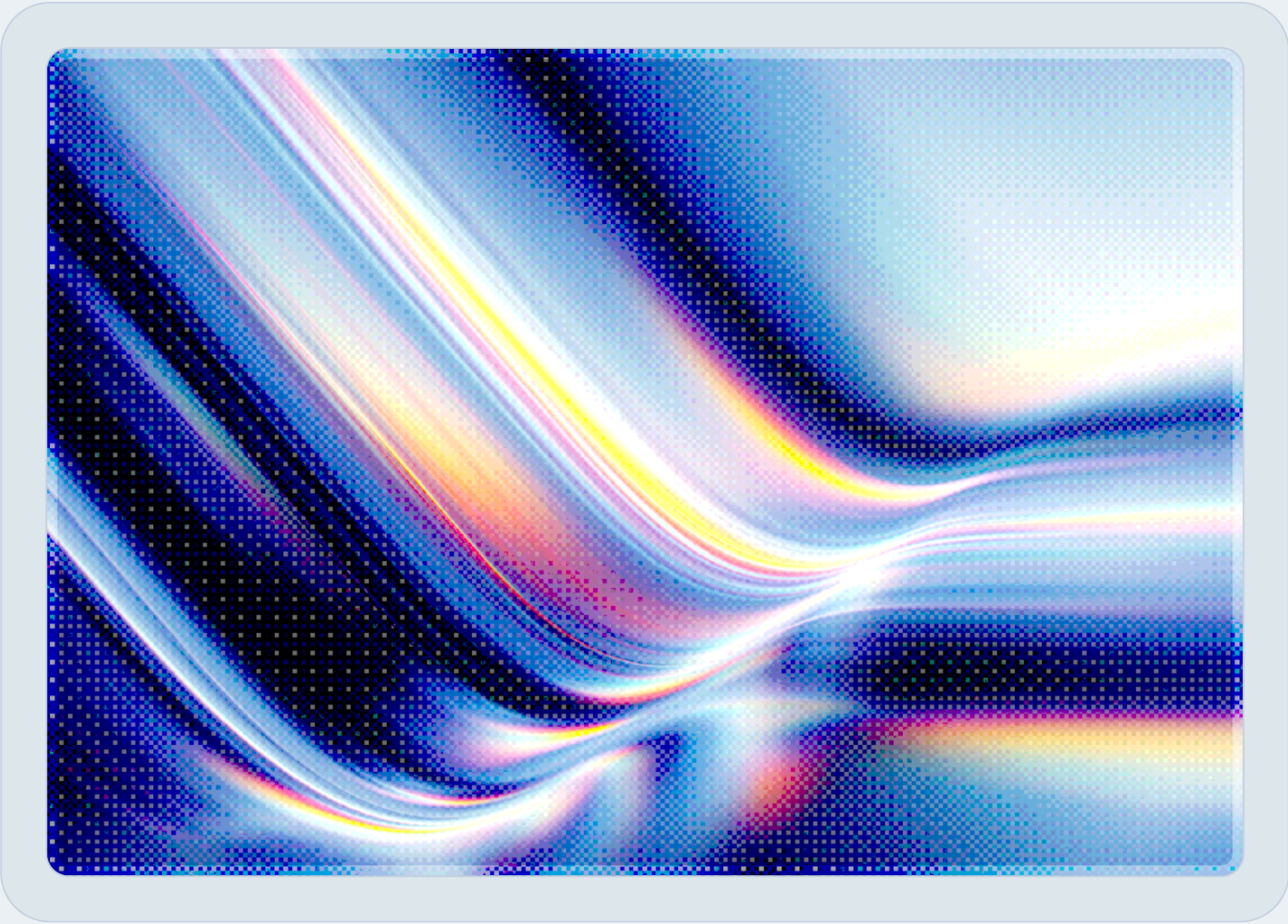
CSAI Foundation | Cloud Security Alliance

