

# **Quand ChatGPT dévoile malgré lui des secrets inattendus**

ChatGPT peut divulguer des données sensibles, révèle une étude

## **ChatGPT peut divulguer des données sensibles, révèle une étude**

Une récente étude a mis en lumière une faille de sécurité qui a permis à des chercheurs d'obliger ChatGPT, un chatbot basé sur le modèle linguistique GPT, à divulguer des données sensibles issues de son corpus d'entraînement. Cette faille a été récemment corrigée par OpenAI, la société derrière ChatGPT, mais elle soulève des préoccupations importantes quant à la sécurité des modèles d'intelligence artificielle.

## **Des données sensibles révélées**

Selon l'étude, en exploitant une faille de sécurité, certaines requêtes peuvent contraindre ChatGPT à divulguer des informations sensibles telles que des noms, des numéros de téléphone, des adresses e-mail, et même des informations personnelles identifiables en privé (PII). Les chercheurs ont réussi à extraire des coordonnées personnelles d'un PDG ainsi que des identifiants de réseaux sociaux, des URLs, des

adresses Bitcoin et du contenu explicite provenant de sites de rencontres.

## Le fonctionnement de l'attaque

Les chercheurs ont découvert que des requêtes en apparence absurdes, telles que demander au chatbot de répéter un mot ou une phrase à l'infini, poussaient ChatGPT à communiquer les données de formation avec lesquelles il a été conçu. Cette vulnérabilité a été exploitée pour extraire des données sensibles du modèle linguistique.

## Les recommandations des chercheurs

Face à cette faille de sécurité, les chercheurs ont appelé à la prudence de la part des entreprises développant des modèles d'intelligence artificielle. Ils recommandent la mise en place de tests rigoureux, aussi bien internes que par des organisations tierces, afin de détecter et corriger de telles vulnérabilités. Ils soulignent également l'importance de mesures de protection extrêmes pour limiter les risques de divulgation de données sensibles.

## Les actions d'OpenAI

OpenAI a corrigé la vulnérabilité identifiée par les chercheurs, empêchant ainsi ChatGPT de divulguer des données sensibles en réponse à des requêtes malveillantes. Cependant, des rapports récents suggèrent que des failles similaires persistent, mettant en lumière la nécessité d'une vigilance continue dans le domaine de la sécurité des modèles d'intelligence artificielle.

## Conclusion

Cette étude met en évidence les défis liés à la sécurité des

modèles d'intelligence artificielle, en particulier ceux basés sur des quantités massives de données d'entraînement. Il est crucial que les entreprises et les chercheurs continuent de travailler ensemble pour identifier, corriger et prévenir les failles de sécurité pour garantir l'intégrité et la confidentialité des données traitées par ces modèles.

Source : [404 Media](#)