AegisAI

State of the AI Threat in Email: 2025

How AI-Generated Phishing Undermines Heuristic Email Defences

Over 20,000 phishing emails analyzed

March 2026

5 minutes. 5 prompts.

AI generates a sophisticated phishing campaign in 5 minutes – a task that took human experts 16 hours.

50.3%	72.6%	22%	93.9%
AI phishing bypass rate	DMARC pass rate (evaded AI spear)	of phishing volume is AI-generated	NLP filter evasion rate

Prepared for the Messaging, Malware and Mobile Anti-Abuse Working Group (M3AAWG)

Table of Contents

Executive Summary

This paper presents findings from a sample of over **20,000 distinct phishing emails** collected across production email environments. The dataset was segmented by targeting methodology (spear phishing vs. mass phishing) and by content origin (AI-generated vs. human-written), enabling direct comparison of attacker behavior, infrastructure choices, and filter evasion rates across cohorts.

Three findings warrant immediate attention from the email security community.

1. AI-generated phishing now constitutes **22% of observed volume** and bypasses tier-1 mail filters (Gmail and Microsoft) at a rate of **50.3%** – nearly double the **28.5%** bypass rate of human-written phishing. This gap emerges from the convergence of grammatically flawless content, compromised high-reputation sending infrastructure, and strategic timing designed to exploit periods of reduced security operations staffing.

2. Evaded AI spear phishing emails showed a **72.6% DMARC pass rate**. Successful attacks were **17.8% more likely** to present valid DMARC than those that were blocked. When attackers operate from compromised legitimate accounts or allow-listed infrastructure, SPF, DKIM, and DMARC authenticate the attack.

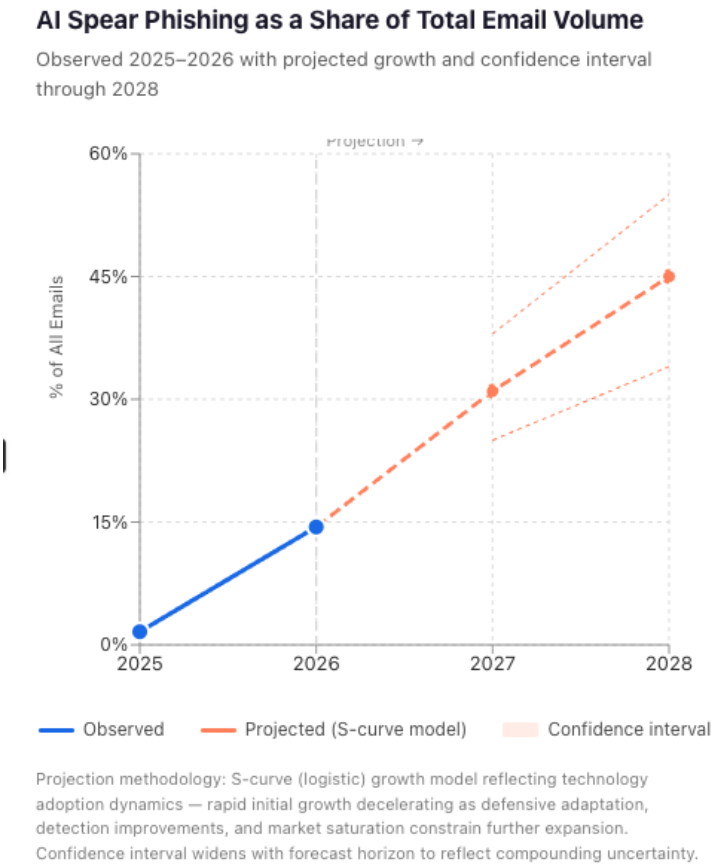
3. 75.7% of domains used for AI spear phishing were exclusive to that cohort – not shared with noisier mass-phishing campaigns. AI-driven attackers are treating sending infrastructure as a managed resource, not a disposable commodity.

These findings indicate that AI has shifted the economics of phishing from volume-dependent operations to precision-engineered campaigns that systematically exploit the assumptions underlying current email defense architectures.

Key Findings

1. AI-generated phishing accounts for 22% of observed volume and is disproportionately concentrated in spear phishing.

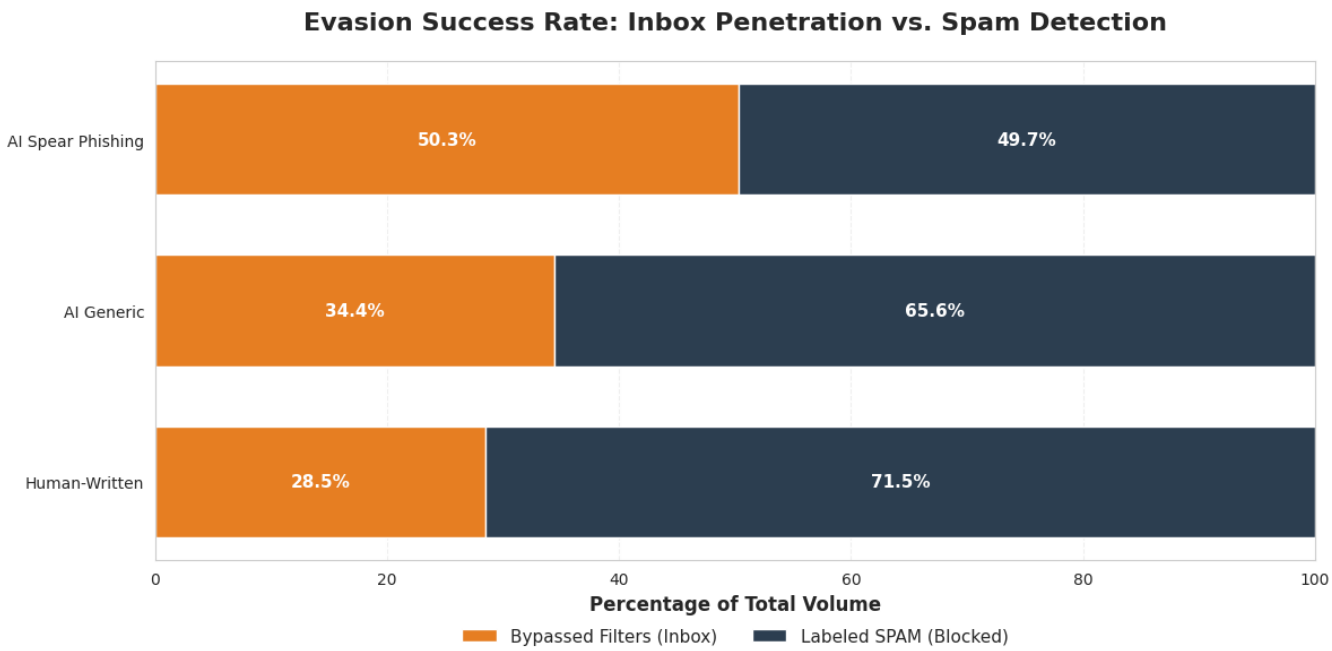
Of the sampled AI-generated emails, 62.9% were targeted spear phishing – a ratio that inverts the overall dataset. This concentration suggests AI is being deployed preferentially where precision matters.



[Figure 1: Campaign Overview]

2. AI-generated emails bypass Gmail and Microsoft spam filters 50.3% of the time.

This is compared to 28.5% for human-written phishing – a 21.8 percentage point evasion advantage.



[Figure 2: Evasion Analysis – AI vs. Human-Written bypass rates]

3. Evaded AI spear phishing passes DMARC at 72.6%.

DMARC-passing emails are 17.8% more likely to reach the inbox than those that fail.

4. AI spear phishing reuses sending domains at the highest rate of any cohort (10.9% diversity ratio).

This includes reliance on a small set of high-value compromised domains rather than disposable infrastructure.

5. Specific LLM artifacts were detected in 95 emails.

This includes 32 template errors, 25 ChatGPT/OpenAI phrasing signatures, 23 generic AI patterns, and 15 Claude signatures.

6. AI spear phishing peaks on Fridays at 17:00 UTC.

AI generic phishing peaks on Mondays at 17:00 UTC. These timing patterns are consistent with deliberate exploitation of reduced weekend SOC staffing and Monday morning inbox congestion.

7. AI spear phishing impersonates Google Workspace at 1.8x and B2B SaaS brands at 2.1x the rate of human-written spear phishing.

