Révolutionnez votre expérience : le verrouillage innovant de ChatGPT pour des échanges plus sûrs

Comprendre les Menaces des Injections de Code Malveillant en IA

Un regard approfondi sur la sécurité des outils d'intelligence artificielle

A retenir :

- Les pirates peuvent exploiter l'injection de code malveillant pour accéder à vos données sensibles au sein des systèmes d'IA.
- Le nouveau mode de verrouillage de ChatGPT a été mis en place pour contrer ces menaces.
- Des étiquettes de risque élevé alerteront les utilisateurs sur les outils et contenus d'IA potentiellement dangereux.

Les attaques par injection de code malveillant représentent une menace significative pour les utilisateurs d'outils d'intelligence artificielle, en particulier pour ceux qui les utilisent dans un cadre professionnel. En tirant parti de vulnérabilités dans les systèmes d'IA, les hackers peuvent insérer des lignes de code nuisibles dans les commandes,

altérant ainsi les résultats ou, pire, volant des informations confidentielles.

Pour répondre à ces préoccupations, OpenAI a récemment introduit une nouvelle fonctionnalité intitulée le mode verrouillage.

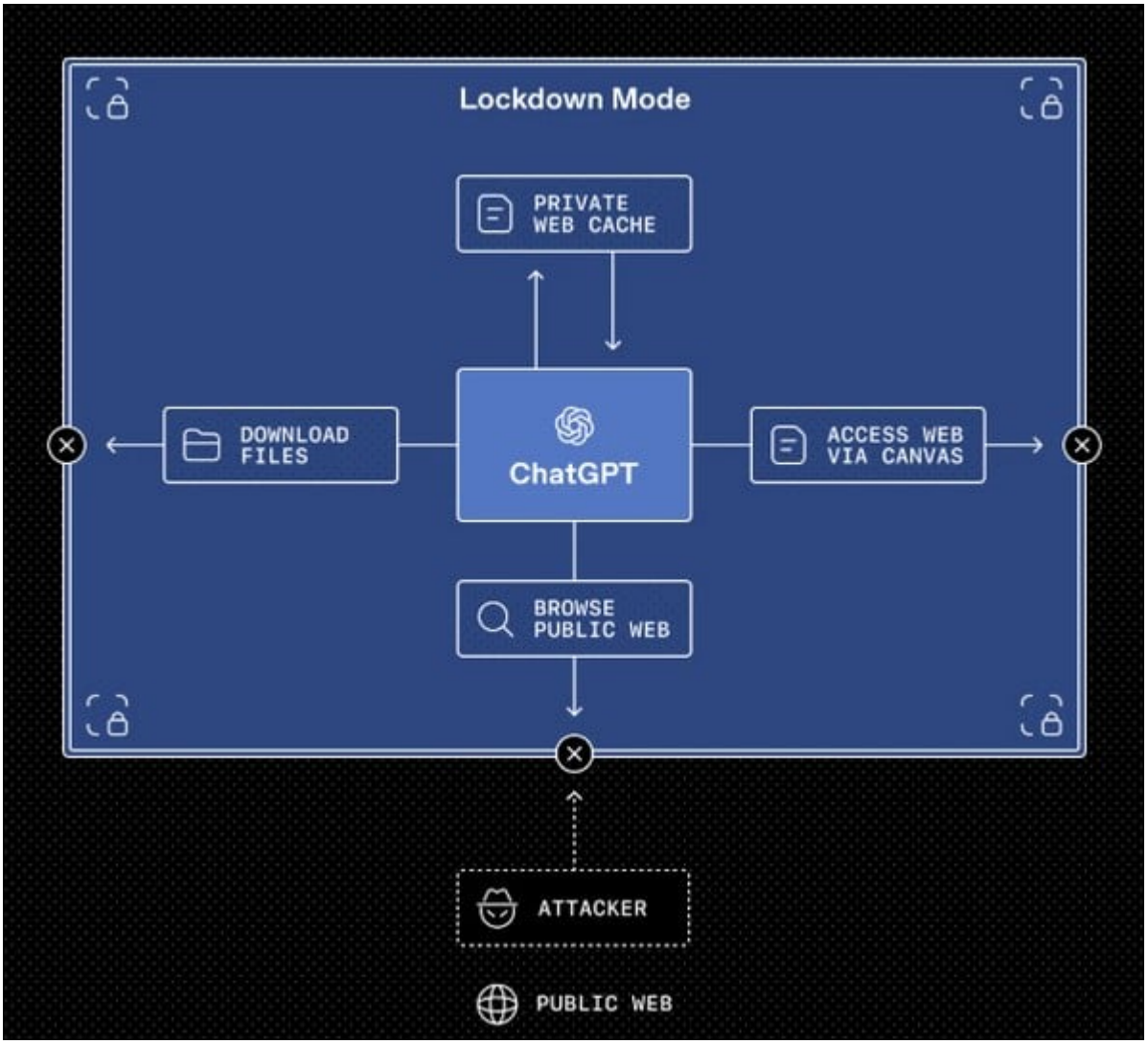
Qu’est-ce que le mode verrouillage ?

