Los nuevos modelos de IA como ChatGPT que no logran ser fiables

ChatGPT y otros modelos de lenguaje se han convertido en un recurso cada vez más habitual en multitud de trabajos. Sin embargo, tienen un problema de fondo que tiende a empeorar: estos sistemas dan a menudo respuestas incorrectas y la tendencia no es positiva. “Los sistemas nuevos mejoran sus resultados en tareas difíciles, pero no en fáciles, así que estos modelos se vuelven menos fiables”, resume Lexin Zhou, coautor de un artículo que este miércoles publica la revista científica Nature, que ha escrito junto a cuatro españoles y un belga del Instituto VRAIN (Instituto Universitario Valenciano de Investigación en Inteligencia Artificial) de la Universitat Politécnica de València y de la Universidad de Cambridge. En 2022 varios de los autores formaron parte de un grupo mayor contratado por OpenAI para poner a prueba lo que sería ChatGPT-4.

El artículo ha estado un año en revisión antes de ser publicado, un periodo común para este tipo de trabajos científicos; pero ya fuera del estudio, los investigadores han probado también si los nuevos modelos de ChatGPT o Claude resuelven estos problemas y han comprobado que no: “Hemos encontrado lo mismo”, zanja Zhou. “Hay algo incluso peor. ChatGPT-o1 [el programa más reciente de OpenAI] no evita tareas y si le das un puñado de cuestiones muy difíciles no dice que no sabe, sino que emplea 100 o 200 segundos pensando la solución, lo que es muy costoso en términos computacionales y de tiempo para el usuario”, añade.

Para un humano no es sencillo detectar cuándo uno de estos modelos puede estar equivocándose: “Los modelos pueden resolver tareas complejas, pero al mismo tiempo fallan en tareas simples”, dice José Hernández-Orallo, investigador de la UPV y otro de los autores. “Por ejemplo, pueden resolver varios problemas matemáticos de nivel de doctorado, pero se pueden equivocar en una simple suma”, añade.

Este problema no será fácil de resolver porque la dificultad de los retos que los humanos pongan a estas máquinas será cada vez más difícil: “Esa discordancia en expectativas humanas de dificultad y los errores en los sistemas, empeorará. La gente cada vez pondrá metas más difíciles para estos modelos y prestará menos atención a las tareas más sencillas. Eso seguirá

así si los sistemas no se diseñan de otra manera”, dice Zhou.

Estos programas evitan cada vez más disculparse por no saber algo. Esa seguridad irreal hace que los humanos se decepcionen más cuando la respuesta acaba siendo errónea. El artículo prueba que los humanos creen a menudo que son correctos resultados incorrectos que ofrecen en tareas difíciles. Esta aparente confianza ciega, unido a que los nuevos modelos tienden a responder siempre, no da muchas esperanzas para el futuro, según los autores.

“Los modelos de lenguaje especializados en áreas sensibles como la medicina podrían diseñarse con opciones de rechazo” a responder, dice el artículo, o colaborar con supervisores humanos para que sepan mejor cuándo deben abstenerse de responder. “Hasta que esto se logre, y dado el alto uso de estos modelos en la población general, llamamos a la concienciación sobre el riesgo que supone depender de la supervisión humana para estos sistemas, especialmente en áreas donde la verdad es crítica”, escriben en el artículo.

En opinión de Pablo Haya, investigador del Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid, en declaraciones a SMC España, el trabajo sirve para entender mejor el alcance de estos modelos: “Desafía la suposición de que escalar y ajustar estos modelos siempre mejora su precisión y alineación”. Y añade: “Por un lado, observan que, aunque los modelos más grandes y ajustados tienden a ser más estables y a proporcionar respuestas más correctas, también son más propensos a cometer errores graves que pasan desapercibidos, ya que evitan no responder. Por otro lado, identifican un fenómeno que denominan ‘discordancia de la dificultad’ y que revela que, incluso en los modelos más avanzados, los errores pueden aparecer en cualquier tipo de tarea, sin importar su dificultad”.

Un remedio casero

Un remedio casero para resolver estos errores, según el artículo, es adaptar el texto de la petición: “Si le preguntas varias veces va a mejorar”, dice Zhou. Este método implica cargar al usuario con la responsabilidad de acertar con su pregunta o adivinar si la respuesta es correcta. Cambios sutiles en el prompt (petición) como “¿podrías responder?”, en lugar de

“por favor, responde a lo siguiente”, dará diferentes niveles de precisión. Pero el mismo tipo de pregunta puede funcionar para tareas difíciles y mal para sencillas. Es un juego de prueba y acierta.

El gran problema futuro de estos modelos es que su presunto objetivo es lograr una superinteligencia capaz de resolver problemas que los humanos son incapaces de asumir por falta de capacidad. Pero, según los autores de este artículo, ese camino no tiene salida: “El modelo actual no nos llevará a una IA superpoderosa que pueda solucionar la mayoría de tareas de una manera fiable”, afirma Zhou.

Ilya Sutskever, cofundador de OpenAI y uno de los científicos más influyentes del sector, acaba de fundar una nueva compañía. En declaraciones a Reuters admitió algo parecido: este camino está agotado. “Hemos identificado una montaña que es un poco diferente de lo que estaba trabajando, una vez subas a su cima, el modelo cambiará y todo lo que sabemos sobre la IA cambiará una vez más”, dijo. Zhou está de acuerdo: “En cierto modo apoya nuestros argumentos. Sutskever ve que el modelo actual no basta y busca nuevas soluciones”.

Estos problemas no implican que estos modelos no sirvan para nada. Los textos o ideas que proponen sin que haya detrás una verdad fundamental siguen siendo válidos. Aunque cada usuario deberá asumir su riesgo: “Yo no me fiaría por ejemplo del resumen de un libro de 300 páginas”, explica Zhou. “Seguro que hay un montón de información útil, pero no me fiaría al 100%. Estos sistemas no son deterministas, sino aleatorios. En esa aleatoriedad pueden incluir algún contenido que se desvíe del original. Es preocupante”, añade.

Este verano se hizo célebre entre la comunidad científica el caso de una investigadora española que mandó al Comité Europeo de Protección de Datos una página de referencias llena de nombres y enlaces alucinados en un documento precisamente sobre auditar la IA. Ha habido otros casos en tribunales y por supuesto multitud que han pasado sin ser detectados.

Por El País, es