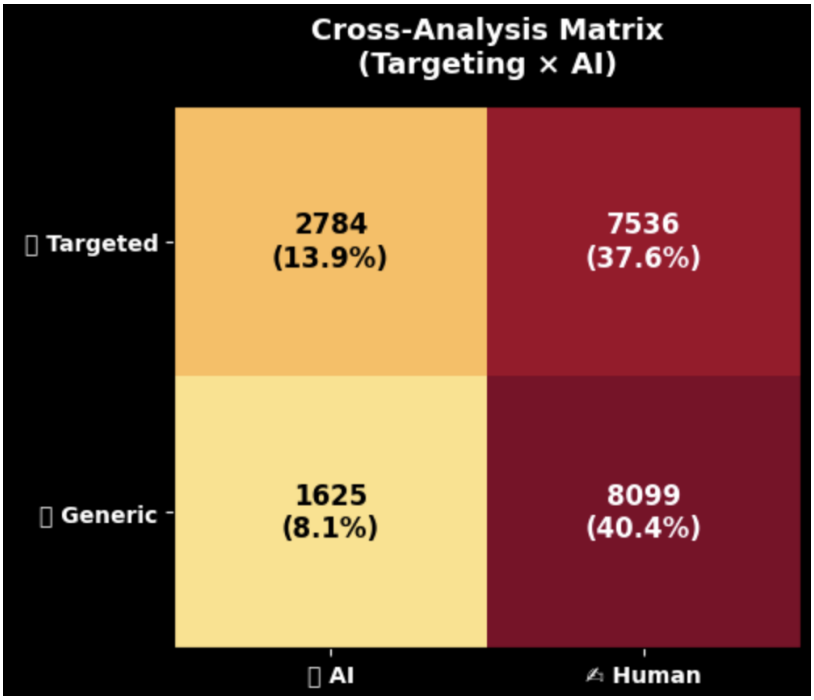
This aligns with SaaS/webmail being the single most-targeted sector at 21.2% of all phishing in Q3 2025.

01

Dataset and Methodology

The analysis examined a sample of over **20,000 phishing emails** collected from production environments protected by an AI-native email defense platform. Each email was classified along two independent axes: **targeting methodology** (spear phishing vs. mass/generic phishing) and **content origin** (likely AI-generated vs. likely human-written).

This produced four analytical cohorts for cross-comparison:



[Figure 1: Campaign Overview – Volume, targeting, and AI-generation breakdown]

Methodological limitations. AI-generation classification is probabilistic, not deterministic. The 22% figure represents emails assessed as likely AI-generated based on observable features. The dataset reflects the threat landscape as observed through one platform's collection.

The paper's finding that **22% of phishing volume is AI-generated** lands between sharply divergent industry estimates (0.7–4.7% from Hoxhunt, ~12% from Mimecast, 40% of BEC from VIPRE, 82.6% from KnowBe4). The 22% figure is most plausible as a measure of *substantially* AI-generated phishing.

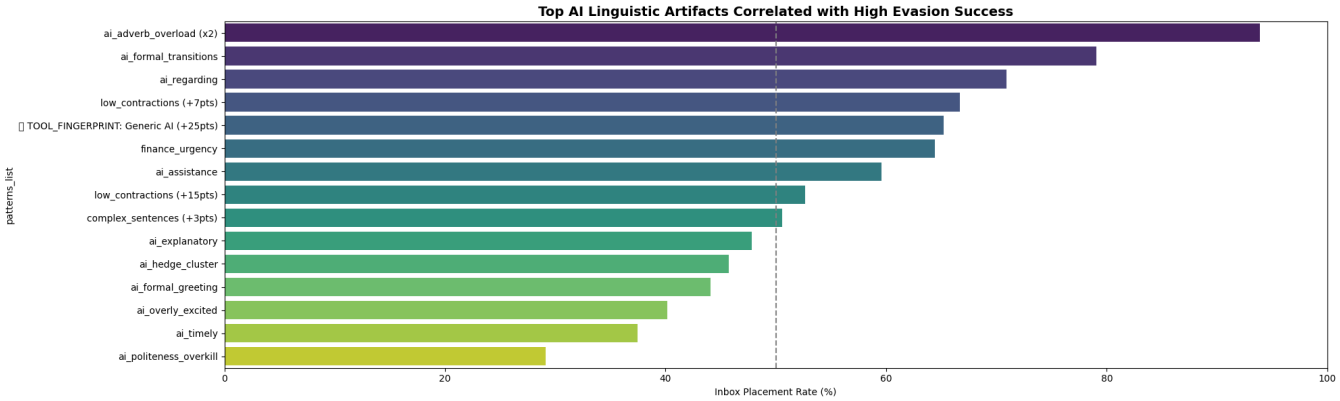
02

Technical Analysis

2.1 The Evasion Gap: AI vs. Human Performance Against Tier-1 Filters

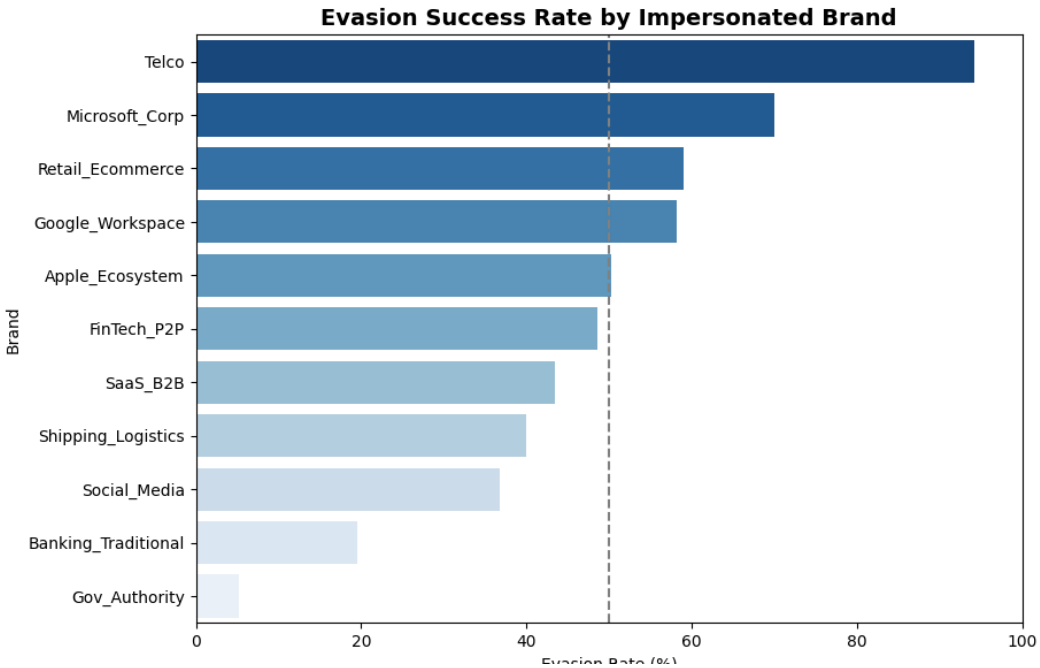
AI spear phishing emails bypassed Gmail and Microsoft spam filters **50.3%** of the time. Human-written phishing bypassed the same filters **28.5%** of the time. The resulting **21.8 percentage point advantage** is not explained by any single variable. Our analysis identifies three reinforcing mechanisms.

Mechanism 1: Content Sanitization. LLMs produce grammatically correct, professionally toned text that avoids the linguistic markers historically used by NLP-based spam filters. In our dataset, AI-generated spear phishing achieved a **93.9% evasion rate** against NLP-based content filters. These emails averaged **562 words** – substantially longer than typical phishing.



