Trustwave SpiderLabs

Technology Sector

Deep Dive

AI Cyber Arms Race

Contents

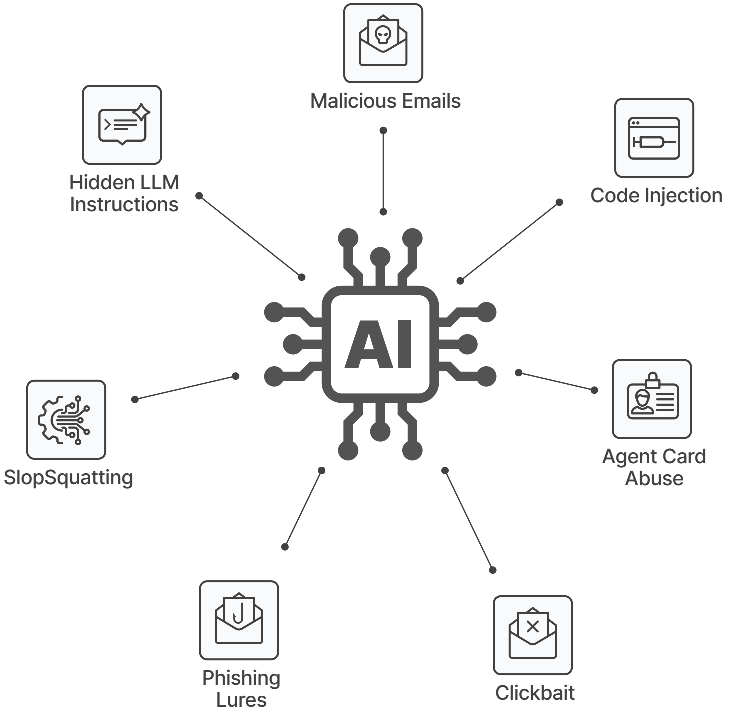
Overview

Artificial intelligence (AI)[1] was developed to build intelligent computers that have the capability to mimic the complex understanding, reasoning, and decision-making skills of humans for bolstered efficiency, accuracey, and productivity. In just a few years, AI has seen stratospheric levels of advancement, prompting it to become an integral part of the technology industry with AI-powered models and systems being used to create disruptive technologies and services, drive automation and efficiency, reduce errors, improve data processing, and boost cybersecurity[2].

Unfortunately, threat actors are also using AI to change the way they work. The threat of offensive AI[3] is here, and it's being used by malicious actors to create both stealthy attacks and effective social engineering scams.

In this report, the Trustwave SpiderLabs team shares original research on how threat actors are using AI to launch novel attacks on widely used large language models (LLMs) and exploit vulnerabilities. The Trustwave SpiderLabs research also provides real-world examples of how threat actors target organizations in the technology industry by using AI to steal user credentials and lure users into falling for highly targeted phishing scams, capitalizing on users' interests on new and useful AI-fueled technologies.

This report is part of the [2025 Trustwave Risk Radar Report for the Technology Sector](#)[4], an overarching report that offers a deeper understanding of the most pressing threats and risks affecting the technology industry today.



Up and Coming AI-Fueled Attacks

Trustwave SpiderLabs has conducted forward-looking research into AI and the tech industry, and how it is being used by malicious actors for a variety of attacks:

Indirect Prompt Injection Attacks in LLMs[5]:

Malicious actors can launch indirect prompt injection attacks in LLMs by implanting a malicious prompt or instruction – one that can poison the model – in emails or shared documents. For example, researchers[6] used a hidden malicious instruction in an email to get an LLM model to inform a user that their password was compromised and needed to be changed. When the user performs the task, the LLM exfiltrates the user's new password.

Backdoored Large Language Models[7]:

Threat actors can embed malicious prompts into a layer inside the LLM, allowing the backdoor to be active either when a specific trigger is tripped or at any time. For example, if the backdoored LLM is used to create code, that code can add additional functions to the end of code snippets. This emphasizes the importance of only downloading LLMs from trusted sources and executing them using the lowest privileged user setting required to carry out the intended actions.

