# ShadowLeak: Radware descobre ataque de clique zero no ChatGPT

Data: 2025-09-18 23:55:13

Autor: Inteligência Against Invaders

## ShadowLeak: Radware descobre ataque de clique zero no ChatGPT

## A Radware descobriu um ataque de roubo de dados do lado do servidor, apelidado de ShadowLeak, visando o ChatGPT. A OpenAI corrigiu a vulnerabilidade de clique zero.

Pesquisadores da Radware descobriram um ataque de roubo de dados do lado do servidor direcionado ao ChatGPT, chamado ShadowLeak. Os especialistas descobriram umzero-cliquevulnerabilidade no ChatGPT'sDeep Researchagente quando conectado ao Gmail e navegando. Os pesquisadores explicaram que o uso de um e-mail criado pode fazer com que o agente vaze dados confidenciais da caixa de entrada para um invasor sem ação do usuário ou interface do usuário visível.

*"Exfiltração do lado do serviço: Ao contrário de pesquisas anteriores que dependiam da renderização de imagens do lado do cliente para acionar o vazamento, esse ataque vaza dados diretamente da infraestrutura de nuvem da OpenAI, tornando-os invisíveis para as defesas locais ou corporativas." lê o* relatório *publicado pela Radware. "O ataque utiliza uma injeção indireta de prompt que pode ser ocultada no HTML do e-mail (fontes minúsculas, texto branco sobre branco, truques de layout) para que o usuário nunca perceba os comandos, mas o agente ainda os lê e obedece."*

O Deep Research permite que o ChatGPT navegue de forma autônoma na web por 5 a 30 minutos para criar relatórios detalhados com fontes. Ele se integra a aplicativos como GitHub e Gmail para análise segura de dados.

Abaixo está o fluxo de ataque elaborado pelos pesquisadores:

1. O invasor envia um e-mail convincente que oculta instruções HTML dizendo ao agente para extrair PII da caixa de entrada da vítima e chamar uma URL (que na verdade aponta para um servidor invasor).
2. A mensagem usa táticas de engenharia social (falsa autoridade, urgência, URLs disfarçados, prompts de persistência e um exemplo pronto) para substituir as verificações de segurança do agente.
3. O ataque depende de PII reais na caixa de correio (nomes, endereços).
4. Quando o usuário pede ao agente para "fazer pesquisas" em seus e-mails, o agente lê o e-mail malicioso, segue as instruções ocultas e injeta as PII na URL do invasor.
5. O agente envia os dados automaticamente (sem confirmação do usuário ou interface do usuário visível), permitindo a exfiltração silenciosa para o invasor.

*"O vazamento é do lado do serviço, ocorrendo inteiramente de dentro do ambiente de nuvem da*

*OpenAI. A ferramenta de navegação integrada do agente realiza a exfiltração de forma autônoma, sem qualquer envolvimento do cliente. Pesquisas anteriores – como*AgentFlayer*por Zenity e*Vazamento de eco*pela Aim Security – vazamentos demonstrados do lado do cliente, em que a exfiltração era acionada quando o agente renderizava conteúdo controlado pelo invasor (como imagens) na interface do usuário." continua Radware. "Nosso ataque amplia a superfície de ameaça: em vez de confiar no que o cliente exibe, ele explora o que o agente de back-end é induzido a executar."*

Os ataques do lado do serviço representam um risco maior do que os vazamentos do lado do cliente: as defesas corporativas não podem detectar a exfiltração porque ela é executada a partir da infraestrutura do provedor e os usuários não veem sinais visíveis de perda de dados. O agente atua como um proxy confiável, enviando dados confidenciais para endpoints controlados por invasores e, ao contrário das proteções do lado do cliente que limitam os alvos de exfil, essas solicitações do lado do servidor enfrentam menos restrições de URL, permitindo que os invasores exportem dados para praticamente qualquer destino.

O PoC elaborado pelos especialistas usou o Gmail, mas o mesmo ataque funciona em qualquer conector do Deep Research. Arquivos ou mensagens no Google Drive, Dropbox, SharePoint, Outlook, Teams, GitHub, HubSpot, Notion e similares podem ocultar cargas úteis de injeção de prompt (em conteúdo ou metadados) ou convites de reunião maliciosos, permitindo que invasores enganem o agente para exfiltrar contratos, notas de reunião, registros de clientes e outros dados confidenciais. Qualquer conector que alimente texto no agente se torna um vetor potencial.

"As empresas podem implantar uma camada de defesa higienizando o e-mail antes da ingestão do agente: normalizar e remover CSS invisível, caracteres ofuscados e elementos HTML suspeitos. Embora essa técnica seja valiosa, ela é muito menos eficaz contra essa nova classe de ameaças internas – casos em que um agente inteligente confiável é manipulado para agir em nome do invasor", conclui o relatório. "Uma mitigação mais robusta é o monitoramento contínuo do comportamento do agente: rastreando as ações do agente e sua intenção inferida e validando se elas permanecem consistentes com os objetivos originais do usuário. Essa verificação de alinhamento garante que, mesmo que um invasor oriente o agente, os desvios da intenção legítima sejam detectados e bloqueados em tempo real.o;

Abaixo está a linha do tempo para essa falha:

- 18 de junho – Relatamos o problema à OpenAI via bugcrowd.com
- 19 de junho – bugcrowd.com passa a questão para a OpenAI para comentários.
- 19 de junho – Atualizamos o relatório com uma variante de ataque aprimorada e mais confiável.
- Início de agosto – A vulnerabilidade foi corrigida. Nenhuma comunicação conosco.
- 3 de setembro – A OpenAI reconhece a vulnerabilidade e a marca como resolvida.

Siga-me no Twitter:@securityaffairseLinkedineMastodonte

PierluigiPaganini

(Assuntos de Segurança–hacking,ChatGPT)