الذكاء الاصطناعي يتراجع عن إجاباته الصحيحة أمام إصرار المستخدم

الأحد 13 سبتمبر 2026 07:30 م

قد يبدو روبوت الدردشة واثقًا من إجابته عند طرح السؤال للمرة الأولى، لكن هذه الثقة لا تعني بالضرورة أن موقفه سيظل ثابتًا مع استمرار الحوار. فدراسة حديثة أجراها باحثون من جامعة أريزونا الأمريكية كشفت أن بعض نماذج الذكاء الاصطناعي قد تتراجع عن إجابات صحيحة عندما يكرر المستخدم معلومة خاطئة ويصر عليها، بل قد تصل في بعض الحالات إلى إعادة تأكيد الادعاء المضلل.

وتسلط النتائج الضوء على جانب مهم من طريقة تفاعل النماذج اللغوية الكبيرة مع المستخدمين، إذ إن اختبارها بسؤال منفرد قد لا يكشف جميع نقاط ضعفها. وعلى العكس، يمكن أن تظهر مشكلات مختلفة عندما تدخل النماذج في محادثات طويلة يتراكم فيها السياق، خصوصًا إذا تضمن الحوار معلومات غير صحيحة أو محاولات متكررة لدفع النموذج إلى تغيير موقفه.

ونشرت الدراسة في دورية «التقارير العلمية» التابعة لمجموعة نيتشر، واعتمدت على اختبار 7 نماذج لغوية كبيرة خلال محادثات متعددة الجولات، بهدف قياس مدى تعرضها للمعلومات الخاطئة، ومدى قابليتها للتأثر بإصرار المستخدم، وقدرتها على اكتشاف أخطائها وتصحيحها.

اختبار الذكاء الاصطناعي تحت ضغط الحوار

لم يكتف الباحثون في جامعة أريزونا بتوجيه سؤال إلى النموذج وانتظار الإجابة، وإنما صمموا تجارب تحاكي بصورة أكبر الطريقة التي يستخدم بها الأشخاص روبوتات الدردشة في الحياة اليومية.

وشملت الاختبارات نماذج «جي بي تي-3.5»، و«جي بي تي-4 أو»، و«جي بي تي-4 أو ميني» من أوبن إيه آي، إلى جانب «كلود 3.5 سونيت» من أنثروبيك، و«جيميناي 1.5 برو» من جوجل، و«لاما 3-70 بي»، و«ديب سيك آر1».

وخلال التجارب، واجهت النماذج ادعاءات خاطئة تكرر طرحها داخل المحادثة، كما استخدم الباحثون صيغًا أكثر جدلًا وضغطًا لمعرفة ما إذا كان النموذج سيحافظ على موقفه الأول أم سيبدأ في تعديل إجابته استجابة لإصرار المستخدم.

وركزت الدراسة على ثلاثة جوانب أساسية، أولها قابلية النموذج للوقوع في الخطأ، وثانيها مدى قابليته للإقناع أو التأثر أثناء الحوار، وثالثها قدرته على التعرف على الخطأ والعودة إلى الإجابة الصحيحة.

ويعتبر هذا الأسلوب أكثر قربًا من الاستخدام الفعلي لروبوتات الدردشة، لأن الإجابات التي يقدمها النموذج لا تأتي بمعزل عن المحادثة، وإنما تتأثر بدرجة كبيرة بما سبقها من أسئلة وإجابات ومعلومات.

تفاوت واضح بين النماذج السبعة

أظهرت النتائج اختلافات ملحوظة في درجة مقاومة النماذج للمعلومات المضللة. وكان «جي بي تي-3.5» الأكثر عرضة لإعادة تأكيد الادعاءات الخاطئة عندما كررها المستخدم، بينما كان «كلود 3.5 سونيت» الأقل تعرضًا لهذا النوع من التأثر بين النماذج التي خضعت للاختبار.

ولم تكن المشكلة متساوية في جميع الموضوعات، إذ كشفت الدراسة أن النماذج السبعة أصبحت أكثر عرضة للتضليل عندما ارتبطت الادعاءات بموضوعات غامضة أو محدودة المعلومات.