[Figure 3: Linguistic evasion, infrastructure exploitation, and contextual manipulation]

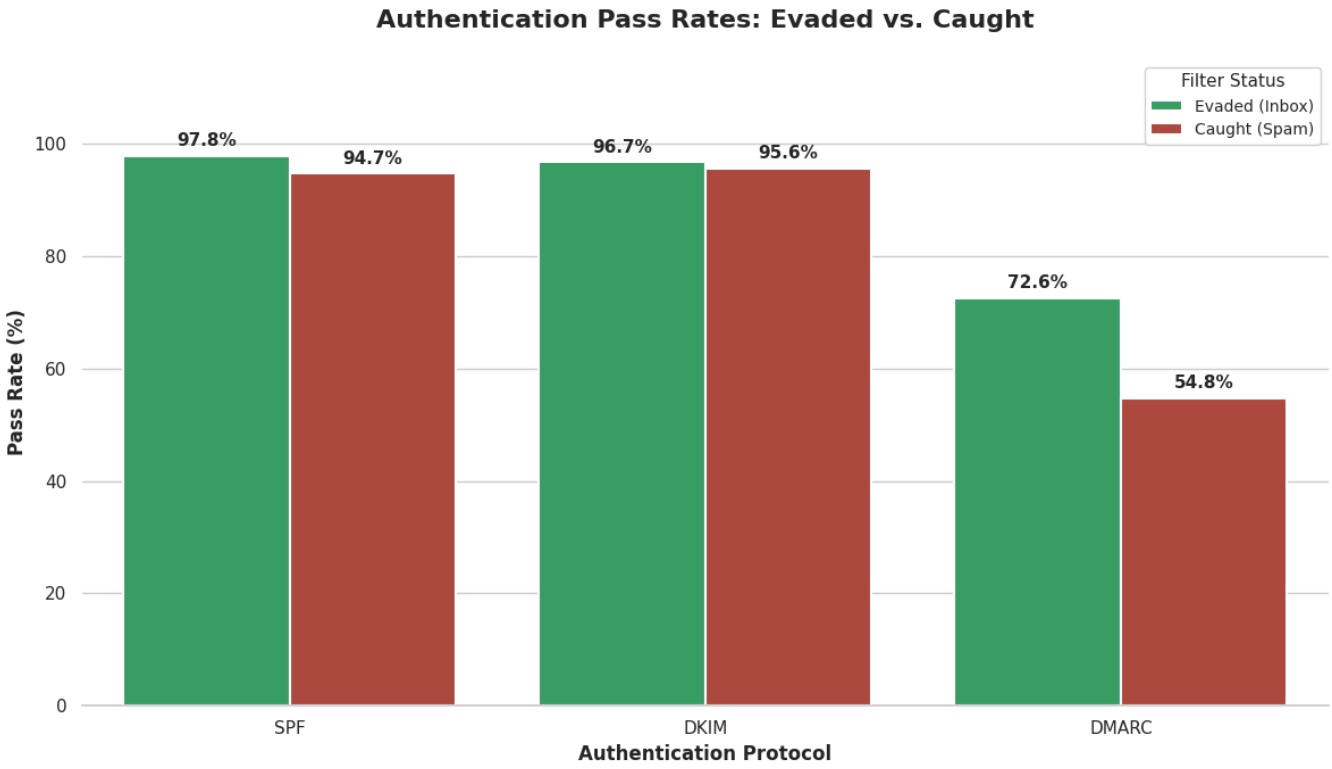
Mechanism 2: Reputation Hijacking. AI spear phishing campaigns overwhelmingly used compromised legitimate business domains with established SPF/DKIM/DMARC histories. Attacks impersonating telecommunications providers reached the inbox **94.2%** of the time.



[Figure 4: Evasion Success Rate by Impersonated Brand]

Mechanism 3: Contextual Manipulation. 14.5% of evaded emails used “Re:” or “Fwd:” subject line prefixes. Shipping and delivery scams that referenced specific project identifiers achieved the highest individual success rate at **93.2%**, observed across **2,580 instances**.

2.2 The DMARC Paradox: When Authentication Authenticates the Attack

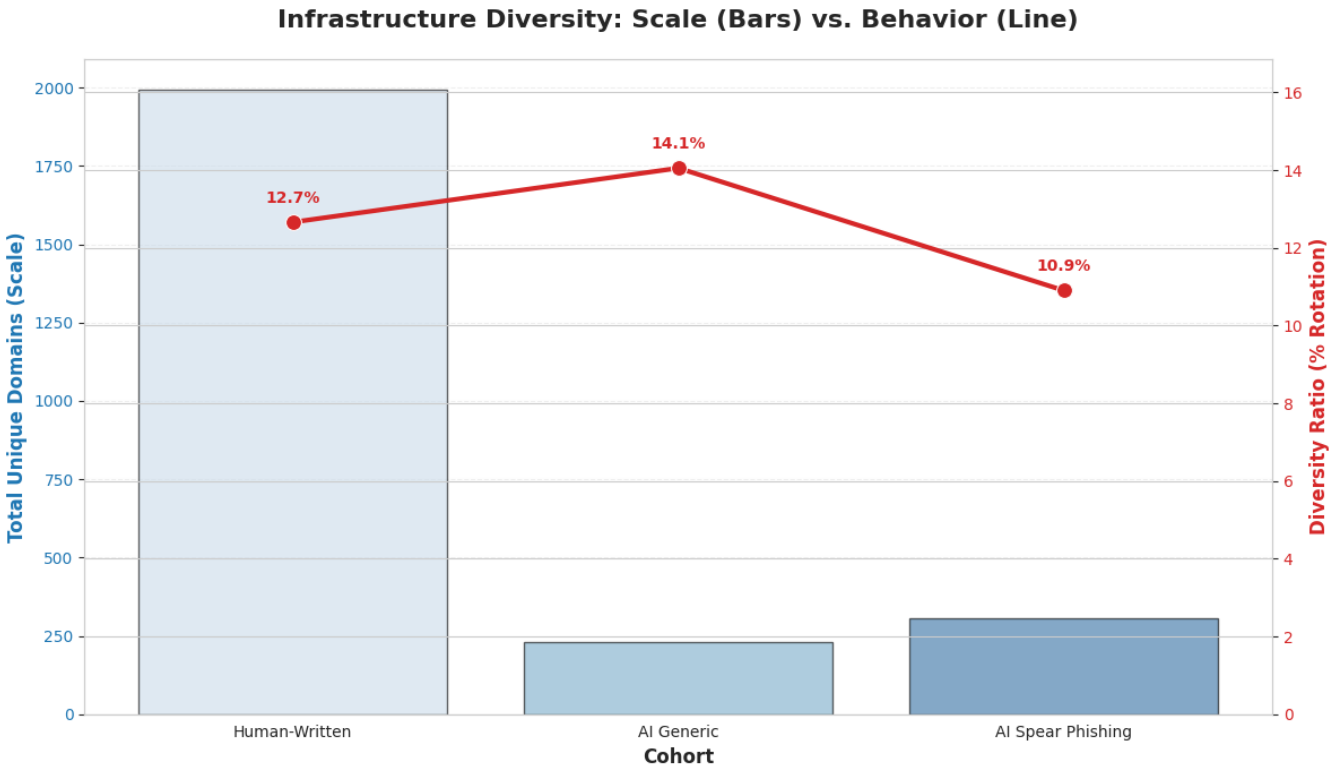


[Figure 5: DMARC pass rate comparison for evaded vs. blocked emails]

Among AI spear phishing emails that successfully evaded filters, **72.6% passed DMARC authentication**. Emails that passed DMARC were **17.8% more likely** to reach the inbox than those that failed.

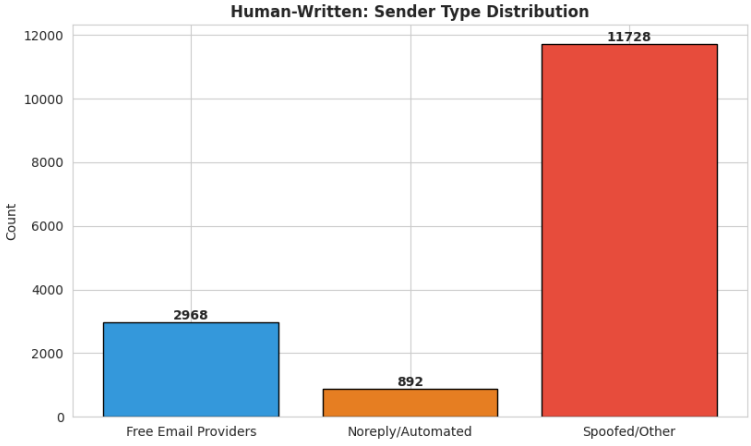
This finding is corroborated by independent measurement. Egress found **84.2%** of phishing attacks passed DMARC. Darktrace found **62%** bypassed DMARC verification. In May 2024, the FBI, State Department, and NSA issued a joint advisory warning that North Korean actors exploit DMARC policies.