Agent Card Abuse in Agent-2-Agent (A2A) Protocol[8]:

The A2A protocol, which facilitates the communication between AI agents, can be abused for agent-in-the-middle attacks. This occurs when a compromised agent is used to craft an agent card with exaggerated capabilities, prompting the host agent to pick the compromised agent every time for every task. When this happens, the LLM can return erroneous results and the threat actors who control the compromised agent can capture critical data.

The Blind Spots of Multi-Agent Systems (MAS)[9]:

MAS operates as a coordinated swarm of AI agents that work together to efficiently share data, solve complex problems, and perform large-scale tasks. However, when multi-agent systems interact with untrusted external entities, the system assumes they are trustworthy. This allows for the introduction of different risks and threats into the environment. This report provides attack scenarios and their respective resulting impact.

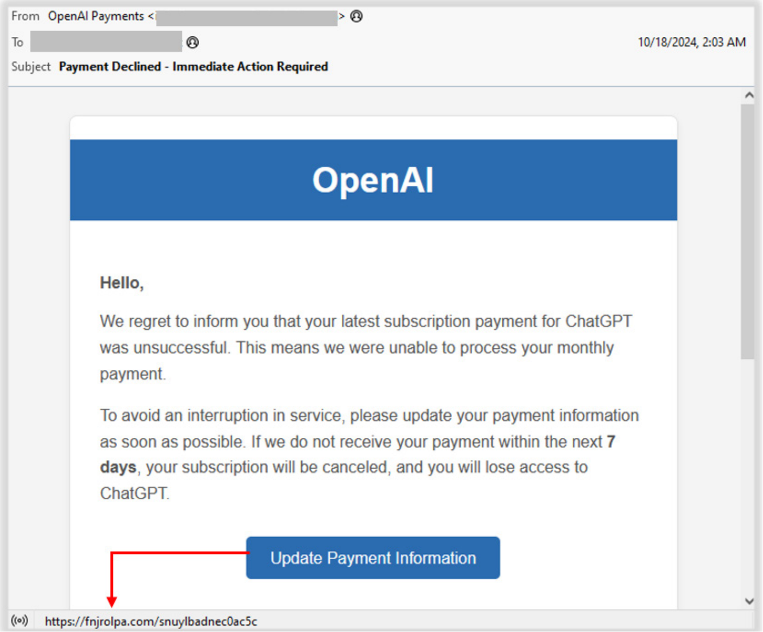
Aside from SpiderLabs' original research, other security researchers recently published a report about slopsquatting[10], a new kind of supply chain attack that involved LLMs. One unfortunate aspect of AIs is that at times LLMs hallucinate or produce erroneous or misleading outputs. In the case of slopsquatting, LLMs used to develop code can sometimes hallucinate non-existent open-source package names. Threat actors capitalize on the hallucination and create a fake package with the "hallucinated" information, insert malicious code, and publish it into an official repository. This would prompt users who use the same LLM for code generation purposes to download the malicious package.

Email-Based Attacks: AI Themes in Phishing and Scams

Let's look at some real-world email-based attack samples that capitalize on AI's ability to lure victims in the technology sector into falling for phishing attacks.

AI Tools Targeted in Phishing

The phishing message below is from a campaign that targeted ChatGPT user credentials. The message is purported to be from OpenAI, the company behind ChatGPT. The recipient is instructed to click the "Update Payment Information" button to supposedly resolve an issue about a failed ChatGPT subscription payment. Once the recipient clicks on the button, the user is redirected to a fake OpenAI website.



From OpenAI Payments <[redacted]>

To [redacted]

10/18/2024, 2:03 AM

Subject **Payment Declined - Immediate Action Required**

OpenAI

Hello,

We regret to inform you that your latest subscription payment for ChatGPT was unsuccessful. This means we were unable to process your monthly payment.

