Adolfo Guzmán Arenas

ChatGPT, el nuevo y asombroso *chatbot* de inteligencia artificial

ChatGPT fue diseñado para producir lenguaje humano bastante natural; pregúntele lo que sea y recibirá una respuesta como si la hubiera escrito alguien más, algo parecido a tener una conversación. Esta novedosa herramienta de inteligencia artificial impresiona rápidamente, pero muchas personas han señalado que tiene algunas dificultades serias. En este artículo explico su funcionamiento, peligros y usos.

ChatGPT es un nuevo y poderoso *chatbot* de inteligencia artificial (IA) diseñado para producir lenguaje humano bastante natural. Pregúntele lo que quiera y recibirá una respuesta que suena como si la hubiera escrito un humano, ya que ha aprendido los conocimientos y habilidades de escritura de las personas al ser entrenado con cantidades masivas de datos en internet. Al igual que tener una conversación con alguien, usted puede hablar con ChatGPT, y éste recordará las cosas que usted le ha dicho en el pasado y, al mismo tiempo, podrá corregirse cuando lo desafíen. Por ello, ChatGPT impresiona rápidamente, pero muchas personas han señalado que tiene algunas dificultades serias.

¿Cómo trabaja ChatGPT?

ChatGPT funciona a partir de su "transformador preentrenado generativo" (GPT, por sus siglas en inglés), que es una red neuronal artificial gigantesca de aprendizaje mecánico que utiliza algoritmos especializados para encontrar patrones dentro de las secuencias de datos (Hetler, 2023). Ante una pregunta, el transformador extrae una cantidad significativa de datos para formular una respuesta. Se basa en el modelo de lenguaje GPT-4, lanzado el 14 de marzo de 2023 por OpenAI, como una versión mejorada de GPT-3.

OpenAI, una empresa de investigación en IA fundada en 2015, creó ChatGPT y lo lanzó en noviembre de 2022. OpenAI cuenta con el respaldo de varios inversionistas, entre los cuales destaca Microsoft. La misma empresa también creó Dall-E, una herramienta de IA que genera imágenes de arte a partir de descripciones de texto.

En este artículo hablaré tanto de ChatGPT como de GPT-3, y por extensión, de GPT-4, que ha sido entrenado con muchos más datos. GPT-2 apareció en 2019; este modelo usó 40 gigabytes (GB) de textos. Luego, GPT-3 se entrenó con 570 GB de datos tomados de Wikipedia, diversos sitios internet y otras fuentes. No obstante, el fabricante no ha compartido qué tan grande es el conjunto de datos usado para GPT-4 (Heikkilä, 2023).

ChatGPT utiliza el aprendizaje profundo, una forma de aprendizaje automático para producir texto similar al lenguaje humano a través de redes neuronales transformadoras. El transformador predice el texto, integrando la siguiente palabra, oración o párrafo, según la secuencia típica de sus datos de entrenamiento. La capacitación comienza con datos genéricos, luego pasa a datos más personalizados para una tarea específica. Así, ChatGPT se entrenó con texto en línea para aprender el lenguaje humano y luego usó transcripciones para aprender los conceptos básicos de las conversaciones.

Las personas que lo entrenaron constantemente le brindaron conversaciones y clasificaron las respuestas. De esta manera, ayudaron a determinar las mejores respuestas. Para seguir entrenando al *chatbot*, los usuarios pueden votar a favor o en contra de la respuesta haciendo clic en los íconos de "pulgar hacia arriba" o "pulgar hacia abajo" que se encuentran a un lado. Los usuarios también pueden escribir comentarios adicionales para mejorar y afinar el diálogo futuro.

En resumen, ChatGPT conoce la sintaxis: no genera frases como "las perro comen carne"; conoce la semántica: no genera frases como "los perros comen polinomios"; tiene buen sentido común: no genera frases como "los perros comen 100 kilos de carne diariamente". Mucho de lo que responde concuerda con la información que hay en internet, pero también inventa respuestas que parecen razonables, y revuelve lo cierto con sus mentiras ("alucinaciones").

