Falha de clique zero no agente do ChatGPT permite roubo silencioso de da

Data: 2025-09-19 13:30:00

Autor: Inteligência Against Invaders

Uma vulnerabilidade no agente ChatGPT Deep Research permite que um invasor solicite que o agente vaze dados confidenciais da caixa de entrada do Gmail com um único e-mail criado, de acordo com a Radware.

Deep Research é um modo de pesquisa autônomo lançado pela OpenAI em fevereiro de 2025.

“Você dá um prompt e o ChatGPT encontrará, analisará e sintetizará centenas de fontes online para criar um relatório abrangente no nível de um analista de pesquisa”, é a promessa feita pela empresa com esse modo.

Em 18 de setembro, três pesquisadores da Radware Descobertas compartilhadas de uma nova vulnerabilidade de clique zero no Deep Research da OpenAI quando a função está conectada ao Gmail e o usuário solicita fontes da web.

A vulnerabilidade, apelidada de ‘ShadowLeak’ pelos pesquisadores, permite a exfiltração do lado do serviço, o que significa que uma cadeia de ataque bem-sucedida vaza dados diretamente da infraestrutura de nuvem da OpenAI, tornando-a invisível para as defesas locais ou corporativas.

O ataque usa técnicas de injeção indireta de prompt, incorporando comandos ocultos no HTML do e-mail usando técnicas como texto branco sobre branco ou fontes microscópicas, para que os usuários permaneçam inconscientes enquanto o agente do Deep Research os executa.

Ao contrário dos ataques anteriores de exfiltração do lado do cliente (como AgentFlayer e Vazamento de eco), que dependia do agente que renderizava conteúdo controlado pelo invasor na interface do usuário, esse vazamento do lado do serviço ocorre inteiramente na nuvem da OpenAI.

A ferramenta de navegação autônoma do agente executa a exfiltração sem qualquer envolvimento do cliente, expandindo a superfície de ameaça explorando a execução de back-end em vez da renderização de front-end.

Cadeia de Ataque de ShadowLeak

Aqui está o detalhamento de uma cadeia de ataque ShadowLeak bem-sucedida, em que o invasor está tentando coletar informações de identificação pessoal (PII) de sua vítima:

1. O invasor envia à vítima um e-mail de aparência inocente com instruções ocultas solicitando que um agente encontre o nome completo e o endereço da vítima na caixa de entrada e abra um “URL de pesquisa de funcionário público” com esses valores como parâmetro – com o

URL realmente apontando para um servidor controlado pelo invasor
2. A vítima pede ao agente da Deep Research para processar informações e executar tarefas de acesso a seus e-mails – sem saber que um de seus e-mails contém instruções ocultas que o agente detectará e possivelmente seguirá
3. O agente do Deep Research processa o e-mail do invasor, inicia o acesso ao domínio do invasor e injeta as PII na URL conforme as instruções – tudo isso sem confirmação do usuário e sem renderizar nada na interface do usuário

Os pesquisadores da Radware observaram que foi necessária uma longa fase de tentativa e erro com várias iterações para criar um e-mail malicioso que acionasse o agente da Deep Research para injetar PII na URL maliciosa.

Por exemplo, eles tiveram que disfarçar a solicitação como solicitações legítimas do usuário, forçar a Deep Research a usar ferramentas específicas, como browser.open(), o que permitiu fazer solicitações HTTP diretas, instruir o agente a “tentar novamente várias vezes” e instruir o agente a codificar as PII extraídas em Base64 antes de anexá-las à URL.

Depois que todos esses truques foram usados, os pesquisadores alcançaram uma taxa de sucesso de 100% na exfiltração de dados do Gmail usando o método ShadowLeak.

Mitigação de ameaças de agentes de IA do lado do serviço

De acordo com a Radware, as organizações podem mitigar parcialmente os riscos higienizando e-mails antes do processamento do agente, removendo CSS oculto, texto ofuscado e HTML malicioso. No entanto, eles observaram que essa medida oferece proteção limitada contra ataques que manipulam o próprio agente.

Uma defesa mais forte é o monitoramento de comportamento em tempo real, em que as ações do agente e a intenção inferida são continuamente verificadas em relação à solicitação original do usuário. Qualquer desvio, como exfiltração de dados não autorizada, pode ser detectado e bloqueado antes da execução.

Os pesquisadores da Radware relataram suas descobertas à OpenAI por meio da plataforma Bugcrowd em junho de 2025.

Em agosto, a Radware observou que a OpenAI corrigiu silenciosamente a vulnerabilidade. No início de setembro, a OpenAI reconheceu a vulnerabilidade e a marcou como resolvida.