann, "Ransomware attacks in the context of generative artificial intelligence—An experimental study," Int. Cybersecur. Law Rev., vol. 4, no. 4, pp. 399–414, 2023.
[13] W. H. Elashmawi, H. Osman, M. Osama, and N. Nader, "Development generative AI for cybersecurity: Evaluating script generation and attack classification in penetration testing," in Proc. Int. Workshop Adv. Inf. Secur. Manage. Appl. Cham, Switzerland: Springer, Sep. 2023, pp. 48–62.
[14] A. Badmus and M. E. Adebayo, "Compliance-aware DevOps for generative AI: Integrating legal risk management, data controls, and model governance to mitigate deepfake and data privacy risks in synthetic media deployment," Scholars Int. J. Law, Crime Justice, vol. 3, no. 12, pp. 468–478, 2020.

[15] O. Krishnamurthy, "Enhancing cyber security enhancement through generative AI," Int. J. Universal Sci. Eng., vol. 9, no. 1, pp. 35–50, 2023.
[16] T. I. Adeyinka and K. I. Adeyinka, "Leveraging generative AI for automated cyber threat simulation and response frameworks," SSRN, 2023, Paper 5334064.
[17] N. Batel, Q. Zhou, and P. Wagner, "Generative AI-assisted cyber threat intelligence analysis and automated incident response," J. Adv. Artif. Intell. Res., vol. 2, no. 1, 2023.
[18] K. Kumar and B. Bhushan, "Augmenting cybersecurity and fraud detection using artificial intelligence advancements," in Proc. Int. Conf. Comput., Commun., Intell. Syst. (ICCCIS), Nov. 2023, pp. 1207–1212.
[19] M. Muthusamy, "Generative AI–powered cloud and machine learning architectures for digital privacy and risk management in banking and trade systems over 5G networks," Int. J. Adv. Res. Comput. Sci. Technol., vol. 6, no. 4, pp. 8681–8690, 2023.
[20] A. Svendsen and B. Garvey, "Prompt-engineering testing ChatGPT4 and Bard for assessing generative-AI efficacy to support decision-making," SSRN, 2023, Paper 4495320.
[21] I. H. Sarker, "Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview," Secur. Privacy, vol. 6, no. 5, Art. no. e295, 2023.
[22] T. C. Truong, Q. B. Diep, and I. Zelinka, "Artificial intelligence in the cyber domain: Offense and defense," Symmetry, vol. 12, no. 3, Art. no. 410, 2020.
[23] M. E. Bonfanti, "Artificial intelligence and the offence–defence balance in cyber security," in Cyber Security: Socio-Technological Uncertainty and Political Fragmentation. London, U.K.: Routledge, 2022, pp. 64–79.