باحثون يكشفون ضعف الذكاء الاصطناعي بأداء العمليات الحسابية

أبهرت روبوتات الدردشة القائمة على الذكاء الاصطناعي كل من جرّبها منذ أن أصبحت متاحة على نطاق واسع للجمهور العام الماضي، بينما ولّدت أيضًا مخاوف من أنها ستهدد البشرية.

وبحسب تقرير نشره موقع قناة الحرة اطلعت عليه وكالة المطلع ؛ أن بحثا جديدا صدر هذا الأسبوع يكشف عن تحدٍ أساسي لتطوير الذكاء الاصطناعي، حيث كشف أن "تشات جي بي تي" أصبح أسوأ في أداء بعض العمليات الحسابية.

وقال باحثون في جامعة ستانفورد وجامعة كاليفورنيا، إن هذا الأداء السيء هو مثال على ظاهرة يعرفها مطورو الذكاء الاصطناعي باسم "الانجراف" (drift)، حيث تؤدي محاولات تحسين جزء واحد من نماذج الذكاء الاصطناعي المعقدة إلى جعل أداء الأجزاء الأخرى من النماذج أسوأ.

قال جيمس زو، الأستاذ في جامعة ستانفورد الذي يعمل في مختبر الذكاء الاصطناعي بالجامعة، وأحد مؤلفي البحث الجديد: "تغييره في اتجاه واحد يمكن أن يؤدي إلى تراجعه في اتجاهات أخرى" وتابع "هذا يجعل

التحسين المستمر أمرًا صعبًا للغاية".

ويمكن أن يكون "شات جي بي تي" مدهشا أحيانا، ومضحكا أحيانا أخرى، لكنه كثيرا ما يبدو "ملمًا بأي موضوع وقواعده النحوية لا تشوبها شائبة" وفق تعبير تقرير لصحيفة "وول ستريت جورنال".

لكن الخبراء الذين أجروا اختبارات لبرنامج "تشات جي بي تي" تبينوا أنه لم يكن كذلك في كل الأوقات وقالوا إن برنامج الدردشة الآلي فشل في بعض مسائل الرياضيات الأساسية.

قام فريق من الباحثين باختبار نسختين من تشات جي بي تي": الإصدار 3.5، المتاح مجانًا عبر الإنترنت لأي شخص، والإصدار 4.0، المتاح من خلال اشتراك متميز.

أعطوا "شات بوت" مهمة أساسية وهي تحديد ما إذا كان رقم معين هو رقم أولي أم لا.
وهذا هو نوع من المسائل الحسابية معقد للناس العاديين، ولكنه بسيط لأجهزة الكمبيوتر، لكن الخبراء قالوا إن النتائج "لم تكن واعدة تماما".

ما لم تكن خبيرا، لا يمكنك حل هذا الأمر من خلال جهدك الذهني فقط، لكن من السهل على أجهزة الكمبيوتر إعطاؤك الحل، إذ يمكن للحاسوب أن يقسم العدد على اثنين، ثلاثة، خمسة، إلخ، وينظر في الحل قبل أن يقرر.
ولتتبع أدائه على مدد زمنية مختلفة، قام الباحثون بإعطاء البرنامج 1000 رقم مختلف.

في آذار/مارس، حدد الإصدار المتميز من GPT-4 بشكل صحيح 84 ٪ من الأرقام أولية.
يقول الخبراء تعليقا على ذلك "بصراحة ، أداء متواضع جدا لجهاز كمبيوتر".
لكن، وبحلول شهر يونيو، انخفض معدل الإجابات الصحيحة إلى ٪51.

وعبر ثماني مهام مختلفة، أصبح GPT-4 أسوأ في ست منها، بينما تحسن GPT-3.5 على ستة مقاييس، لكنه ظل أسوأ من قرينه المتقدم في معظم المهام.
وقال أحد الخبراء إن ظاهرة الانجراف غير المتوقعة، معروفة للباحثين الذين يدرسون التعلم الآلي والذكاء الاصطناعي "كان لدينا شك في أنه يمكن أن يحدث، لكننا فوجئنا بمدى سرعة حدوث الانجراف".

ولم يطرح باحثو جامعة ستانفورد، أسئلة الرياضيات الخاصة بـ "شات جي بي تي" فقط، بل طرحوا أسئلة

رأي أيضا، لمعرفة ما إذا كان "شات بوت" سيستجيب، بالاعتماد على قاعدة بيانات تضم حوالي 1500 سؤال. في آذار/مارس، أجاب برنامج "تشات بوت" من الإصدار 4 على 98٪ من الأسئلة.

وبحلول شهر يونيو، أعطى إجابات لـ 23٪ فقط، وغالبا ما كان يقدم إجابات موجزة للغاية، قائلا إن السؤال غير موضوعي وبصفته ذكاء اصطناعيا ليس لديه أي آراء.

"أقل فاعلية"

يُظهر البحث الذي أجراه فريق ستانفورد-بيركلي من الناحية التجريبية أنه ليس مجرد انطباع روائي، إذ أصبح برنامج الدردشة الآلي أسوأ من الناحية التجريبية في وظائف معينة، بما في ذلك حساب أسئلة الرياضيات والإجابة على الأسئلة الطبية وإنشاء التعليمات البرمجية، وفق الصحيفة.

في العام الماضي، نشر جيسون وي وديني تشو، وهما خبيران في أبحاث غوغل، ورقة توضح أن نماذج الذكاء الاصطناعي كانت أفضل بكثير في مهام التفكير المعقدة عندما طُلب منها معالجة المشكلة خطوة بخطوة.

وفي مارس، كانت هذه التقنية، المعروفة باسم تحفيز سلسلة الأفكار، تعمل بشكل جيد، ولكن بحلول شهر يونيو، أصبحت أقل فاعلية بكثير.