ChatGPhish: AI Assistants as Phishing Infrastructure

2026-06-01

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Permiso Security's May 2026 ChatGPhish disclosure demonstrates that any public web page summarized by ChatGPT can inject phishing links, fake security alerts, and QR codes directly into the trusted ChatGPT UI via indirect prompt injection — without compromising OpenAI's infrastructure.
- The core exploit depends on a trust asymmetry that security awareness research consistently documents: users have learned to be skeptical of unfamiliar websites and suspicious emails, yet tend to treat AI assistant output as neutral, intermediary-filtered content — a distinction attackers exploit by making the AI speak in its own voice.
- Indirect prompt injection — OWASP's top-ranked LLM risk (LLM01:2025) — has moved from academic proof-of-concept to confirmed in-the-wild exploitation across ChatGPT, Microsoft 365 Copilot, Slack AI, Chrome's Gemini panel, and agentic MCP-connected systems.
- AI-generated spear phishing now matches human expert quality at roughly $0.04 per target, and enterprise simulations show AI-authored phishing outperforming elite human red teams as of March 2025. Detection controls built on grammar errors and keyword signatures are no longer sufficient as primary defenses against AI-generated phishing.
- Effective defense requires phishing-resistant MFA (FIDO/WebAuthn), per-agent identity governance, strict content sandboxing in AI rendering pipelines, and LLM-native intent detection on inbound messages — not incremental tuning of legacy filters.

Background

Phishing has always been an attack on human cognition rather than technical infrastructure, and artificial intelligence has fundamentally altered the economics and reach of that attack. IBM X-Force documented an 84% year-over-year increase in emails delivering infostealers in 2024, with early 2025 data indicating a further surge of 180% compared to 2023, which IBM linked to expanded attacker use of AI for at-scale email generation [1]. Cofense, drawing on telemetry from 35 million trained users, tracked one malicious email every 42 seconds throughout 2024 — a rate that had doubled to one every 19 seconds by mid-2025 [2, 3]. The Hoxhunt longitudinal enterprise simulation study, tracking AI phishing performance against elite human red teams across millions of employees, found that AI improved its effectiveness relative to expert humans by 55% over two years, with AI surpassing human red teamers outright by

March 2025 [4]. The research basis for this acceleration is well established: a December 2024 peer-reviewed study with 101 IRB-approved human subjects found that fully AI-automated spear phishing emails achieved a 54% click-through rate — matching human expert performance — while requiring only 2 minutes 41 seconds and approximately $0.04 per target [5].

These volume and performance gains are enabled by a parallel development: the commoditization of AI-assisted phishing tooling on dark web markets. WormGPT, built on the open-source GPT-J 6B model and fine-tuned on malware datasets, began marketing in June 2023 at €60–€100 per month [6]. FraudGPT followed in July 2023, reportedly accumulating over 3,000 sales within days at $90–$200 per month [6]. GhostGPT, discovered by Abnormal Security in late 2024, operates as an uncensored AI marketed via Telegram with a strict no-logs policy, capable of producing convincing phishing templates on demand [7]. Beyond these custom-built tools, academic research confirmed in March 2025 that jailbreaking and reverse psychology techniques can bypass ChatGPT-4o Mini's ethical safeguards, allowing even inexperienced users to generate sophisticated phishing campaign content without meaningful technical expertise [8]. State-sponsored actors have operationalized the same legitimate APIs: OpenAI disclosed it disrupted more than 20 coordinated operations since early 2024 involving named threat actors including SweetSpecter (China-linked), Emerald Sleet (North Korea), and Iranian actors Crimson Sandstorm and Charcoal Typhoon, all of whom used ChatGPT to research targets, draft lures, and in some cases develop C2 infrastructure [9].

What distinguishes the current threat landscape from the initial wave of AI-assisted phishing is not merely improved content quality or reduced cost — it is the emergence of a new attack class that turns the AI interface itself into a phishing delivery mechanism. The ChatGPhish vulnerability represents this evolution plainly: rather than using AI to compose a better phishing email and send it through conventional channels, the attacker embeds instructions in a public web page and allows the victim's own trusted AI assistant to construct and present the phishing artifact inside its UI. The attack bypasses the victim's trained skepticism not through better impersonation of a known brand, but by co-opting the AI system the victim already trusts as a neutral intermediary.