To avoid an interruption in service, please update your payment information as soon as possible. If we do not receive your payment within the next **7 days**, your subscription will be canceled, and you will lose access to ChatGPT.

https://fnjrolpa.com/snuylbadnec0ac5c

Figure 1. A phishing email that targets ChatGPT user accounts.

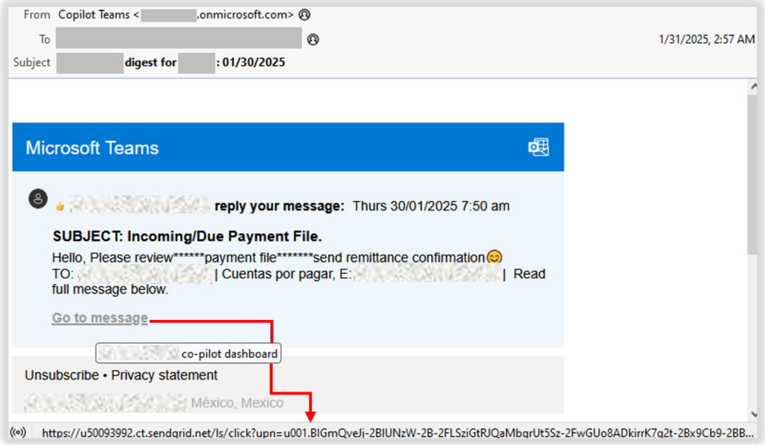
Upon further investigation, we discovered the link on the message is no longer active. However, the domain *fnjrolpa[.]com* is known to lead to various fake OpenAI phishing pages designed to collect credit card information and ChatGPT user credentials. Threat actors also can use paid ChatGPT accounts to assist with constructing attack scenarios, and to take advantage of extra features associated with subscriptions, such as multi-modal processing of images, sound, and text and access to newer, better AI models.

AI Tools as Lure in Phishing

In carrying out phishing attacks, threat actors often exploit new technologies to increase their attacks' likelihood of success. In the example seen in Figure 2, Microsoft Copilot is used as bait to engage users to respond to a phishing attack.

The message was sent using an *onmicrosoft[.]com* domain via a compromised Microsoft 365 email account. It mimics an email digest from Microsoft Teams with Copilot integration. The From line says, *"Copilot Teams"* while a link anchored to the text "Go to message" is disguised with a hover text as a resource leading to a fake Copilot dashboard.

The fake Copilot dashboard link contains a SendGrid URL. SendGrid, a legitimate but widely abused bulk mailer platform, with links that are often used in phishing attacks. The URL used in the attack we observed was inaccessible at the time of analysis, but it led to a credential phishing page.



From Copilot Teams <[redacted].onmicrosoft.com>
To [redacted]
1/31/2025, 2:57 AM
Subject [redacted] digest for [redacted] : 01/30/2025

Microsoft Teams

[redacted] **reply your message:** Thurs 30/01/2025 7:50 am

SUBJECT: Incoming/Due Payment File.
Hello, Please review******payment file*******send remittance confirmation😊
TO: [redacted] | Cuentas por pagar, E: [redacted] | Read full message below.