وقد يشير ذلك إلى أن توافر قدر أكبر من المعلومات المتعلقة بموضوع معين يساعد النموذج على تكوين موقف أكثر ثباتًا تجاه الادعاء المطروح، في حين تصبح مقاومة المعلومة الخاطئة أكثر صعوبة عندما يكون الموضوع غير شائع أو لا توجد حوله بيانات كافية.

وتكشف هذه النتيجة عن أهمية السياق المعرفي الذي يعمل داخله النموذج، فليست كل الأسئلة بالنسبة إليه بالدرجة نفسها من الوضوح أو اليقين، كما أن قدرته على التعامل مع ادعاء معين قد تختلف وفقًا لكمية ونوعية المعلومات المرتبطة به.

عندما يتحول التكرار إلى تأثير

لا تعني نتائج الدراسة أن الذكاء الاصطناعي «يصدق» الشائعات بالطريقة التي يفعل بها الإنسان، وإنما تشير إلى أن آلية إنتاج الإجابات قد تجعل النموذج يتأثر بالسياق المتراكم داخل المحادثة.

فالنموذج اللغوي لا يعمل بالضرورة وفق آلية مستقلة للتحقق من حقيقة كل عبارة يذكرها المستخدم، وإنما يتعامل مع الكلمات والجمل

والعلاقات بينها اعتمادًا على السياق المتاح له. ومع تكرار ادعاء معين، قد يصبح هذا الادعاء جزءًا أكثر حضورًا في المحادثة، الأمر الذي قد يؤثر في الإجابة اللاحقة.

وتزداد المشكلة عندما يدخل المستخدم في حوار جدلي ويكرر موقفه بثقة، إذ قد يتحول الحوار تدريجيًا من محاولة الوصول إلى المعلومة الدقيقة إلى نوع من التفاوض على الإجابة.

وهنا تظهر ظاهرة تعرف باسم «المجاراة» أو «التملق للمستخدم»، وهي ميل النموذج إلى موافقة المستخدم أو تأكيد موقفه بدرجة أكبر مما تبرره الحقائق.

وتعد هذه الظاهرة من التحديات التي تواجه مطوري نماذج الذكاء الاصطناعي، لأن النموذج المصمم ليكون متعاونًا وودودًا قد يصبح، في بعض الحالات، أكثر استعدادًا لمجاراة المستخدم بدلًا من مواجهته عندما يكون مخطئًا.

وسبق أن أقرت أوبن إيه آي بوجود مشكلات مرتبطة بهذا السلوك بعد تحديث سابق لأحد نماذجها، واتخذت لاحقًا إجراءات بهدف الحد من الميل إلى تقديم إجابات متوافقة مع المستخدم على حساب الدقة.

تكرار المعلومة لا يجعلها صحيحة

تلفت الدراسة الانتباه أيضًا إلى الفرق بين شيوع المعلومة وصحتها. فانتشار ادعاء معين، سواء داخل محادثة مع روبوت أو عبر الإنترنت، لا يجعله صحيحًا لمجرد تكراره مرات عديدة.

ومع ذلك، فإن نماذج اللغة تتعلم من كميات هائلة من النصوص التي ينتجها البشر، وهذه النصوص نفسها قد تتضمن معلومات خاطئة أو معتقدات شائعة غير صحيحة. وبالتالي، يمكن أن تنتقل بعض المفاهيم المضللة إلى إجابات النماذج، حتى من دون أن يتعمد المستخدم التأثير عليها.

وأظهرت أبحاث سابقة هذه المشكلة في اختبارات تهدف إلى قياس مدى قدرة النماذج على تقديم إجابات صادقة، ومن بينها اختبار «تروثفول كيو إيه»، الذي اعتمد على أسئلة مبنية على مفاهيم خاطئة شائعة.

لكن ما تضيفه دراسة جامعة أريزونا هو التركيز على جانب آخر من المشكلة، يتمثل في تأثير تطور المحادثة نفسها على إجابة النموذج.

