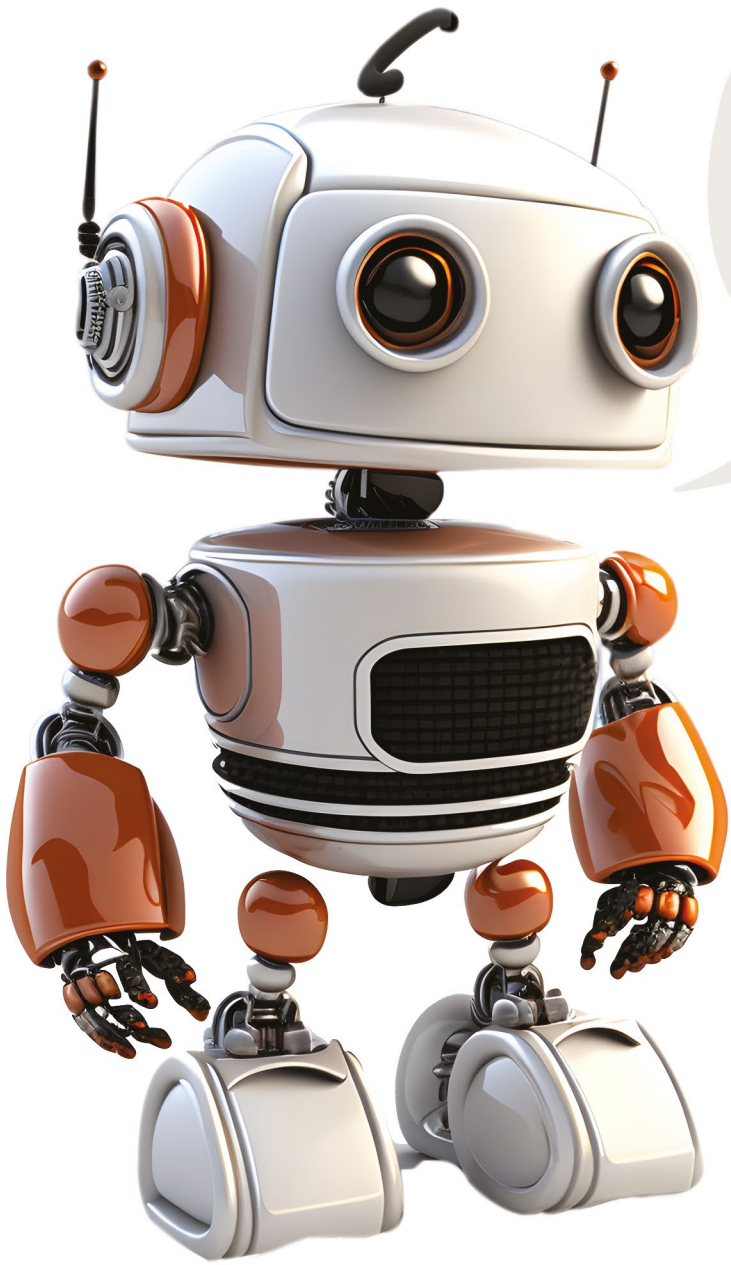
¿ChatGPT o humano?



En busca de la **marca de agua** de textos **generados por IA**

MIT Technology Review
Publicado por Opinno

MELISSA HEIKKILÄ
TRADUCIDO POR **ANA MILUTINOVIC**
14 FEBRERO, 2023

La herramienta podría permitir a los profesores descubrir plagios, o ayudar a combatir los bots de desinformación en redes sociales.