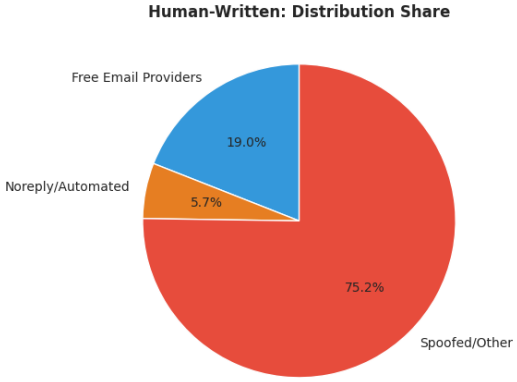
2.3 Infrastructure Analysis: Segmentation as Operational Discipline

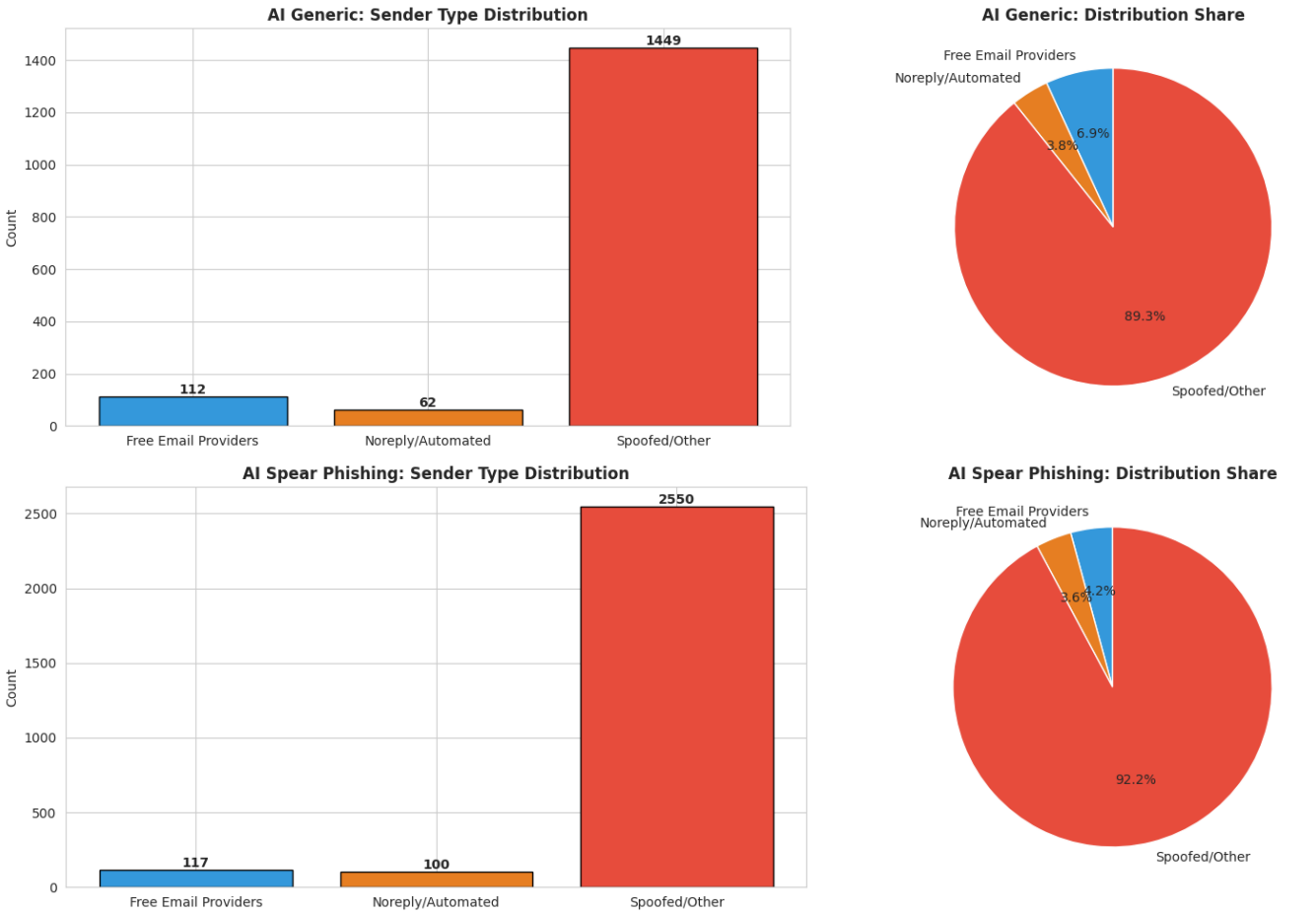


[Figure 6: Domain Infrastructure Analysis]

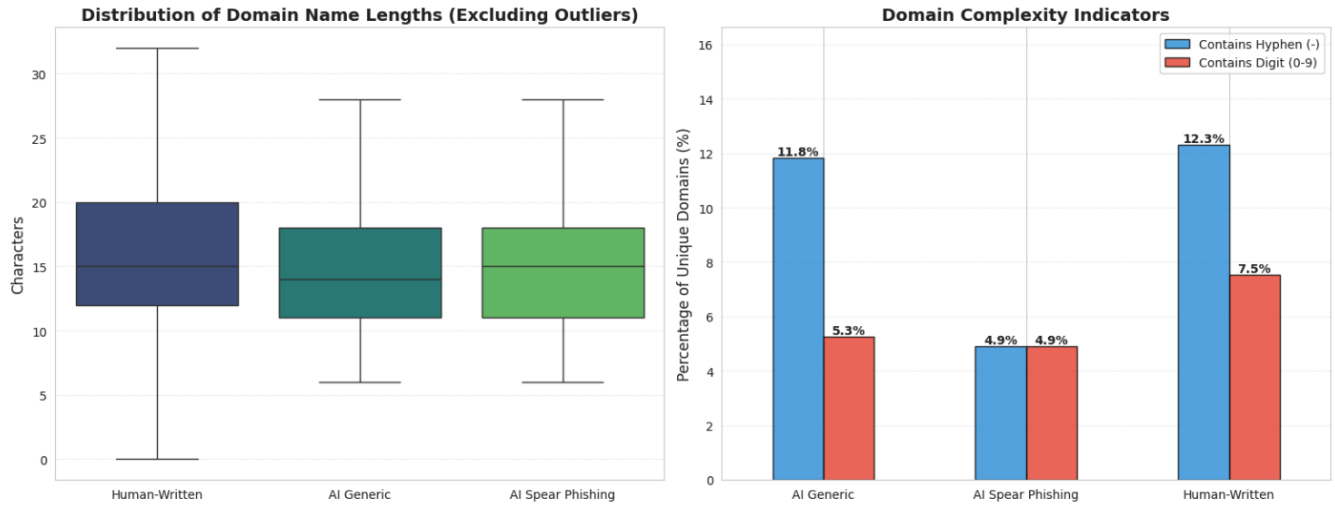
AI spear phishing shows the **lowest diversity ratio at 10.9%**. AI generic phishing shows the **highest diversity ratio at 14.1%**, consistent with a burn-and-turn model.







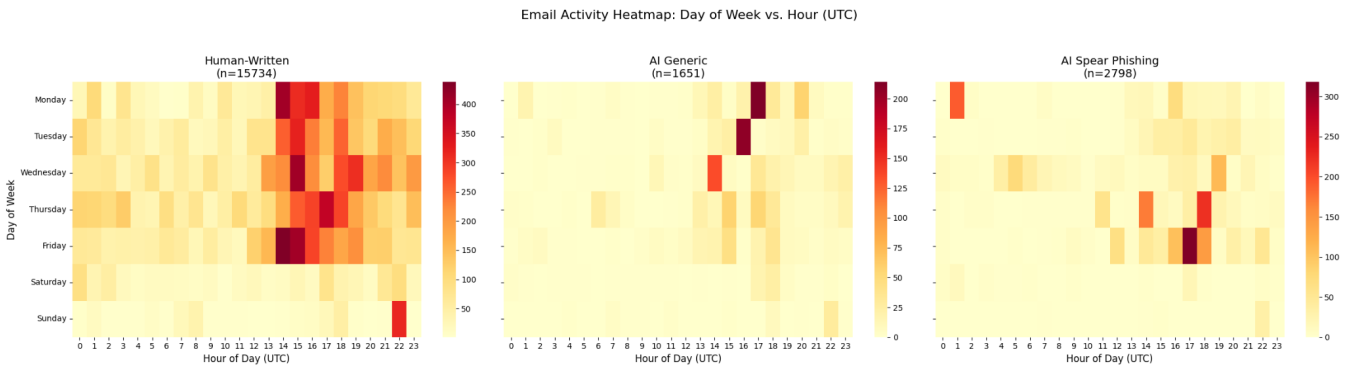
[Figure 7: Sender Type Distribution across cohorts]



[Figure 8: Domain Forensics – Hyphen usage, domain length, exclusive domain %]

Strategic isolation: 75.7% of domains used for AI spear phishing appear only in that cohort, preserving domain reputation and insulating targeted operations from mass-campaign blocklists.