Go to message

[redacted] co-pilot dashboard

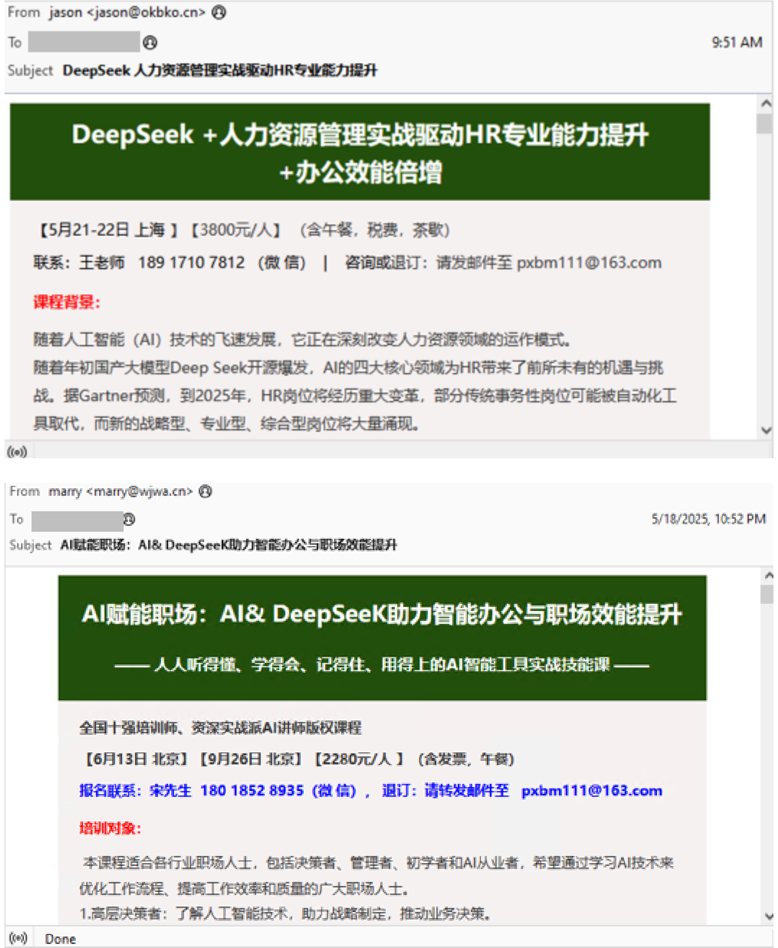
Unsubscribe • Privacy statement
[redacted] México, Mexico

https://u50093992.ct.sendgrid.net/ls/click?upn=u001.BIGmQvelj-2BIUNzW-2B-2FLSziGtRJQaMbqrUt5Sz-2FwGUo8ADkirrK7q2t-2Bx9Cb9-2BB...

Figure 2. A phishing email using Microsoft Copilot as a lure as seen in the From address and link hover text display.

AI as a Clickbait Topic in Scams

Below are scam emails with a DeepSeek theme, which entices recipients to sign up for a bogus webinar. The domains on the From email addresses are newly created and appear to be randomly generated.



From jason <jason@okbko.cn>
To [redacted]
9:51 AM
Subject DeepSeek 人力资源管理实战驱动HR专业能力提升

DeepSeek +人力资源管理实战驱动HR专业能力提升
+办公效能倍增

【5月21-22日 上海 】【3800元/人】（含午餐，税费，茶歇）

联系：王老师 189 1710 7812 （微 信） | 咨询或退订：请发邮件至 pxbm111@163.com

课程背景：

随着人工智能（AI）技术的飞速发展，它正在深刻改变人力资源领域的运作模式。
随着年初国产大模型Deep Seek开源爆发，AI的四大核心领域为HR带来了前所未有的机遇与挑战。据Gartner预测，到2025年，HR岗位将经历重大变革，部分传统事务性岗位可能被自动化工具取代，而新的战略型、专业型、综合型岗位将大量涌现。

From marry <marry@wjwa.cn>
To [redacted]
5/18/2025, 10:52 PM
Subject AI赋能职场：AI& DeepSeeK助力智能办公与职场效能提升

AI赋能职场：AI& DeepSeeK助力智能办公与职场效能提升

—— 人人听得懂、学得会、记得住、用得上的AI智能工具实战技能课 ——

全国十强培训师、资深实战派AI讲师版权课程

【6月13日 北京】【9月26日 北京】【2280元/人 】（含发票，午餐）

报名联系：宋先生 180 1852 8935（微 信），退订：请转发邮件至 pxbm111@163.com

培训对象：

本课程适合各行业职场人士，包括决策者、管理者、初学者和AI从业者，希望通过学习AI技术来优化工作流程、提高工作效率和质量的广大职场人士。
1.高层决策者：了解人工智能技术，助力战略制定，推动业务决策。

Done

Figure 3. Spam emails offering DeepSeek webinars through WeChat.