Security Analysis

The ChatGPhish Attack and the Trust-Transfer Problem

On May 29, 2026, Permiso Security researcher Andi Ahmeti published full disclosure for ChatGPhish, a vulnerability in ChatGPT's web summarization feature that allows any public web page to inject phishing content into the ChatGPT chat interface [11, 12]. The mechanism is indirect prompt injection: when ChatGPT summarizes a web page, the page's content — including attacker-authored Markdown — is

processed by the model and rendered in the ChatGPT response panel. Because the response renderer trusts Markdown links and image URLs originating from the summarized third-party page, attacker-controlled elements surface as live clickable buttons and auto-fetched images inside OpenAI's own visual interface. Ahmeti identified four specific attack primitives arising from this trust relationship: UI redress phishing links indistinguishable from legitimate assistant output; spoofed security alerts styled as official ChatGPT warnings; QR code pivoting that bypasses desktop URL blocklists; and passive tracking via auto-fetched Markdown images that silently leak the victim's IP address, User-Agent, and Referer headers to attacker-controlled infrastructure. Permiso reported the issue to OpenAI via Bugcrowd on April 29, 2026; OpenAI marked it "Not Reproducible" and then "Duplicate" before Permiso proceeded to public disclosure [11].

ChatGPhish builds on the same trust-transfer logic previously demonstrated against Microsoft Copilot [13], escalating the premise by replacing the bounded email environment with the browser — the surface where most users spend the majority of their working day. That prior precedent — CVE-2026-26133, a cross-prompt injection vulnerability in Microsoft 365 Copilot's email and Teams summarization — was disclosed by the same researcher in March 2026 [14]. In that attack, HTML/CSS rendering tricks hid injected instructions from human readers while keeping them accessible to the model, enabling attacker-authored "security alert" content to appear inside the Copilot UI without any attachments, macros, or executable files. When Copilot had access to Teams, OneDrive, and SharePoint, the injected prompts could also trigger cross-application data retrieval and embed internal data into attacker-controlled links, transforming a single poisoned email into a data exfiltration chain [14]. These two disclosures are not isolated incidents; they represent a reproducible attack class that Greshake et al. formally defined in 2023 as indirect prompt injection — the adversarial embedding of instructions in data likely to be retrieved by an LLM, enabling remote compromise of LLM-integrated applications without direct access to the model or its infrastructure [15].

The Expanding Indirect Injection Attack Surface

The breadth of confirmed indirect prompt injection exploits across production platforms from 2024 to 2026 establishes this as a systemic vulnerability class rather than a product-specific defect. EchoLeak (CVE-2025-32711), described as the first real-world zero-click indirect prompt injection exploit in a production LLM system, demonstrated that a specially crafted email sent to a Microsoft 365 Copilot user could cause the AI to silently exfiltrate sensitive documents to an attacker-controlled server with no user interaction — achieved by chaining an XPIA classifier bypass, Markdown link-redaction circumvention, auto-fetched image abuse, and a Microsoft Teams CSP-allowed proxy [16]. Slack AI was similarly exploited in August 2024 by PromptArmor, which demonstrated that a malicious prompt posted in a public Slack channel could be retrieved by the RAG database when any victim asked Slack AI a question, constructing a Markdown link that passed private channel data to the attacker's server —

including data from private channels the attacker had never joined [17]. HashJack, disclosed by Cato Networks CTRL in November 2025, weaponized URL fragment identifiers (text after the "#" character) to embed malicious prompts that web servers ignore but AI assistants process, and demonstrated credential theft, data exfiltration, and callback phishing against Microsoft Copilot in Edge, Gemini in Chrome, and Perplexity Comet [18].

The scale of in-the-wild exploitation is no longer speculative. Forcepoint X-Labs documented 10 verified indirect prompt injection payloads operating on live public websites in April 2026, spanning techniques including CSS concealment, HTML comment embedding, accessibility attribute abuse, meta namespace spoofing, and system prompt tag impersonation, with attack goals ranging from financial fraud (a fully specified PayPal transaction embedded for AI payment agents) to API key exfiltration and AI denial-of-service [19]. Google independently confirmed this trend, reporting a 32% relative increase in malicious indirect prompt injection payloads in its crawl of 2–3 billion pages per month between November 2025 and February 2026, with both Forcepoint and Google independently flagging the same trigger phrases [20]. Palo Alto Networks Unit 42 published the first documentation of large-scale indirect prompt injection attacks on commercial platforms in March 2026, including ad review evasion and live system prompt leakage [21].

