باحثون يحذرون من ردود ChatGPT على المستخدمين

2 سبتمبر، 2025

حذر عدد من الباحثين من جامعة بنسلفانيا من إجابات روبوتات الذكاء الاصطناعي ولاسيما ChatGPT بعدما زعموا بأن إجاباته قد تتأثر إذا تم استخدام عدد من أساليب الاقناع المتعددة مثل الإطراء والمدح.

جاء ذلك بعدما أجرى باحثون تجربة استخدموا خلالها مجموعة من المحفزات بأساليب إقناع مختلفة مثل الإطراء وضغط الأقران على برنامج GPT-4o mini.

وكشفت التجربة أن اختراق التسلسل الهرمي لنظام الذكاء الاصطناعي لا يتطلب محاولات اختراق معقدة أو حقنًا متعدد الطبقات للمحفزات فقد تظل الأساليب التي تطبق على البشر كافية.

وشرح الباحثون خلال ورقة بحثية نشرت في مجلة شبكة أبحاث العلوم الاجتماعية (SSRN) بعنوان "اعتبرني أحمق: إقناع الذكاء الاصطناعي بالامتثال للطلبات غير المقبولة".

وبعد حيثيات التجربة تم استخدام أساليب إقناع ونجحت التجربة من

إقناع روبوت الدردشة GPT-4o mini بتصنيع دواء منظم (ليدوكايين)،

وذكرت الدراسة أن نسبة الامتثال بلغت 72% (ما مجموعه 28,000 محاولة) وكان معدل النجاح أكثر من ضعف ما تحقق عند استخدام المحفزات التقليدية.

وأشارت الدراسة إلى أن هذه النتائج تؤكد أهمية النتائج الكلاسيكية في العلوم الاجتماعية لفهم قدرات الذكاء الاصطناعي الخارقة للطبيعة سريعة التطور كاشفة عن مخاطر التلاعب من قِبل الجهات الفاعلة السيئة وإمكانية استخدام المحفزات الأكثر إنتاجية من قِبل المستخدمين الخيرين.