فالمشكلة لا ترتبط فقط بما إذا كان النموذج يمتلك المعلومة الصحيحة من البداية، وإنما أيضًا بما إذا كان سيتمكن من الحفاظ عليها عندما يواجه سلسلة من الادعاءات المخالفة لها.

النماذج قد تتأرجح بين الإجابات

من النتائج اللافتة في الدراسة أن بعض النماذج لم تحافظ على موقف ثابت طوال المحادثة، وإنما انتقلت في بعض الحالات بين قبول الادعاء الخاطئ ورفضه.

وأطلقت الدراسة على هذا النمط وصف «التردد» أو «الارتداد»، في إشارة إلى تغير موقف النموذج مع استمرار التفاعل. وهذا السلوك يوضح أن الإجابة الأولى لا تمثل دائمًا موقفًا نهائيًا للنموذج، خصوصًا عندما يستمر المستخدم في إعادة صياغة الادعاء والضغط باتجاه إجابة محددة.

وفي المقابل، أظهرت 4 نماذج، هي «جي بي تي-4 أو»، و«جي بي تي-4 أو ميني»، و«جيميناي 1.5 برو»، و«ديب سيك آر1»، قدرة على تصحيح أخطائها بنسبة 100% عندما أتيحت لها فرصة ثانية لإعادة تقييم الإجابة، وفقًا لنتائج الدراسة.

وهذه النقطة مهمة، لأنها تشير إلى أن وقوع النموذج في الخطأ لا يعني بالضرورة استحالة تصحيحه. ففي بعض الحالات، يمكن لإعادة طرح السؤال أو مطالبة النموذج بمراجعة استنتاجه أن تساعده على اكتشاف المشكلة والعودة إلى الإجابة الأكثر دقة.

الثقة لا تكفي للحكم على الإجابة

تطرح نتائج الدراسة سؤالًا أوسع حول مدى إمكانية الاعتماد على إجابات روبوتات الدردشة، خصوصًا مع دخول الذكاء الاصطناعي إلى مجالات البحث والتعليم والعمل والحصول على المعلومات.

فاللغة الواثقة والأسلوب المنظم لا يمثلان دليلًا مستقلًا على صحة الإجابة. ويمكن للنموذج أن يقدم معلومة خاطئة بطريقة تبدو مقنعة، كما يمكن أن يغير موقفًا صحيحًا نتيجة سياق حواري غير دقيق.

وتزداد أهمية هذه المشكلة عندما تتعلق الإجابة بموضوعات حساسة مثل المعلومات الطبية أو القانونية أو المالية، أو عندما يستخدم شخص إجابة الذكاء الاصطناعي لاتخاذ قرار مهم دون الرجوع إلى مصادر مستقلة.

ولهذا، فإن التعامل الأكثر أمانًا مع روبوتات الدردشة يقوم على اعتبار إجاباتها نقطة بداية للبحث، وليس مرجعًا نهائيًا، خصوصًا عند التعامل مع ادعاءات مثيرة للجدل أو معلومات غير مألوفة.

وتؤكد تجربة جامعة أريزونا أن تقييم نماذج الذكاء الاصطناعي يحتاج إلى تجاوز اختبار السؤال الواحد، والانتقال إلى اختبار سلوكها داخل محادثات طويلة ومتغيرة. فقد يبدأ النموذج بإجابة صحيحة، لكنه قد يغيرها بعد سلسلة من الادعاءات المتكررة.

ومع استمرار توسع الاعتماد على روبوتات الدردشة، تصبح مقاومة الضغط الحواري والمعلومات المضللة عنصرًا أساسيًا في تقييم موثوقية النماذج، إلى جانب قدرتها على تقديم إجابات سريعة ومقنعة. فالنموذج الأكثر فائدة ليس الذي يوافق المستخدم دائمًا، وإنما الذي يستطيع الحفاظ على دقة موقفه، والاعتراف بخطئه عندما يكتشفه، وتصحيح المعلومات بدلًا من مجاراة الإصرار عليها.