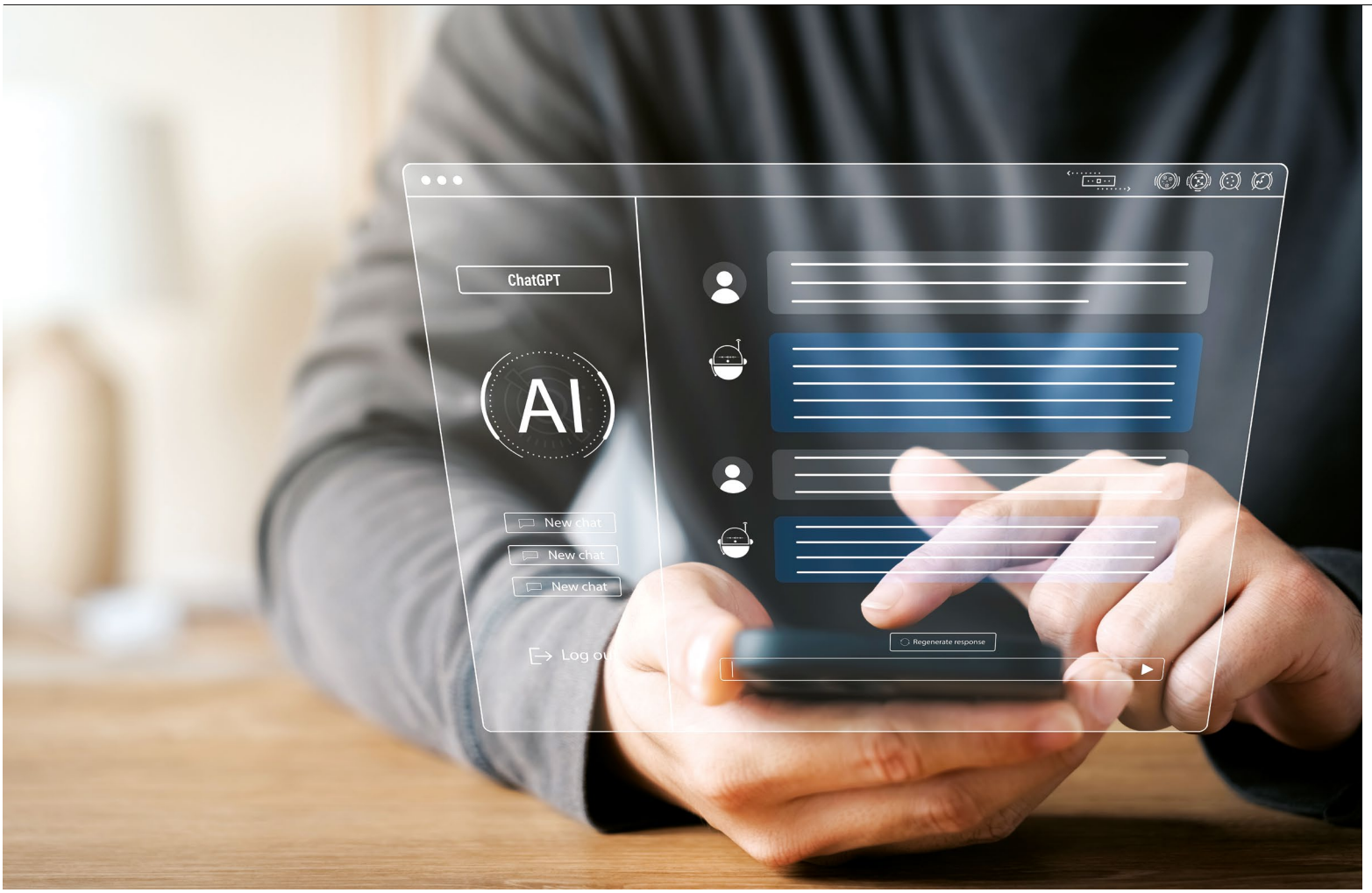
Los patrones ocultos escondidos a propósito en los textos generados por inteligencia artificial (IA) podrían ayudar a identificarlos, permitiéndonos saber si las palabras que estamos leyendo están escritas por un ser humano o no.

Estas marcas de agua son invisibles para el ojo humano, pero consiguen que los ordenadores detecten que un texto ha sido escrito por un sistema de IA. Si se integran en los grandes modelos de lenguaje, podrían ayudar a prevenir algunos problemas causados por estos modelos.

Por ejemplo, desde que OpenAI lanzó ChatGPT en noviembre del año pasado, algunos estudiantes ya han empezado a utilizarlo para escribir trabajos. La de noticias CNET empezó a utilizar ChatGPT para escribir artículos, aunque tuvo que publicar correcciones tras ser acusados de plagio. Esta técnica de marca de agua, introducida en dichos sistemas antes de su lanzamiento, podría ayudar a abordar tales problemas.

