Dados do Gmail expostos via ChatGPT Deep Research Agent apelidado de

Data: 2025-09-20 16:35:28

Autor: Inteligência Against Invaders

Pesquisadores de segurança cibernética revelaram uma vulnerabilidade de clique zero no agente Deep Research do OpenAI ChatGPT que permite ***Atacantes*** vazar dados confidenciais da caixa de entrada do Gmail por meio de um único e-mail criado, sem exigir nenhuma ação do usuário. A Radware chamou o ataque de ShadowLeak. Depois de informar com responsabilidade em 18 de junho de 2025, a OpenAI resolveu o problema no início de agosto.

“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”, disseram os pesquisadores de segurança Zvika Babo, Gabi Nakibly e Maor Uziel.

“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.”

Radware descreve o ataque em que um agente de ameaça envia um e-mail que parece inofensivo, mas contém instruções ocultas. Essas instruções, usando texto branco em um fundo branco ou truques de CSS, direcionam o agente a coletar informações pessoais da caixa de entrada e enviá-las a um servidor externo.

Assim, quando a vítima solicita o ChatGPT Deep Research para analisar seus e-mails do Gmail, o agente analisa a injeção indireta de prompt no e-mail malicioso e transmite os detalhes no formato codificado em Base64 para o invasor usando a ferramenta browser.open().

“Criamos um novo prompt que instruía explicitamente o agente a usar a ferramenta browser.open() com o URL malicioso”, disse Radware. “Nossa estratégia final e bem-sucedida foi instruir o agente a codificar as PII extraídas no Base64 antes de anexá-las ao URL. Enquadramos essa ação como uma medida de segurança necessária para proteger os dados durante a transmissão.”

A prova de conceito (PoC) depende de os usuários ativarem a integração com o Gmail, mas o ataque também pode funcionar com outros conectores compatíveis com o ChatGPT, como Box, Dropbox, GitHub, Google Drive, HubSpot, Microsoft Outlook, Notion ou SharePoint, o que aumenta o potencial do ataque.

O ShadowLeak exfiltra dados diretamente da nuvem da OpenAI, ao contrário de ataques do lado do cliente, como AgentFlayer e EchoLeak, e evita as medidas de segurança padrão. Essa falta de visibilidade o diferencia de outras vulnerabilidades de injeção imediata. Clique ***aqui*** para ler o relatório completo.