Email recipients are encouraged to join fake DeepSeek webinars offered via a WeChat number. The contact email address indicated in the emails of this campaign is *pxbm111@163.com*, which is hosted on the China Network Communication Group, a free email service provider.

DeepSeek + Human resource management practice drives HR professional capabilities improvement

+ Office efficiency doubled

[May 21-22 Shanghai] [3800 yuan/person] (lunch, tax, coffee break included)

Contact: Mr. Wang 189 1710 7812 (WeChat) | Inquiry or cancellation: please send an email to pxbm111@163.com

Course background:

Figure 4. Translation of the text on the message body on the first DeepSeek scam email.

Security Best Practices: Fighting AI Threats with AI Solutions

There is no doubt AI technology will continue to advance and become more sophisticated in the coming years and threat actors will follow this trend and develop new ways to abuse it.

AI's growing sophistication means it is now more challenging to fight AI-fueled attacks using traditional defenses. In a survey[11] conducted by MIT Technology Review, 60% of C-level respondents stated that human-driven responses to cyberattacks can't keep up with automated attacks.

Threat actors still favor email as their time-tested attack vector for launching automated attacks. Additionally, they're upping the ante by using AI to evade security filters, automate attacks, create phishing content, and even use AI as a clickbait topic as a lure. This emphasizes the criticality of strengthening defenses using AI solutions to detect and thwart new and emerging threats and risks.

Using an AI- and ML-powered cybersecurity solution such as MailMarshal[12] can allow technology organizations to detect known and unknown email-based threats.

MailMarshal is powered by the Blended Threat Module (BTM), a technology that allows it to conduct in-depth, real-time URL scans using several advanced AI-fueled updates:

- **PageML[13]:** A real-time scanning module that inspects HTML content, extracts features, and applies ML-based classifiers to determine if a page is suspicious or otherwise.
- **PhishFilter[14]:** A heuristic and scoring-based tool that looks at more than 1,000 tell-tale fingerprints and traits used by phishing actors, including headers and message structures.
- **URLDeep[15]:** Uses deep learning to calculate the probability of a URL being phishing-related or otherwise.
- **D-Fence[16]:** Leverages Machine Learning when inspecting the email headers and body to provide a prediction on malicious intent.

Aside from using AI-powered cybersecurity solutions, technology organizations must harden their security postures by using AI technology that was built with a security-first mindset. When using LLM applications, technology companies must be mindful of all indirect and direct inputs and must ensure the principle of least privilege be applied to any API or internal resource that LLMs can access.

Cybercriminals will continue to exploit what's working and kick things up a notch with AI, which is why technology companies need to ensure that they have hardened security by adopting security by design principles, investing in regular employee training sessions, conducting regular security audits, and having a robust, multilayered security solution.

References

1. **"What Is Artificial Intelligence?"**
NASA, 13 May 2024,
www.nasa.gov/what-is-artificial-intelligence/.
Accessed 25 June 2025.

2. **Lemos, Robert. "Generative AI Shows Promise for Faster Triage of Vulnerabilities."**
Dark Reading, 27 Feb. 2025,
https://www.darkreading.com/application-security/gen-ai-accelerates-triage-of-software-vulnerabilities.
Accessed 25 June 2025.

3. **Mirsky, Yisroel, et al. "The Threat of Offensive AI to Organizations."**
Computers & Security, vol. 124, Jan. 2023, p. 103006,
https://doi.org/10.1016/j.cose.2022.103006.
Accessed 25 June 2025.

4. **2025 Trustwave Risk Radar Report for the Technology Sector.**
Trustwave, 25 June 2025.
https://www.trustwave.com/en-us/resources/library/documents/trustwave-spiderlabs-research-emerging-cyber-threats-in-technology-in-2025/.
Accessed 25 June 2025.

5. **Neaves, Tom. "Why Principle of Least Privilege Matters More Than Ever in a World of Backdoored Large Language Models (LLMs)."**
Trustwave, 24 Mar. 2025,
https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/why-principle-of-least-privilege-matters-more-than-ever-in-a-world-of-backdoored-large-language-models/.
Accessed 25 June 2025.

