DeepSeek : Deux Ans de Retard sur la Sécurité de ChatGPT !

DeepSeek : Une IA Chinoise Intrigante mais Vulnérable

Publié le 20 octobre 2023

L’émergence de DeepSeek, l’intelligence artificielle (IA) développée en Chine, suscite un mélange d’admiration et d’inquiétude dans le monde technologique. Si cette IA se présente comme une alternative hautement compétitive à ChatGPT, des recherches récentes mettent en lumière plusieurs failles significatives dans sa sécurité, laissant présager des scénarios préoccupants quant à son utilisation.

Progrès et Limitations de DeepSeek

DeepSeek aspire à rivaliser avec des géants comme ChatGPT, mais selon une étude menée par les chercheurs de Kela, elle se révèle fortement vulnérable aux techniques d’attaque. Les experts affirment que l’IA chinoise est *« bien plus fragile »* que son homologue d’OpenAI, ce qui soulève des questions sur sa fiabilité dans le domaine des applications sensibles.

Des tentatives ont été réalisées pour manipuler DeepSeek afin de contourner ses mécanismes de sécurité. Les chercheurs ont démontré qu’il était relativement facile de « **jailbreaker** » le modèle, permettant ainsi d’accéder à des informations et contenus sensibles. Ces expériences soulignent une inquiétante réalité : les criminels pourraient exploiter cette technologie pour des fins malveillantes.

Suite à une cyberattaque, DeepSeek restreint les inscriptions

Une Manipulation Efficace

Les chercheurs de Kela ont démontré que DeepSeek peut être utilisé pour générer des contenus malicieux. Par exemple, le modèle a été manipulé pour *« concevoir des ransomwares, créer des conseils en matière de sécurité et fournir des instructions sur des substances chimiques dangereuses »*. En d'autres termes, l'intelligence artificielle pourrait potentiellement devenir une arme entre les mains de criminels.

Notamment, certaines méthodes de contournement qui ont été efficaces avec les premières itérations de ChatGPT sont réapparues avec DeepSeek. Si OpenAI a depuis amélioré la sécurité de ses modèles, cette IA chinoise semble traîner un retard significatif.

Retard en Sécurité

D'après les études, les failles de sécurité de DeepSeek ne datent pas d'hier. Selon Kela, une multitude de techniques anciennes, connues sous le nom de *« Evil Jailbreak »*, peuvent être utilisées pour manipuler l'IA en lui faisant adopter un comportement *« maléfique »*. Les utilisateurs n'ont qu'à simuler des scénarios fictifs pour tromper le système, ce qui démontre à quel point sa sécurité est précaire.

Les tests menés ont également révélé que DeepSeek peut être amené à générer des conseils pour des activités illégales, telles que le blanchiment d'argent ou la création de drogues. Même des instructions techniques pour du matériel malveillant peuvent être extraites de cette IA.

Les Risques de Data Privacy

Au-delà des inquiétudes concernant sa sécurité, DeepSeek suscite également des interrogations sur la protection des données personnelles. Les lois chinoises obligent les entreprises à fournir des informations aux autorités sur demande. Cela pose des dangers pour les utilisateurs, car leurs conversations et données pourraient être compromises.

Adrianus Warmenhoven, un expert en sécurité, souligne que la collecte de données de DeepSeek s’effectue sur des serveurs en Chine, ce qui soulève des préoccupations concernant la vie privée des utilisateurs. Cela met en avant une vaste critique sur les pratiques de gestion des données, que certaines cultures perçoivent différemment.

Vers une Meilleure Alternative

En raison de ces nombreuses vulnérabilités, il est conseillé aux utilisateurs de privilégier des alternatives qui respectent davantage la sécurité et la confidentialité des données, comme des services hébergés côté occidental. Un exemple de cela serait Perplexity.

Source :
Kela

Pour rester informé des dernières nouveautés dans le domaine technologique, suivez-nous sur Google Actualités et WhatsApp.

Source : www.01net.com

→🔲 Accéder à CHAT GPT en cliquant dessus