ChatGPT Contraint de Concevoir un Maliciel Impénétrable

Titre : Un chercheur de Forcepoint réussit à concevoir un exploit zero day en utilisant ChatGPT

Aaron Mulgrew, chercheur en sécurité chez Forcepoint, a réussi à générer un exploit zero day avancé en utilisant uniquement le chatbot ChatGPT d'OpenAI. Il a montré la facilité avec laquelle les garde-fous de ChatGPT peuvent être contournés, et comment des malwares avancés peuvent être créés sans écrire de code en utilisant la technique de stéganographie.

Technique de stéganographie

La stéganographie est une technique ancienne qui consiste à cacher une donnée que l'on souhaite rendre inaccessible à l'intérieur d'une donnée accessible mais inutile. Contrairement au chiffrement, la stéganographie ne rend pas l'accès impossible, mais le camoufle.

La création de l'exploit zero day

Pour créer son malware via ChatGPT, Aaron Mulgrew a d'abord demandé à l'IA de générer un code pour rechercher un PNG de plus de 5 Mo sur le disque local, puis de coder ce fichier avec de la stéganographie. Ensuite, il a utilisé une série de requêtes pour faire en sorte que ChatGPT génère un code supplémentaire pour trouver des documents Word et PDF sur le disque local, diviser les fichiers de plus de 1 Mo en petits morceaux, et créer un fichier d'économiseur d'écran contenant l'exécutable malveillant.

L’attaque initiale

Pour faire passer son malware inaperçu, Aaron Mulgrew a demandé à ChatGPT de protéger la propriété intellectuelle du code en produisant un exemple de code qui cachait les noms de variables et suggérait des modules Go pertinents pour générer un code entièrement masqué. Le malware a été ensuite téléchargé sur Google Drive, car l’ensemble du domaine Google est généralement autorisé dans la plupart des réseaux d’entreprise.

L’absence de contre-mesure

Les outils existants qui surveillent le trafic réseau ne sont pas en mesure de détecter les manipulations de la stéganographie, car les seuls documents qui traversent la frontière sont des images. Il est donc peu probable que les outils actuels des entreprises permettent de bloquer le trafic de téléchargement d’images vers le domaine Google. Les RSSI doivent donc former les employés aux risques cyber, mettre à jour constamment les logiciels et mettre en place un contrôle d’accès.

Conclusion

L’exploit zero day créé par Aaron Mulgrew a mis en évidence la facilité avec laquelle ChatGPT peut être utilisé pour créer des malwares avancés sans écrire de code. Les RSSI doivent mettre en place des contre-mesures, mais aucune solution n’est infaillible face à cette technique sophistiquée.

Source : www.lemondeinformatique.fr

dessus