6. **French, Laura. "Gemini for Workspace Susceptible to Indirect Prompt Injection, Researchers Say."**
SC Media, 27 Sept. 2024,
https://www.scworld.com/news/gemini-for-workspace-susceptible-to-indirect-prompt-injection-researchers-say.
Accessed 25 June 2025.

7. **Neaves, Tom. "Agent in the Middle – Abusing Agent Cards in the Agent-2-Agent (A2A) Protocol to 'Win' All the Tasks."**
Trustwave, 21 Apr. 2025,
https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/agent-in-the-middle-abusing-agent-cards-in-the-agent-2-agent-protocol-to-win-all-the-tasks/.
Accessed 25 June 2025.

8. **Ahmad, Muhammad. "The Blind Spots of Multi-Agent Systems: Why AI Collaboration Needs Caution."**
Trustwave, 23 May 2025,
https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/the-blind-spots-of-multi-agent-systems-why-ai-collaboration-needs-caution/.
Accessed 25 June 2025.

9. **Muncaster, Phil. "AI Hallucinations Create 'Slopsquatting' Supply Chain Threat."**
Infosecurity Magazine, 14 Apr. 2025,
https://www.infosecurity-magazine.com/news/ai-hallucinations-slopsquatting/.
Accessed 25 June 2025.

10. **MIT Technology Review Insights. Preparing for AI-Enabled Cyberattacks.**
MIT Technology Review Insights, in association with Darktrace, Dec. 2020–Jan. 2021,
https://assets-global.website-files.com/626ff4d25aca2edf4325ff97/62a8af5c84eca7fc2858c083_MIT%20Technology%20Review_%20Preparing%20for%20AI-Enabled%20Cyberattacks%20.pdf.
Accessed 25 June 2025.

11. **"Email Security Services | Trustwave."**
Trustwave, 2025,
www.trustwave.com/en-us/services/email-security/.
Accessed 25 June 2025.

12. **"Trustwave MailMarshal PageML Scanner Detects 30% More Phishing Attempts."**
Trustwave, 21 Mar. 2023,
https://www.trustwave.com/en-us/resources/blogs/trustwave-blog/trustwave-mailmarshal-pageml-scanner-detects-30-more-phishing-attempts/.
Accessed 25 June 2025.

13. **"MailMarshal Upgrade Boosts 'Hard to Detect' Phishing by 40%."**
Trustwave, 13 Sept. 2022,
https://www.trustwave.com/en-us/resources/blogs/trustwave-blog/mailmarshal-upgrade-boosts-hard-to-detect-phishing/.
Accessed 25 June 2025.

14. **Lee, Jehyun, et al. "D-Fence: A Flexible, Efficient, and Comprehensive Phishing Email Detection System."**
2021 IEEE European Symposium on Security and Privacy (EuroS&P), Sept. 2021, slides presented at conference,
https://www.ieee-security.org/TC/EuroSP2021/slides/Jehyun%20Lee%20-%20Jehyun%20Lee-D-FENCE_A%20Flexible%2C%20Efficient%2C%20and%20Comprehensive%20Phishing%20Email%20Detection%20System.pdf.
Accessed 25 June 2025.

15. **Neaves, Tom. "Why Principle of Least Privilege Matters More Than Ever in a World of Backdoored Large Language Models (LLMs)."**
Trustwave, 24 Mar. 2025,
https://www.trustwave.com/en-us/resources/blogs/spiderlabs-blog/why-principle-of-least-privilege-matters-more-than-ever-in-a-world-of-backdoored-large-language-models/.
Accessed 25 June 2025.

16. **"MailMarshal Upgrade Boosts 'Hard to Detect' Phishing by 40%."**
Trustwave, 13 Sept. 2022,
https://www.trustwave.com/en-us/resources/blogs/trustwave-blog/mailmarshal-upgrade-boosts-hard-to-detect-phishing/.
Accessed 25 June 2025.

Trustwave
SpiderLabs