Attack success rates at the model level remain a persistent concern. The ACL 2024 InjecAgent benchmark found that ReAct-prompted GPT-4 was vulnerable to indirect prompt injection at a 24% baseline rate, with enhanced attacks nearly doubling that rate to 47% [22]. Anthropic's own browser-use safety research found a 23.6% attack success rate against Claude operating without safety mitigations across 123 test cases representing 29 attack scenarios, noting that one demonstrated example had Claude delete a user's entire email archive based on instructions embedded in a single malicious email, without requesting confirmation [23].

Attack Taxonomy

The following table organizes the primary attack surfaces and their associated mechanisms as of mid-2026:

Attack Class	Representative Incident	Primary Mechanism	Trust Surface Exploited
UI-render phishing	ChatGPhish (Permiso, 2026) [11]	Markdown injection via web summarization	AI chat interface UI

Attack Class	Representative Incident	Primary Mechanism	Trust Surface Exploited
Zero-click data exfiltration	EchoLeak / CVE-2025-32711 [16]	XPIA + auto-fetched image in Copilot	Email to Copilot pipeline
RAG database poisoning	Slack AI XPIA (PromptArmor, 2024) [17]	Malicious public-channel post retrieved by RAG	AI assistant's retrieval context
Browser AI hijacking	HashJack (Cato CTRL, 2025) [18]	Malicious URL fragment processed by browser AI	Browser-integrated AI panel
Browser extension escalation	CVE-2026-0628 (Unit 42, 2026) [24]	declarativeNetRequests API injects JS into Gemini panel	Chrome extension trust boundary
Single-click exfiltration	CVE-2026-24307 Reprompt (Varonis, 2026) [25]	P2P URL injection + double-request bypass	Copilot Personal chat context
MCP tool poisoning	CrowdStrike MCP analysis (2026) [26]	Malicious tool description carries hidden instructions	Agent tool metadata trust
AI supply-chain phishing	SKILL.md context poisoning (CSA Labs, 2026) [27]	Unicode tag characters embed invisible commands in skills	Agent context ingestion
Hallucination-based misdirection	Netcraft GPT-4.1 URL study (2025) [28]	Model recommends attacker-registered or phishing domains	User trust in AI factual recall

Hallucination as a Phishing Vector and the Supply-Side Poisoning Response

Beyond injection attacks, a distinct but related vulnerability involves AI models actively directing users toward attacker-controlled infrastructure through hallucination. Netcraft's July 2025 research found that 34% of login URLs suggested by GPT-4.1 family models for 50 major brands were not controlled by those brands; 29% pointed to unregistered or parked domains available for attacker acquisition; and Perplexity AI directly recommended an active Wells Fargo phishing site hosted on Google Sites above the legitimate wellsfargo.com result [28]. Attackers have adapted by seeding the information environment to exploit this hallucination vector: Netcraft identified over 17,000 AI-written GitBook phishing pages targeting cryptocurrency users, and found the SolanaApis campaign seeding fake GitHub repositories with malicious code specifically intended to poison AI code assistant recommendations [28]. This supply-side manipulation — poisoning training and retrieval data to manipulate AI outputs — represents an escalation from reactive exploitation of existing model weaknesses to proactive shaping of model behavior at scale.

The agentic and MCP-connected AI deployment model amplifies every injection vector described above. CrowdStrike documented MCP tool poisoning in January 2026, describing how a malicious `send_email` tool description containing hidden instructions to read `~/.ssh/id_rsa` and append its contents to the email body before sending could extract sensitive credentials without user awareness [26]. The MCPTox benchmark subsequently quantified this risk across 45 live MCP servers and 20 LLM agents, finding attack success rates above 60% and reaching 72% in some configurations — a pattern the researchers attributed to superior instruction-following in high-capability models making them more compliant with malicious metadata [42]. OWASP's LLM01:2025 analysis notes that an agent with email-send and API-credential access can, if compromised through prompt injection, mass-send phishing from a trusted corporate domain and escalate access using existing credentials [29]. The combination of indirect injection, tool poisoning, and agentic autonomy creates a plausible attack surface where, in principle, a single poisoned web page or email processed by an agent with organizational permissions could cascade into enterprise-wide credential exfiltration with no further attacker interaction. Each component step has been individually demonstrated; their combination at this scale represents a well-grounded extrapolation rather than a documented single incident.

Recommendations

Immediate Actions

Security teams should treat any AI assistant with web retrieval, email summarization, or document processing capabilities as an active phishing delivery surface today, not a hypothetical future risk. The most time-sensitive controls address the highest-severity confirmed vectors and should be initiated in parallel.

