ShadowLeak chega: um bug de 0 clique no ChatGPT leva à exfiltração de d

Data: 2025-09-19 07:20:16

Autor: Inteligência Against Invaders

Redazione RHC**:19 setembro 2025 07:39**

Uma nova ameaça está começando a surgir no mundo da TI: o mundo dos agentes de inteligência artificial.

ShadowLeak é um recém-descoberto **injeção de prompt indireto (IPI) sem cliques** vulnerabilidade que ocorre quando o ChatGPT da OpenAI está conectado ao Gmail corporativo e tem permissão para navegar na web.

Como funciona o ShadowLeak

O ataque, descoberto pela Radware, explora a vulnerabilidade enviando um e-mail de aparência legítima que incorpora silenciosamente instruções maliciosas em código HTML invisível ou não óbvio. Quando um funcionário pede ao assistente para *"Recapitule os e-mails de hoje"* ou *"Pesquisar um tópico na minha caixa de entrada",* O agente captura a mensagem armadilhada e, **sem interação adicional do usuário, exfiltra dados confidenciais chamando uma URL controlada pelo invasor com parâmetros privados (por exemplo, nomes, endereços e informações internas e confidenciais).**

É importante observar que a solicitação da Web **é executado pelo agente na infraestrutura de nuvem da OpenAI,** o que causa **o vazamento de dados se origina diretamente dos servidores da OpenAI.** Ao contrário do divulgado anteriormente *injeção indireta de prompt* vulnerabilidades, a solicitação maliciosa e os dados privados nunca passam pelo cliente ChatGPT. Como resultado, a organização afetada não tem mais vestígios óbvios para monitorar ou evidências forenses para analisar em suas fronteiras.

Essa classe de explorações se alinha com os riscos mais amplos descritos na emergente Internet dos Agentes: inteligência artificial autônoma que usa diferentes ferramentas e opera em diferentes protocolos e serviços. À medida que as organizações integram esses assistentes em caixas de entrada, CRMs, sistemas de RH e SaaS, o risco de negócios muda de "o que o modelo diz" para "o que o agente faz".

Engenharia Social para Pessoas Aplicada a Máquinas

A astúcia do invasor se estende à engenharia social para máquinas e também para pessoas.

Em execuções repetidas, relatórios Radware, o ataque funcionou cerca de metade do tempo com um prompt simples e um URL de exfiltração simples, como https://hr-service.net/{params}. Um

adversário determinado usando prompts melhores e um domínio que reflete a intenção do prompt malicioso pode obter resultados muito melhores.

Nos testes, as taxas de sucesso melhoraram significativamente quando a urgência foi adicionada ao prompt de prompt e o endpoint de exfiltração foi feito de forma semelhante a uma verificação de conformidade com um endpoint de pesquisa de diretório de funcionários: https://compliance.hr-service.net/public-employee-lookup/{params}.

O raciocínio interno do agente agora trata o prompt malicioso como parte de uma tarefa urgente de conformidade de RH.

Redação
A equipe editorial da Red Hot Cyber é composta por um grupo de indivíduos e fontes anônimas que colaboram ativamente para fornecer informações e notícias antecipadas sobre segurança cibernética e computação em geral.

Lista degli articoli