En algunos estudios, estas marcas de agua ya han sido utilizados para identificar textos generados por IA con una certeza casi total. Por ejemplo, los investigadores de la Universidad de Maryland (EE UU) lograron detectar texto redactado por OPT-6.7B, el modelo de lenguaje de código abierto creado por Meta, gracias a

Los modelos de lenguaje de IA funcionan al predecir y generar una palabra a la vez.



un algoritmo de detección que habían desarrollado. Su trabajo se describe en un artículo que aún no ha sido revisado por colegas, y el código estará disponible de forma gratuita para el 15 de febrero.

Los modelos de lenguaje de IA funcionan al predecir y generar una palabra a la vez. Después de cada palabra, este algoritmo de marca de agua divide el vocabulario del modelo de lenguaje en palabras entre una lista verde y una lista roja. Después, le pide al modelo que elija palabras en la lista verde.

Cuantas más palabras de la lista verde haya en un fragmento de texto, más probable es que el texto haya sido generado por una máquina. El texto escrito por una persona suele contener una combinación más aleatoria de palabras. Por ejemplo, para la palabra «bonita», el algoritmo de marca de agua podría clasificar la palabra «flor» como verde y «orquídea» como roja. Es más probable que el modelo de IA (con algoritmo de marca de agua) utilice la palabra «flor» y no «orquídea», explica Tom Goldstein, profesor asistente de la Universidad de Maryland, que participó en esta investigación.

La investigación básica debe escuchar a las futuras necesidades de la sociedad.

ChatGPT es solo un ejemplo de la nueva generación de grandes modelos de lenguaje, que generan textos tan fluidos que podrían confundirse con la redacción humana. Estos modelos de IA regurgitan datos con confianza, pero son conocidos por mostrar falsedades y sesgos. Para el ojo inexperto, puede ser casi imposible distinguir un fragmento escrito por un modelo de IA que uno escrito por una persona. La impresionante velocidad del desarrollo de la IA implica que los modelos nuevos y más potentes pronto harán que el conjunto de herramientas existente para detectar texto sintético sean cada vez menos eficaces. Es una carrera constante entre los desarrolladores de IA para crear nuevas herramientas de seguridad que puedan seguir la última generación de modelos de IA.

«Ahora mismo, esto es el salvaje Oeste», señala John Kirchenbauer, investigador de la Universidad de Maryland, que participó en el desarrollo de la marca de agua y espera que esta tecnología pueda dar una ventaja a los esfuerzos de detección de la IA. La herramienta desarrollada por su equipo podría ajustarse para funcionar con cualquier modelo de lenguaje de IA que prediga la siguiente palabra, destaca Kirchenbauer.

Los hallazgos son prometedores y oportunos, opina Irene Solaiman, directora de Políticas de Hugging Face, la *start-up* de IA., Solaiman trabajó en el estudio de la detección de los resultados de IA como investigadora de IA en OpenAI, pero no participó en esta investigación.

«A medida que los modelos se implementan a gran escala, más personas fuera de la comunidad de IA, sin capacitación en informática, tendrán que acceder a los métodos de detección», indica Solaiman.

Sin embargo, existen limitaciones en este nuevo método. Esta marca de agua solo funciona si sus creadores la integran desde el principio en el gran modelo de lenguaje. Aunque se dice que OpenAI está trabajando en métodos para detectar textos generados por IA, con marcas de agua incluidas, la investigación sigue siendo secreta. La empresa no suele dar mucha información a terceros sobre el funcionamiento de ChatGPT, o cómo fue entrenado, y mucho menos ofrece acceso para analizarlo. OpenAI no respondió a nuestra solicitud de comentarios.

Segñun Solaiman, tampoco está claro cómo se aplicará el nuevo trabajo a otros modelos, además del de Meta, como ChatGPT. El modelo de IA donde se probó la marca de agua también es más pequeño que los modelos populares, como ChatGPT.

Se necesitan más pruebas para explorar las diferentes formas en las que alguien podría luchar contra estos métodos de marca de agua, pero los investigadores afirman que las opciones son limitadas. «Tendría que cambiar la mitad de las palabras en un fragmento de texto antes de poder eliminar la marca de agua», resalta Goldstein.

«Es peligroso subestimar a los alumnos de secundaria, así que no lo haré. Pero, en general, la persona promedio no podrá manipular este tipo de marca de agua», concluye Solaiman. **</>**

«A medida que los modelos se implementan a gran escala, más personas fuera de la comunidad de IA, sin capacitación en informática, tendrán que acceder a los métodos de detección», indica Solaiman.