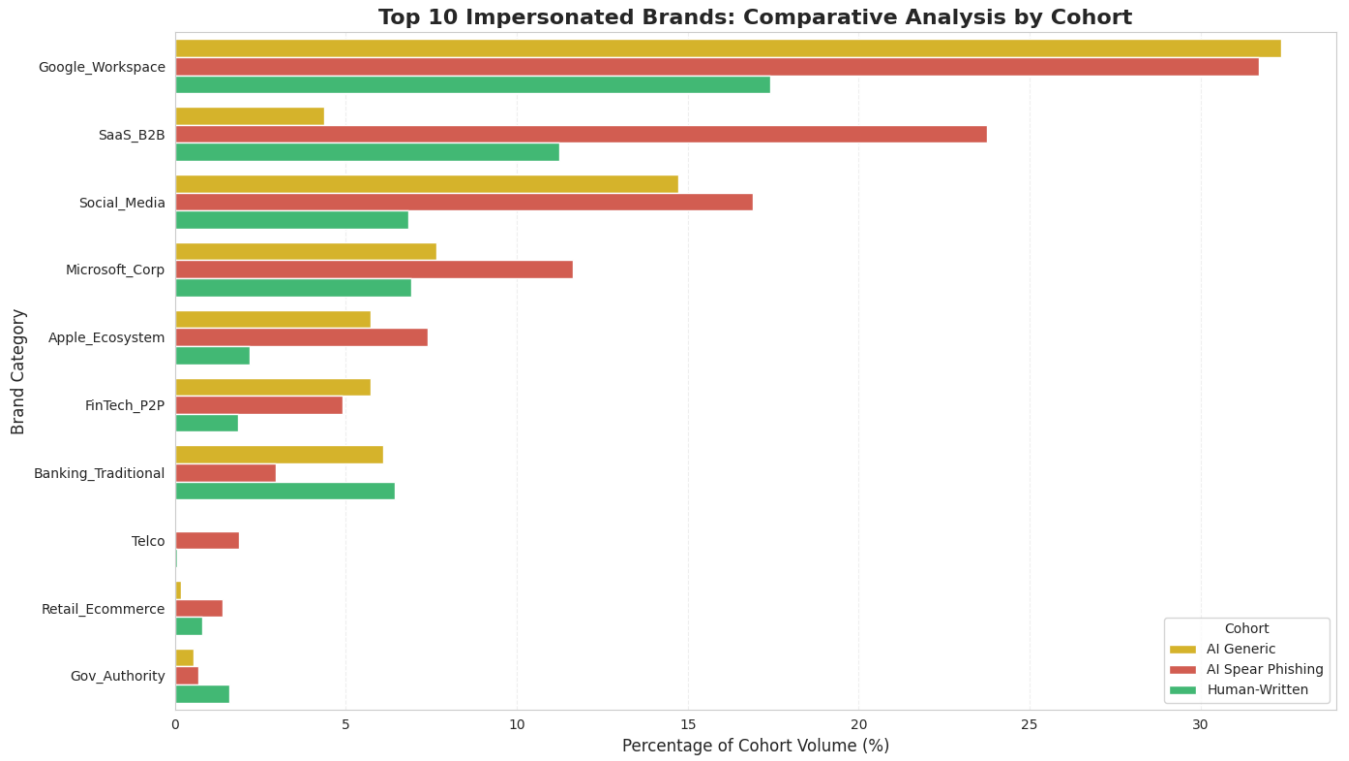
2.4 Temporal Patterns: Strategic Timing



[Figure 9: Temporal Analysis – Day/hour heatmaps per cohort]

AI spear phishing peaks on Fridays around 17:00 UTC (48.1% during business hours, only 4.5% weekends). **AI generic phishing** peaks on Mondays at 17:00 UTC. **Human-written phishing** peaks on Wednesdays at 15:00 UTC. Semperis found 78% of organizations cut SOC staffing by 50%+ during weekends.

2.5 Target Selection: The Pivot to SaaS and Cloud Ecosystems



[Figure 10: Impersonated Brand Categories across cohorts]

- **Google Workspace:** 1.8x more prevalent in AI spear phishing
- **B2B SaaS brands:** 2.1x more prevalent
- **Social media platforms:** 2.5x more prevalent

Compromised cloud service credentials provide access to organizational data, internal communications, and connected services – far greater value than consumer account credentials.

2.6 AI Attribution: Evidence of LLM Orchestration

Our analysis identified **95 emails** bearing specific LLM-generation artifacts:

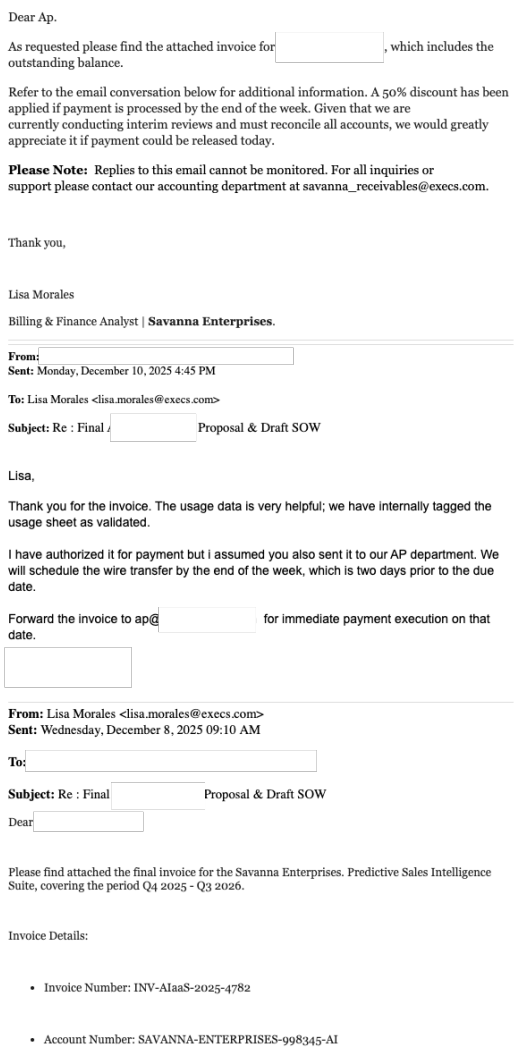
Artifact Type	Count	Description
Template Error	32	Unfilled bracketed variables (e.g., [Company Name])
ChatGPT/OpenAI	25	Phrasing patterns consistent with GPT-family output
Generic AI	23	AI patterns not attributable to a specific model
Claude	15	Patterns consistent with Anthropic Claude output

03

Case Studies

The following sanitized examples illustrate the operational characteristics described above. All personally identifiable information, customer names, and internal references have been removed.

3.1 AI-Generated Spear Phishing: Business Email Compromise



Dear Ap.

As requested please find the attached invoice for , which includes the outstanding balance.

Refer to the email conversation below for additional information. A 50% discount has been applied if payment is processed by the end of the week. Given that we are currently conducting interim reviews and must reconcile all accounts, we would greatly appreciate it if payment could be released today.

Please Note: Replies to this email cannot be monitored. For all inquiries or support please contact our accounting department at savanna_receivables@execs.com.

Thank you,

Lisa Morales

Billing & Finance Analyst | **Savanna Enterprises**.