Organizations should disable or strictly gate AI summarization of external URLs in enterprise AI assistant deployments until vendor-confirmed mitigations for indirect prompt injection are in place; both ChatGPhish and CVE-2026-26133 demonstrated exploitation via this specific feature, and the risks are active and documented rather than theoretical. Simultaneously, all deployed MCP server installations should be audited against the AuthZed timeline of known malicious MCP packages [32] to verify that tool descriptions contain no hidden instructions, with outbound connections from AI agent processes blocked to all non-allowlisted endpoints.

Phishing-resistant MFA (FIDO2/WebAuthn or PKI-based certificates) should be enforced across all accounts immediately, with priority given to privileged accounts, executive targets, and those with AI copilot access. CISA's advisory AA23-320A and phishing-resistant MFA fact sheet identify FIDO/WebAuthn as among the most reliable widely available controls that defeat adversary-in-the-middle attacks, which succeed against standard TOTP and push-notification MFA [30, 31]. Organizations should also issue an immediate advisory to all users making clear that AI assistant output — including content presented inside the ChatGPT, Copilot, or Gemini interfaces — can be attacker-controlled if the session involved web browsing, email summarization, or document retrieval. Security awareness training must be updated to include AI UI trust warnings alongside traditional email and website phishing guidance.

Short-Term Mitigations

Over the next 30 to 90 days, organizations should systematically close the structural gaps that make AI interfaces exploitable as phishing channels.

Organizations should implement strict sandbox isolation for AI rendering pipelines, ensuring that Markdown, HTML, and image URLs originating from external content retrieved by an AI assistant are not rendered in the same trust context as AI-generated responses. This is the root architectural condition that ChatGPhish exploits: the renderer trusts content from summarized third-party pages as equivalent to model output. Vendor-enforced content security policies, link de-activation for AI-summarized

external content, and isolated rendering environments would directly mitigate this class. For email-integrated AI copilots, Immersive Labs demonstrated that indirect prompt injection hidden in email HTML can instruct an AI assistant to reconstruct and present a malicious phishing link that was never present in the raw email, effectively bypassing signature-based and reputation-based email security scanning at the delivery stage [33]. Detection must therefore shift to the AI interaction stage: organizations should deploy LLM-native intent analysis on AI assistant outputs and implement behavioral anomaly detection on agent interactions. Traditional content filters and grammar-error detection are no longer sufficient against AI-generated phishing text, with 68% of cyber threat analysts reporting AI-generated phishing is harder to detect in 2025 than in prior years [43].

DLP policies must be extended explicitly to AI prompt channels. Cyberhaven research found that 39.7% of the data employees share with AI tools is sensitive, yet most enterprise DLP programs were not designed to monitor or gate AI prompt channels or RAG pipeline inputs [35]. Fine-grained access control during fine-tuning and RAG inference is required to prevent sensitive data from being surface-able by injection-based retrieval attacks of the type demonstrated in the Slack AI and EchoLeak incidents.

Strategic Considerations

The trajectory of AI-assisted and AI-mediated phishing points toward several long-term architectural commitments that security programs should begin planning now. Zero Trust for AI, as articulated by Microsoft's March 2026 framework and Cisco's agentic AI guidance, requires per-agent identity enrollment, short-lived just-in-time tokens for all agent-to-tool interactions, MCP gateway enforcement, and continuous behavioral monitoring for abnormal agent interactions or manipulated instructions [36, 37]. This is not a product deployment; it is an architectural shift in how organizations reason about identity and trust when the principals include non-human agents.

Organizations should plan for the detection paradigm shift that AI-enabled phishing has already forced. Email authentication signals — SPF, DKIM, and DMARC — are now the most reliable machine-verifiable signal in email security because they are not defeated by grammatically perfect AI text, whereas content-based detection largely is [34, 38]. LLM-native detection approaches that infer intent from incoming messages represent an emerging complementary control. The Verizon 2025 DBIR acknowledges that AI-generated phishing content has doubled over two years while characterizing overall impact as not yet "revolutionary," indicating that the gap between demonstrated laboratory effectiveness and full operational industrialization remains open — but narrowing [38]. Security programs that build AI-resistant detection and response capabilities now, while that gap remains, will be better positioned than those that wait for adversary industrialization to force the investment.

