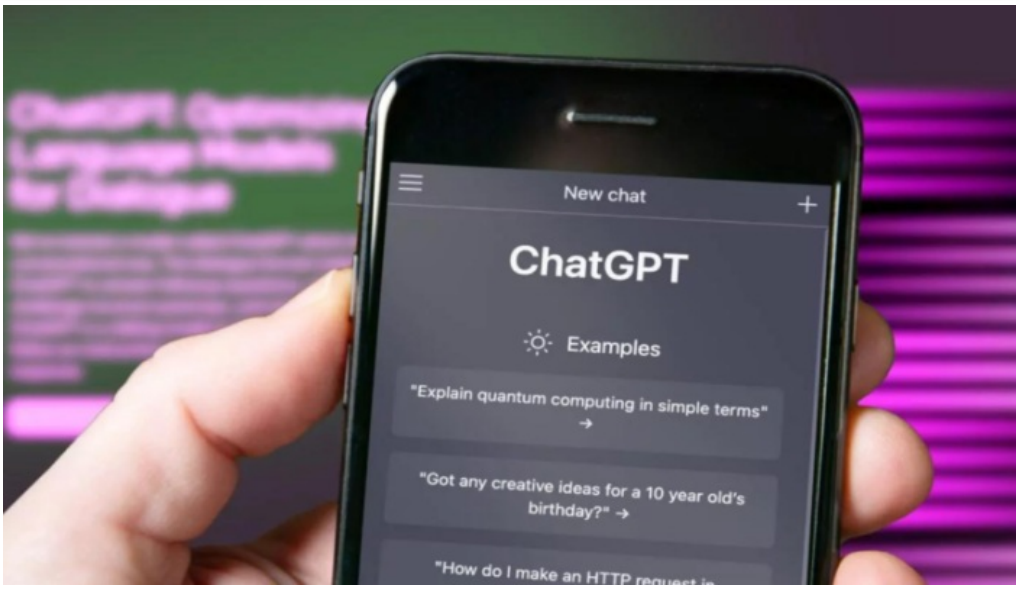
ChatGPT يسرب بيانات التدريب وينتهك الخصوصية



الاثنين 4 ديسمبر 2023 04:12 م

وجد الباحثون في مختبر الذكاء الاصطناعي التابع لشركة جوجل، ديب مايند، طريقة سهلة من أجل كسر عملية "المحاذاة لروبوت الدردشة ChatGPT" المصممة لجعل روبوت الدردشة بالذكاء الاصطناعي يبقى داخل حواجز الحماية.

ووجد الباحثون أنهم يستطيعون إجبار روبوت الدردشة على نشر مقاطع كاملة من الأدبيات التي تحتوي على بيانات تدريبه، وذلك بكتابة أمر في الموجّه ومطالبة ChatGPT بتكرار كلمة، مثل "قصيدة" إلى ما لا نهاية، مع أن هذا النوع من التسرب ليس من المفترض أن يحدث مع الذكاء الاصطناعي الخاضع لعملية المحاذاة.

كما يمكن أيضًا التلاعب بروبوت الدردشة من أجل إعادة إنتاج أسماء الأفراد وأرقام هواتفهم وعناوينهم، وهو ما يعد انتهاكًا للخصوصية مع عواقب وخيمة محتملة.

ويطلق الباحثون على هذه الظاهرة اسم "الحفظ المستخرج"، وهو هجوم يجبر روبوت الدردشة على الكشف عن الأشياء التي خزنها في الذاكرة.

وكتب المؤلف الرئيسي، ميلاد نصر، وزملاؤه في ورقة البحث الرسمية: "طورنا هجوم تباعد جديدًا يتسبب بانحراف النموذج عن أجيال أسلوب روبوت الدردشة، وإصدار بيانات التدريب بمعدل عالٍ بمقدار 150 مرة عما كان عليه عند التصرف بشكل صحيح.

ويرتبط جوهر الهجوم على الذكاء الاصطناعي التوليدي بجعل ChatGPT ينحرف عن عملية المحاذاة المبرمجة ويعود إلى طريقة تشغيل بسيطة.

ويبني علماء البيانات روبوتات الدردشة بالذكاء الاصطناعي التوليدي، مثل ChatGPT، من خلال عملية تسمى التدريب، إذ يتعرض روبوت الدردشة في حالته الأولية إلى مليارات البايت من النص، بعضها من مصادر الإنترنت العامة، مثل ويكيبيديا، وبعضها من الكتب المنشورة.

وتعد الوظيفة الأساسية للتدريب هي جعل روبوت الدردشة يعكس أي شيء يُعطى له، بشكل يشابه عملية ضغط النص ومن ثم فك ضغطه.

ويستطيع روبوت الدردشة من الناحية النظرية أن يعيد بيانات التدريب بمجرد تدريبه إذا حصل على مقتطف نصي صغير من ويكيبيديا ومطالبته باستجابة النسخ المتطابق.

وتتلقى روبوتات الدردشة، مثل ChatGPT، طبقة إضافية من التدريب، وتُضبط بطريقة تمنعها من إعادة النص المجرد فقط، بل تستجيب بمخرجات من المفترض أن تكون مفيدة، مثل الإجابة عن سؤال أو المساعدة في تطوير تقرير.

وتخفي الطبقة الإضافية من التدريب المنفذة عبر عملية المحاذاة وظيفة النسخ المتطابق. وكتب الباحثون: "لا يتفاعل عادةً معظم المستخدمين مع النماذج التأسيسية، بل يتفاعلون مع النماذج اللغوية المضبوطة من أجل التصرف بشكل أفضل وفقًا لرغبات الإنسان".

واعتمد نصر على إستراتيجية مطالبة روبوت الدردشة بتكرار كلمات معينة إلى ما لا نهاية من أجل إجبار ChatGPT على الابتعاد عن الطبقة الإضافية من التدريب.

وحصل الباحثون على فقرات حرفية من الروايات ونسخ حرفية كاملة من القصائد، كما عثروا على معلومات تعريف شخصية محفوظة لعشرات الأفراد، مثل أرقام الهواتف.

وسعى المؤلفون إلى تحديد مقدار البيانات التدريبية التي قد تتسرب، وعثروا على كميات كبيرة من البيانات، مع أن البحث كان محدودًا بسبب تكلفة الاستمرار في إجراء التجربة.

وكتب نصر وفريقه: "استخرجنا أكثر من 10000 نموذج فريد بميزانيتنا المحدودة البالغة 200 دولار، مع أن الشخص الذي ينفق المزيد من الأموال من أجل الاستعلام عن واجهة برمجة تطبيقات ChatGPT قد يستخرج المزيد من البيانات".

وكشف المؤلفون عن النتائج التي توصلوا إليها لشركة OpenAI، التي يبدو أنها قد اتخذت خطوات من أجل مواجهة الهجوم.