Divulga UAB

Revista de difusió de la recerca de la Universitat Autònoma de Barcelona

29/01/2025

Tu asistente virtual favorito te responde, ¿pero realmente te entiende?

Las conversaciones con asistentes virtuales son cada vez más frecuentes y se incorporan incluso en empresas privadas. Mucha gente piensa que las respuestas son lo bastante convincentes para parecer humanas. Pero ¿desarrollan realmente el lenguaje como las personas? Un nuevo artículo examina la responsabilidad lingüística y el razonamiento en la creación de oraciones comparando a 400 personas y 7 modelos de lenguaje avanzados.

Imagen generada por GPT

Los Modelos de Lenguaje Extensos (“Large Language Models”, LLM) suponen uno de los avances tecnológicos más impresionantes que hemos presenciado en los últimos años. Se utilizan con éxito en varias aplicaciones que incluyen la predicción de la siguiente palabra, como autocompletar el texto y las respuestas automáticas a preguntas mediante agentes artificiales (asistentes virtuales o bots). Actualmente, estas habilidades lingüísticas se presentan como lo suficientemente convincentes para el ojo no entrenado que mucha gente argumenta que los resultados de estos modelos parecen una respuesta humana.

Pero ¿los LLM desarrollan realmente el lenguaje como las personas? ¿Tene esperar que lo hagan? Una compañía aérea fue requerida a pagarle los daños a un que recibió información incorrecta mientras mantenía una conversación con el bot de la

compañía. Según un representante de la empresa, el bot incluyó “palabras engañosas” en las respuestas de las preguntas del cliente. Finalmente, el juez emitió una sentencia de representación distorsionada negligente a favor del pasajero, aunque la compañía continuó argumentando que el bot **es** y **debería ser el responsable** de sus propias palabras. Entonces, los chats con bots, asistentes virtuales y otras aplicaciones interactivas relacionadas con estas tecnologías predictivas, ¿tienen una comprensión del lenguaje parecida a la humana o su habilidad está inherentemente limitada?

Para poder responder a esta pregunta, personal investigador de la Universidad Rovira i Virgili, la Universidad de Pavía, la Universidad Humboldt de Berlín, la Universidad de Nueva York, la Universidad Autónoma de Barcelona y la Institució Catalana de Recerca i Estudis Avançats (ICREA) han comparado 400 personas y 7 modelos de última generación en un nuevo punto de referencia que implica indicios lingüísticos muy simples. El objetivo era ofrecer a los modelos las mejores condiciones posibles para responder correctamente. La prueba involucraba procesar y responder oraciones como “John engañó a Mary y Lucy engañó a Mary. En este contexto, ¿Mary engañó a Lucy?”

Como era de esperar, las personas completaron la tarea con éxito. Los LLM, en cambio, presentaron mucha variedad en sus respuestas, de forma que algunos modelos respondieron mejor que otros. Así, los LLM como clase resultan peor que las personas. Más notable es el hecho que los modelos cometieron tipos de errores que eran completamente ausentes de las respuestas humanas. Por ejemplo, al enunciado “Franck se lee a sí mismo y John se lee a sí mismo, a Anthony y a Frank. En este contexto, ¿Franck fue leído?”, algunos de los modelos probados respondieron que “en este contexto, es imposible decir con seguridad quien leía a Franck” y que, para contestar, necesitaríamos saber “información adicional sobre la situación específica, como el material de lectura de John”.

El resumen de estos resultados por la autora principal de este estudio, la Prof. Evelina Leivada (UAB e ICREA), revela que cuando se rasca la superficie de un aparente buen rendimiento lingüístico, el rendimiento lingüístico de los LLM puede esconder defectos inherentes a la modelización del lenguaje como método. El mensaje principal es que la inteligencia, el razonamiento y el ancoraje de palabras en las condiciones del mundo real no puede emerger como un producto secundario de la inferencia estadística. Como remarca otro autor del estudio, el Prof. Gary Marcus, en su libro publicado el 2024: “Taming Silicon Valley. How We Can Ensure that AI Works for Us”, los sistemas de Inteligencia Artificial son indiferentes a la verdad que se esconde detrás de sus palabras, hecho que genera preocupaciones sobre la desinformación masiva, la difamación, la contaminación del mercado y la magnificación de los sesgos que se producen a gran escala.

Evelina Leivada
Personal investigador ICREA, Departament de Filología Catalana
Universitat Autònoma de Barcelona
Evelina.Leivada@uab.cat

Referencias

Dentella, V., Günther, F., Murphy, E., Marcus, G. & Leivada, E. **Testing AI on language comprehension tasks reveals insensitivity to underlying meaning.** *Scientific Reports* 14, 28083 (2024). https://doi.org/10.1038/s41598-024-79531-8

View low-bandwidth version