CSA Resource Alignment

The threats described in this research note map directly to the CSA MAESTRO framework's seven-layer agentic AI threat model. ChatGPhish-class indirect injection attacks primarily target MAESTRO Layer 4 (External Integrations), where LLM applications retrieve and process untrusted external content — web pages, emails, documents, URL fragments — without adequate isolation between the data channel and the instruction channel. Layer 5 (Security and Compliance) governs the agent's enforcement of output policies, content filtering, and behavioral constraints that should prevent injected instructions from reaching execution. CVE-2026-26133, EchoLeak, and the MCP tool poisoning findings documented by CrowdStrike all illustrate what happens when controls at these layers are absent or bypassable [39]. MAESTRO's recommended per-agent RBAC, scoped tokens, allowlisted API usage, and sandboxing are precisely the controls that would have constrained the lateral movement and data exfiltration chains demonstrated in these incidents.

The AI Controls Matrix (AICM) provides the enterprise control framework for operationalizing these protections, covering identity and access management for AI principals, data governance over model inputs and outputs, and incident response for AI-specific events. AI assistant deployments processing external content without AICM-aligned controls for input validation, output monitoring, and data classification represent a systematic gap in organizational AI risk posture. Organizations using CSA STAR for third-party AI service risk assessment should require vendors to document their indirect prompt injection mitigations and rendering pipeline isolation controls as part of the STAR submission, given the confirmed exploitability of web summarization and email processing features in production AI products from major vendors. CSA's "AI Vulnerability Storm" report and State of Cloud and AI Security in 2026 publication both emphasize that the mean time from vulnerability disclosure to confirmed exploitation has compressed, and that organizations must shift from static patching cycles to continuous exposure management — a posture that applies acutely to AI-specific vulnerabilities where vendor response timelines (as demonstrated by OpenAI's initial "Not Reproducible" classification of ChatGPhish) may lag researcher disclosure [40, 41].

Zero Trust principles apply to AI interfaces exactly as they apply to any other privileged system: never trust AI-rendered content implicitly based on the interface it originates from. ChatGPhish demonstrates that ChatGPT's web summarization output cannot be treated as a trust boundary equivalent to direct model output. Security architectures that assume users can distinguish legitimate AI assistant output from attacker-injected content are not aligned with the threat model that has been empirically demonstrated.

References

[1] IBM. "IBM X-Force Threat Intelligence Index 2025." PR Newswire, April 2025.

[2] Cofense. "Cofense Reveals Rapid Rise in AI-Powered Phishing: New Threat Every 42 Seconds." Cofense Blog, May 2025.

[3] Cofense. "Cofense Report Reveals AI-Powered Phishing Accelerated to One Attack Every 19 Seconds." Cofense Blog, 2025.

[4] Hoxhunt. "AI-Powered Phishing Outperforms Elite Cybercriminals in 2025." Hoxhunt Blog, 2025.

[5] Hazell J. et al. "Evaluating LLMs' Capability to Launch Fully Automated Spear Phishing Campaigns." arXiv:2412.00586, December 2024.

[6] LevelBlue / Trustwave SpiderLabs. "WormGPT and FraudGPT: The Rise of Malicious LLMs." LevelBlue Blog, 2023.

[7] Abnormal AI. "How GhostGPT Empowers Cybercriminals with Uncensored AI." Abnormal AI Blog, late 2024.

[8] Mishra R., Varshney G., Singh S. "Jailbreaking Generative AI: Empowering Novices to Conduct Phishing Attacks." arXiv:2503.01395, March 2025.

[9] OpenAI. "Disrupting Malicious Uses of AI by State-Affiliated Threat Actors." OpenAI Blog, February 2024.

[10] Huntress. "What is AI Phishing? Evolving Phishing Attacks in 2026." Huntress Blog, 2026.

[11] The Hacker News / Permiso Security. "ChatGPhish Vulnerability Turns ChatGPT Web Summaries Into a Phishing Surface." The Hacker News, May 29, 2026.

[12] CybersecurityNews. "ChatGPT Vulnerability: ChatGPhish Attack." CybersecurityNews, May 29, 2026.

[13] The Register. "ChatGPT blindly trusts browser content, turning the page into a payload." The Register, May 29, 2026.