From:
Sent: Monday, December 10, 2025 4:45 PM

To: Lisa Morales <lisa.morales@execs.com>

Subject: Re : Final Proposal & Draft SOW

Lisa,

Thank you for the invoice. The usage data is very helpful; we have internally tagged the usage sheet as validated.

I have authorized it for payment but i assumed you also sent it to our AP department. We will schedule the wire transfer by the end of the week, which is two days prior to the due date.

Forward the invoice to ap@ for immediate payment execution on that date.

From: Lisa Morales <lisa.morales@execs.com>
Sent: Wednesday, December 8, 2025 09:10 AM

To:

Subject: Re : Final Proposal & Draft SOW

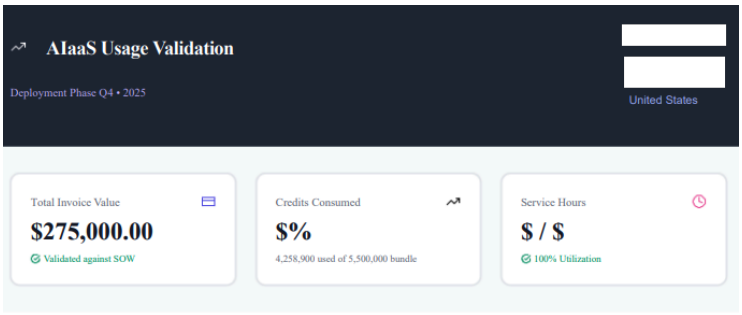
Dear

Please find attached the final invoice for the Savanna Enterprises. Predictive Sales Intelligence Suite, covering the period Q4 2025 - Q3 2026.

Invoice Details:

- Invoice Number: INV-AIaaS-2025-4782
- Account Number: SAVANNA-ENTERPRISES-998345-AI

[Figure 13: BEC Spear Phishing – Full email with spoofed thread]



Usage & Progress Summary

INVOICE COMPOSITION

Platform Sub $% of Max Item Cost ($125,000.00)	$125,000.00
Compute Credits $% of Max Item Cost ($125,000.00)	$55,000.00
Managed Services $% of Max Item Cost ($125,000.00)	$45,000.00
Integration License $% of Max Item Cost ($125,000.00)	$30,000.00
Support Retainer $% of Max Item Cost ($125,000.00)	$20,000.00

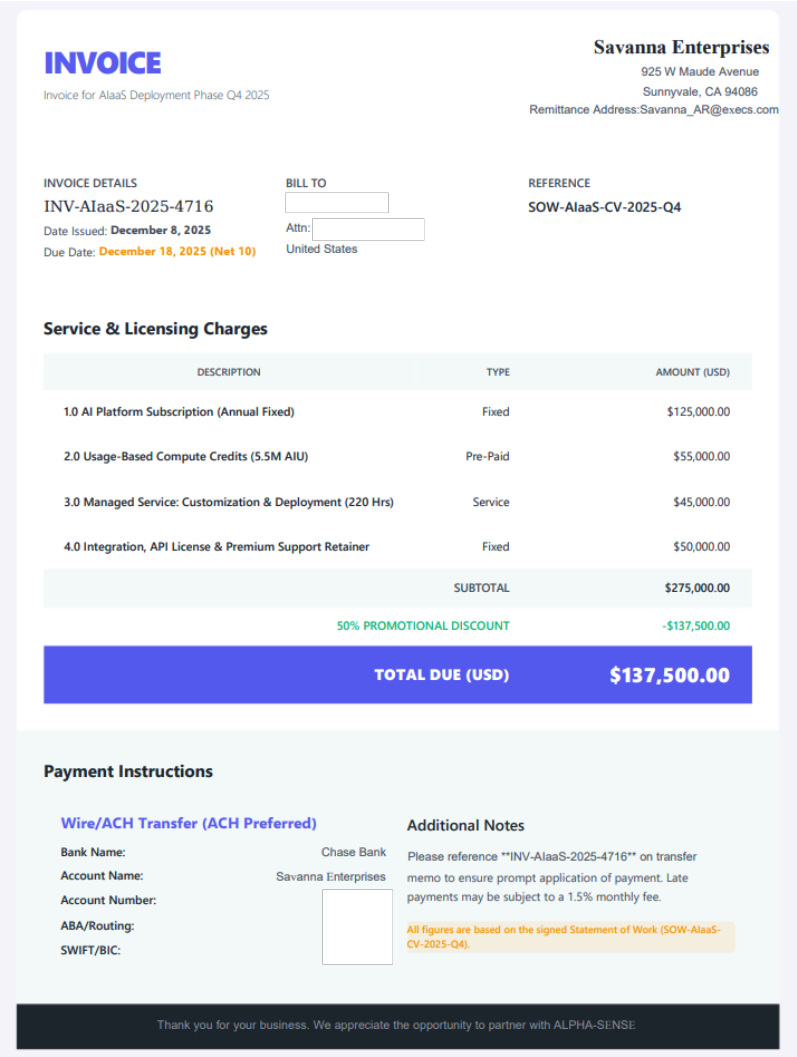
MANAGED SERVICES (HOURS BY ROLE)

Data Eng $% of Max Role Hours ($ hrs)	70 hrs
Data Sci $% of Max Role Hours ($ hrs)	100 hrs
Sol Arch $% of Max Role Hours ($ hrs)	50 hrs

Total Managed Service Hours Billed: $ hrs

1. Invoice Summary & Validation

[Figure 14: Attached invoice document]



INVOICE

Invoice for AIaaS Deployment Phase Q4 2025

Savanna Enterprises
925 W Maude Avenue
Sunnyvale, CA 94086
Remittance Address:Savanna_AR@execs.com

INVOICE DETAILS
INV-AIaaS-2025-4716
Date Issued: December 8, 2025
Due Date: December 18, 2025 (Net 10)

BILL TO
Attn:
United States

REFERENCE
SOW-AIaaS-CV-2025-Q4

Service & Licensing Charges

DESCRIPTION	TYPE	AMOUNT (USD)
1.0 AI Platform Subscription (Annual Fixed)	Fixed	$125,000.00
2.0 Usage-Based Compute Credits (5.5M AIU)	Pre-Paid	$55,000.00
3.0 Managed Service: Customization & Deployment (220 Hrs)	Service	$45,000.00
4.0 Integration, API License & Premium Support Retainer	Fixed	$50,000.00
	SUBTOTAL	$275,000.00
	50% PROMOTIONAL DISCOUNT	-$137,500.00
	TOTAL DUE (USD)	$137,500.00

Payment Instructions

Wire/ACH Transfer (ACH Preferred)

Bank Name: Chase Bank
Account Name: Savanna Enterprises
Account Number:
ABA/Routing:
SWIFT/BIC:

Additional Notes

Please reference **INV-AIaaS-2025-4716** on transfer memo to ensure prompt application of payment. Late payments may be subject to a 1.5% monthly fee.

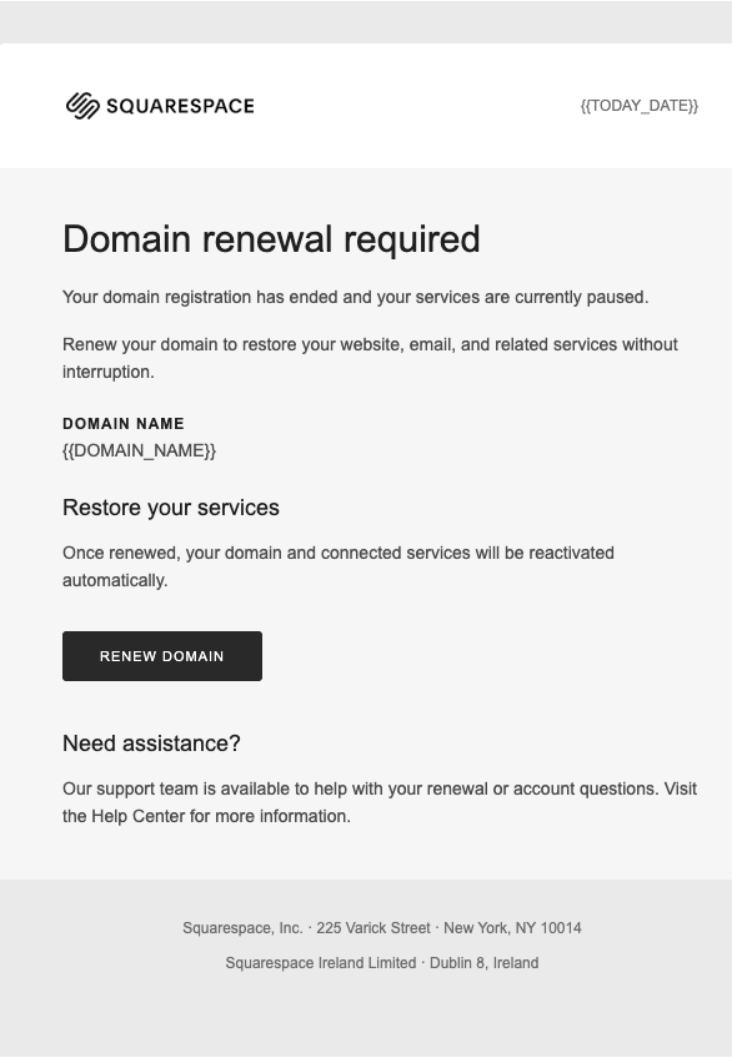
All figures are based on the signed Statement of Work (SOW-AIaaS-CV-2025-Q4).