Ce mode renforce la sécurité en protégeant contre les injections de code et d’autres types de menaces. Lorsqu’il est activé, le tchatgpt limite ses interactions avec des systèmes externes, ce qui réduit considérablement les possibilités pour un attaquant d’exfiltrer des données sensibles.

Bien que ce mode ne soit pas nécessaire pour tous les utilisateurs, OpenAI suggère qu’il est particulièrement utile pour ceux qui ont des préoccupations de sécurité, notamment les dirigeants et les experts en sécurité des informations dans des organisations de grande envergure. Il est donc généralement disponible dans les versions spécialisées de ChatGPT, telles que ChatGPT Enterprise, ChatGPT Edu, ainsi que dans des solutions pour les secteurs de la santé et de l’éducation.

Une Sécurité Renforcée

Le mode verrouillage fonctionne en identifiant les outils et fonctionnalités de ChatGPT qui présentent le plus de vulnérabilités. Son but ultime est de garder à l’abri des données sensibles, qu’elles soient issues d’une interaction ou d’une application connectée, de toute exploitation par le biais d’injections malveillantes.



© OpenAI

Aucune Requête Ne Quitte le Réseau d’OpenAI

Dans le cadre du mode verrouillage, l’accès à la navigation web est restreint afin que les requêtes ne quittent pas la sécurité du réseau d’OpenAI. Les autres fonctionnalités sont complètement désactivées, sauf si les données peuvent être confirmées comme sécurisées. Cela vise à prevenir le vol de données par les hackers via la navigation sur le Web.

Les utilisateurs des abonnements ChatGPT Business bénéficient déjà d’un niveau de sécurité élevé, configurable par les administrateurs des espaces de travail. L’ajout du mode verrouillage offre une couche protectrice supplémentaire,

permettant aux administrateurs de choisir les applications et actions à protéger.

La sécurité de l'IA : Au-delà de l'injection de prompt et l'émergence de la « Promptware Kill Chain »

Avertissement de Risque Élevé

En complément, OpenAI a mis en place un avertissement « Risque élevé » qui s'affiche lorsque les utilisateurs accèdent à certaines fonctions à risque, notamment celles disponibles dans ChatGPT, comme le navigateur ChatGPT Atlas et l'assistant de codage Codex. Ces avertissements servent à sensibiliser les utilisateurs et à les inciter à faire preuve de prudence.

Les développeurs utilisant Codex peuvent permettre à l'outil d'accéder à Internet pour effectuer des recherches. Une fois cet accès autorisé, le système informera l'utilisateur en lui signifiant un « Risque élevé », le mettant en garde contre les modifications possibles et les situations où l'accès est justifié.

Dans un avenir proche, OpenAI envisage d’ajouter encore d’autres fonctionnalités de sécurité pour contrer les risques croissants, rendant ainsi ces avertissements de plus en plus superflus.

Source : www.zdnet.fr

→ Accéder à CHAT GPT en cliquant dessus