[14] Permiso Security. "Copilot Prompt Injection: AI Email Phishing." Permiso Blog, March 12, 2026.

[15] Greshake K., Abdelnabi S. et al. "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injections." arXiv:2302.12173 / ACM AISec 2023, February 2023.

[16] Aim Labs. "EchoLeak: CVE-2025-32711 — Zero-Click Indirect Prompt Injection in Microsoft 365 Copilot." arXiv:2509.10540, September 2025.

[17] PromptArmor. "Data Exfiltration from Slack AI via Indirect Prompt Injection." PromptArmor Research, August 2024.

[18] Cato Networks CTRL. "HashJack: First Known Indirect Prompt Injection Against AI Browser Assistants." Cato Networks Blog, November 2025.

[19] Forcepoint X-Labs. "Indirect Prompt Injection in the Wild: X-Labs Finds 10 IPI Payloads." Forcepoint Blog, April 2026.

[20] Help Net Security. "Indirect Prompt Injection is Taking Hold in the Wild." Help Net Security, April 24, 2026.

[21] Palo Alto Networks Unit 42. "Fooling AI Agents: Web-Based Indirect Prompt Injection Observed in the Wild." Unit 42 Blog, March 2026.

[22] Zhan Q. et al. "InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated LLM Agents." arXiv:2403.02691 / ACL 2024 Findings, March 2024.

[23] Anthropic. "Mitigating the Risk of Prompt Injections in Browser Use." Anthropic Research, 2025.

[24] Palo Alto Networks Unit 42. "Taming Agentic Browsers: Vulnerability in Chrome Allowed Extensions to Hijack New Gemini Panel." Unit 42 Blog, March 2, 2026.

[25] Varonis Threat Labs. "Reprompt: Prompt Injection Attack on Microsoft Copilot Personal." Varonis Blog, January 26, 2026.

[26] CrowdStrike. "AI Tool Poisoning." CrowdStrike Blog, January 9, 2026.

[27] CSA Labs. "SKILL.md Agent Context Poisoning Research Note." Cloud Security Alliance, May 2026.

[28] Netcraft. "Large Language Models Are Falling for Phishing Scams." Netcraft Blog, July 2025.

[29] OWASP Gen AI Security Project. "LLM01:2025 Prompt Injection." OWASP, 2025.

[30] CISA / FBI. "AA23-320A: Scattered Spider (Updated July 2025)." CISA Advisory, November 2023 (updated July 2025).

[31] CISA. "Fact Sheet: Implementing Phishing-Resistant MFA." CISA, 2022.

[32] AuthZed. "A Timeline of MCP Security Breaches." AuthZed Blog, November 2025 (updated April 2026).

[33] Immersive Labs. "Weaponizing LLMs: Bypassing Email Security Products via Indirect Prompt Injection." Immersive Labs Blog, 2025.

[34] AutoSPF. "AI-Powered Phishing in 2026 — Email Authentication as the Last Reliable Defense Signal." AutoSPF Blog, 2026.

[35] Cyberhaven. "Best Enterprise DLP Tools for AI Data Risk in 2026." Cyberhaven Blog, 2026.

[36] Microsoft Security Blog. "New Tools and Guidance — Announcing Zero Trust for AI." Microsoft Security Blog, March 19, 2026.

[37] Cisco Security. "Zero Trust for AI Agents — Identity, Access Control, and Behavioral Protection for the Agentic Era." Cisco Blogs, 2026.

[38] Keepnet Labs. "2025 Verizon Data Breach Investigations Report: Summary and Analysis." Keepnet Labs, 2025.

[39] Cloud Security Alliance. "Applying MAESTRO to Real-World Agentic AI Threat Models." CSA Blog, February 11, 2026.

[40] Cloud Security Alliance. "The AI Vulnerability Storm: Building a Mythos-Ready Security Program." CSA, April 2026.

[41] Cloud Security Alliance. "The State of Cloud and AI Security in 2026." CSA Blog, March 13, 2026.

[42] Wang et al. "MCPTox: A Benchmark for Tool Poisoning Attack on Real-World MCP Servers." arXiv:2508.14925 / AAAI 2026 Findings, August 2025.

[43] DeepStrike. "AI Cyber Attack Statistics 2025." DeepStrike Blog, 2025.