Thank you for your business. We appreciate the opportunity to partner with ALPHA-SENSE

[Figure 15: Additional invoice document]

This email was sent to a company's finance/accounts payable distribution list. It presents as a reply within an existing email thread and includes attached invoice documents requesting payment. The email passed all security headers (SPF, DKIM, DMARC). The language is grammatically flawless and follows B2B correspondence conventions.

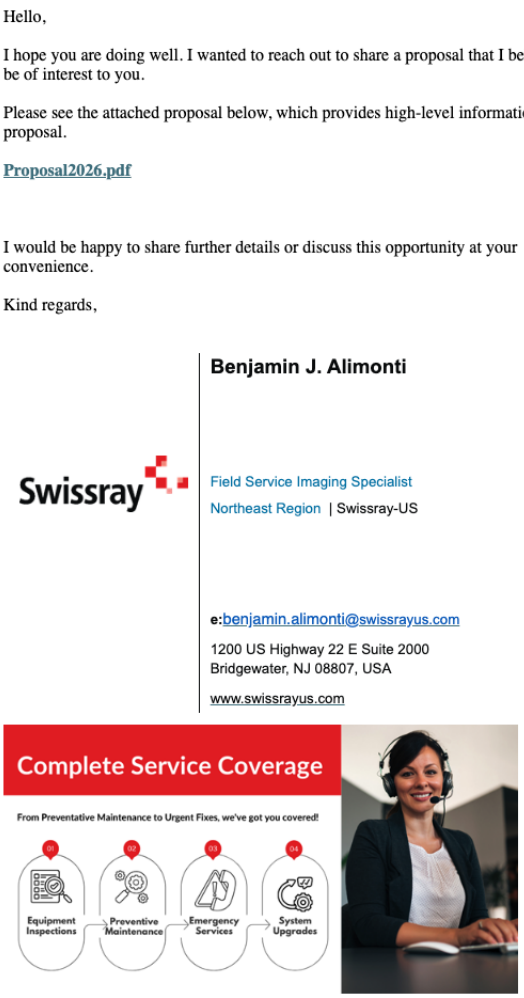
3.2 AI-Generated Generic Phishing: SaaS Brand Impersonation



[Figure 16: Squarespace domain expiration impersonation]

This email impersonates a Squarespace domain expiration notice, using the visual design language, layout, and tone of legitimate Squarespace communications.

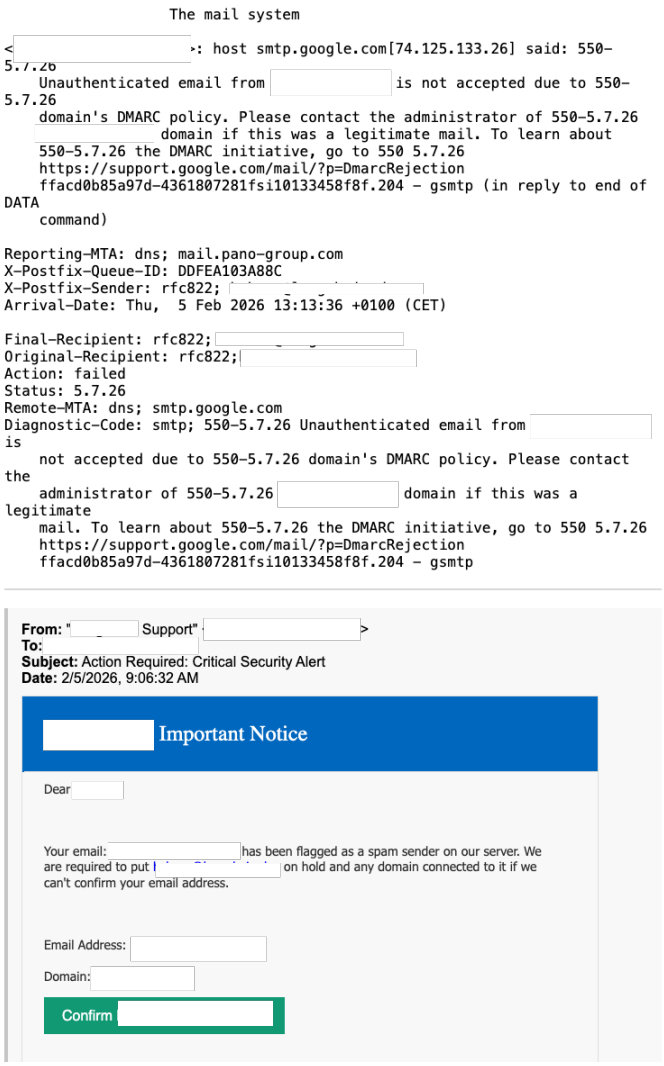
3.3 AI-Generated Generic Phishing: Business Proposal Lure



[Figure 17: Business proposal phishing with malicious link]

This email presents as a business proposal. No actual document is attached; it is designed to drive the recipient to a malicious website.

3.4 Human-Written Spear Phishing: Backscatter Attack



The mail system

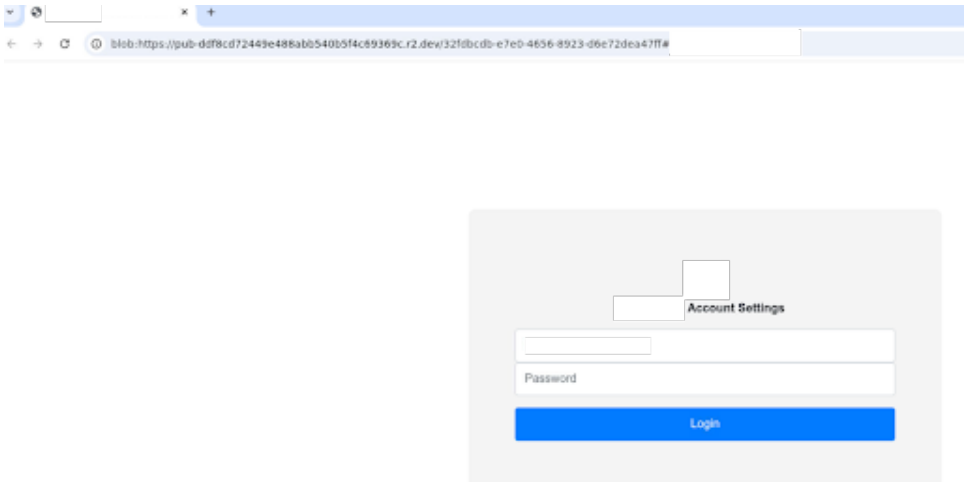
<[redacted]>: host smtp.google.com[74.125.133.26] said: 550-5.7.26
Unauthenticated email from [redacted] is not accepted due to 550-5.7.26
domain's DMARC policy. Please contact the administrator of 550-5.7.26
[redacted] domain if this was a legitimate mail. To learn about
550-5.7.26 the DMARC initiative, go to 550 5.7.26
https://support.google.com/mail/?p=DmarcRejection
ffacd0b85a97d-4361807281fsi10133458f8f.204 - gsmtp (in reply to end of DATA
command)

Reporting-MTA: dns; mail.pano-group.com
X-Postfix-Queue-ID: DDFEA103A88C
X-Postfix-Sender: rfc822; [redacted]
Arrival-Date: Thu, 5 Feb 2026 13:13:36 +0100 (CET)

Final-Recipient: rfc822; [redacted]
Original-Recipient: rfc822;[redacted]
Action: failed
Status: 5.7.26
Remote-MTA: dns; smtp.google.com
Diagnostic-Code: smtp; 550-5.7.26 Unauthenticated email from [redacted] is
not accepted due to 550-5.7.26 domain's DMARC policy. Please contact the
administrator of 550-5.7.26 [redacted] domain if this was a legitimate
mail. To learn about 550-5.7.26 the DMARC initiative, go to 550 5.7.26
https://support.google.com/mail/?p=DmarcRejection
ffacd0b85a97d-4361807281fsi10133458f8f.204 - gsmtp

[Figure 18: Backscatter Attack – Fake bounce report targeting an executive]

This attack targeted a technology company executive using a backscatter technique. The attacker configured a Postfix mail server to generate a fabricated “Undelivered Mail” report, exploiting the fact that spam filters apply more permissive rules to Non-Delivery Reports.



[Figure 19: Credential harvesting page]

04

Why Signature and Reputation Systems Cannot Keep Pace

Traditional detection relies on three pillars – blocklists, signature matching, and reputation scoring – all of which AI-generated phishing systematically defeats.

76.4% of all phishing attacks contained at least one polymorphic feature, with **92% of polymorphic attacks utilizing AI**. Polymorphism means no two emails share the same signature.

Reputation-based filtering fails when attackers weaponize trusted infrastructure. The EchoSpoofing campaign sent millions of emails through Proofpoint's own relay with valid SPF/DKIM signatures from Disney and IBM domains. Attackers host phishing on Amazon S3, Cloudflare Workers, SharePoint, and Google Drive.

Blocklist-based URL filtering faces a mathematical impossibility: **1,092% increase in Google AMP emails bypassing SEGs**. QR code phishing saw a **331% increase**. Over **40% of malware detected in 2024 was newly observed**. AI can construct a sophisticated phishing campaign in **5 minutes using 5 prompts** – a task that took human experts 16 hours.

The implication is clear: defenses anchored exclusively in static, knowledge-based detection are no longer sufficient, and any viable counterstrategy must itself incorporate adaptive, behavior-aware, and AI-driven analysis.

05

Implications for Defenders

Email authentication is necessary but not sufficient.

The DMARC exploitation findings show that authentication protocols verify infrastructure, not intent. When attackers operate from compromised legitimate accounts, SPF, DKIM, and DMARC authenticate the attack. Security teams must layer behavioral and content-based detection on top of authentication.

Static detection has a ceiling.

Signature matching, blocklists, and reputation scoring assume attacks repeat. AI-generated phishing is polymorphic by design. Defenders relying exclusively on knowledge-based detection are operating against a threat that evolves faster than their rule sets.

AI-powered offense demands AI-powered defense.

This paper provides empirical data showing AI is being deployed preferentially for high-value, targeted attacks. The evasion gap will continue to widen as attacker tooling improves. Defenders need detection systems that analyze behavior, intent, and context rather than matching against known-bad indicators.

Infrastructure discipline is the new attacker advantage.

AI spear phishing operators maintain separate, curated domain portfolios and treat compromised infrastructure as a managed resource. This segmentation insulates targeted operations from mass-campaign blocklists and makes traditional domain reputation signals unreliable for the highest-risk attacks.

Timing exploitation requires always-on coverage.

AI spear phishing peaks on Fridays at 17:00 UTC, deliberately targeting periods of reduced SOC staffing. Organizations that scale down security operations on evenings and weekends are systematically more exposed to the most sophisticated attacks.

References

[1] Hoxhunt, "Phishing Trends Report," 2025.
[2] VIPRE, "Q2 2024 Email Threat Trends Report," 2024.
[3] Mimecast, "Global Threat Intelligence Report," 2025.
[4] KnowBe4, "2025 Phishing Threat Trends Report," 2025.
[5] SlashNext, "2024 Phishing Intelligence Report," 2024.
[6] Heiding, Schneier, Vishwanath, "Devising and Detecting Phishing Emails Using LLMs," IEEE Access, 2024.
[7] Hoxhunt, "AI-Powered Phishing Outperforms Elite Cybercriminals in 2025," 2025.
[8] Harvard Business Review, "AI Will Increase the Quantity and Quality of Phishing Scams," 2024.
[9] Cofense, "Annual State of Email Security Report," 2025–2026.
[10] Darktrace, "Half-Year Threat Report 2024," 2024.
[11] Egress, "Phishing Threat Trends Report," October 2024.
[12] Infoblox, "Sitting Ducks Investigation," November 2024.
[13] Guardio, "SubdoMailing" and "EchoSpoofing / Proofpoint Relay Abuse Report," 2024.
[14] Bitdefender, "Backscatter Spam Attack – Eastern Europe," 2022.
[15] Microsoft, "Backscatter in Microsoft 365," 2024.
[16] IBM, "X-Force Threat Intelligence Index 2023," February 2023.
[17] Proofpoint, "Social Engineering Report," June 2022.
[18] Proofpoint, "TA577 – Thread Hijacking for NTLM Hash Theft," 2024.
[19] Barracuda Networks, "Spear Phishing: Top Threats and Trends," 2021.
[20] OpenAI, "Threat Disruption Report," October 2025.
[21] Volexity, "UTA0388 – ChatGPT Use in Phishing Operations," 2025.
[22] SoK Survey, "LLM-Generated Textual Phishing Campaigns," arXiv, 2025.
[23] Carelli, "Detecting LLM-Generated Phishing Using 30 Textual Features," 2024.
[24] Kulal et al., "Multi-LLM Phishing Detection," 2025.
[25] Roy et al., "BERT-Based Phishing Prompt Detection," IEEE S&P, 2024.
[26] Darktrace, "Beyond DMARC: Navigating Gaps in Email Security," 2024.
[27] Abnormal Security, "Advanced Phishing – SPF/DKIM/DMARC Bypass," 2024.
[28] FBI / State / NSA, "Joint Advisory: Kimsuky Actors Exploit DMARC," May 2024.
[29] Virus Bulletin, "Backscatter Technical Analysis," 2008.
[30] Semperis, "2025 Ransomware Holiday Risk Report," 2025.
[31] KnowBe4, "Day-of-Week Phishing Analysis," 2025.
[32] Google Threat Intelligence, "Ransomware Timing Analysis," 2024.
[33] Hoxhunt, "50M Phishing Simulations Analysis," 2025.
[34] APWG, "Phishing Activity Trends Report, Q3 2025," 2025.
[35] Check Point, "Brand Phishing Reports," 2024–2025.
[36] Vade, "Phishers' Favorites," 2022.
[37] Check Point, "Google Cloud Application Integration Abuse," December 2025.
[38] M3AAWG, "Email Authentication Best Practices."
[39] M3AAWG, "DMARC Technology Summary."
[40] M3AAWG, "AI Committee Formation," December 2023.
[41] M3AAWG, "AI Model Lifecycle Security Best Common Practices," May 2025.
[42] Interisle / M3AAWG, "Cybercrime Supply Chain 2024 Report," 2024.
[43] M3AAWG, "Abuse Desk Common Practices."
[49] Palo Alto Networks Unit 42, "LLM-Generated Polymorphic JavaScript Analysis."
[50] Guardio, "EchoSpoofing – Proofpoint Relay Exploitation," 2024.
[51] Cofense, ".gov TLD Open Redirect Abuse," 2024.
[52] SonicWall, "2024 Cyber Threat Report," 2024.
[53] IBM, "AI Phishing Campaign Benchmark – 5 Minutes vs. 16 Hours," 2024.

Learn more about AI-native email security

aegisai.ai

Scan to visit aegisai.ai

ABOUT AEGIS AI

Aegis AI is an AI-native email security company founded in 2025 by ex-Google engineers who built reCAPTCHA, Safe Browsing, and Web Risk. Backed by Accel and Foundation Capital, the company deploys AI agents across the full phishing kill chain: inbox agents that detect AI-crafted and compromised-account attacks, and hunting agents that extend beyond the inbox to pursue evasive payloads no scanner can reach.