Los peligros de ChatGPT

ChatGPT genera respuestas incorrectas, falla en matemáticas básicas, parece que no puede responder preguntas lógicas simples e incluso llega a argumentar hechos completamente incorrectos (Wu, 2023; Agomuoh, 2023). A diferencia de otros asistentes de IA, como Siri o Alexa, ChatGPT no utiliza internet para encontrar las respuestas. En su lugar, construye una oración palabra por palabra, seleccionando el "símbolo" más probable que debería venir a continuación, en función de su entrenamiento. En otras palabras, ChatGPT llega a una respuesta haciendo una serie de conjeturas, lo cual es parte de la razón por la cual puede argumentar respuestas incorrectas como si fueran completamente ciertas.

Si bien es excelente para explicar conceptos complejos, lo que lo convierte en una herramienta poderosa para el aprendizaje, es importante no creer todo lo que dice ChatGPT. Recuerde que la herramienta no siempre está en lo correcto, pues tiene un conocimiento limitado de los eventos mundiales después de 2021.

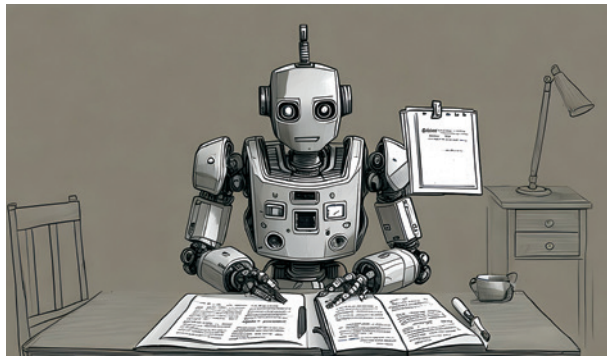
Hay demasiadas formas en que se puede abusar de estos sistemas, ya que se distribuyen gratuitamente y no existe una revisión o regulación para evitar los daños. Por lo tanto, han surgido muchas críticas y preocupaciones al respecto de su uso.

ChatGPT tiene un sesgo integrado en su sistema

Este *chatbot* se entrenó con la escritura colectiva de las personas en todo el mundo, del pasado y del presente. Por esta razón, los mismos sesgos que existen en el mundo real también pueden aparecer en el modelo de IA.

ChatGPT podría generar los ensayos que se les pide a estudiantes de secundaria

En realidad, cualquier estudiante puede pedirle a ChatGPT que revise su ensayo o le indique cómo



mejorar un párrafo, o bien puede abstenerse totalmente de redactarlo y pedirle a ChatGPT que escriba todo. Esta herramienta incluso genera respuestas mejores que las que muchos alumnos podrían hacer, como escribir cartas de presentación o describir temas importantes de una obra literaria famosa.

ChatGPT podría causar daños en el mundo real

La información incorrecta de ChatGPT puede causar daños en el mundo real, por ejemplo, al dar un consejo médico incorrecto. Pero también hay otras preocupaciones. La velocidad a la que puede generar un texto que suena natural hace que sea muy fácil para los estafadores hacerse pasar por alguien más en las redes sociales digitales. Del mismo modo, será difícil detectar un correo electrónico de *phishing*, diseñado para extraer detalles confidenciales. Además, ChatGPT produce texto sin errores gramaticales, lo que solía alertar de que el correo era mal intencionado.

ChatGPT puede difundir información falsa

La escala a la que ChatGPT puede producir texto, junto con la capacidad de hacer que incluso la información incorrecta suene convincentemente correcta, sin duda hará que la información en internet sea aún más cuestionable. Es decir, si muchos usuarios de ChatGPT deciden publicar en internet los resultados de sus consultas al *chatbot*, internet se contaminará aún más con mentiras y noticias falsas: antes sólo los humanos las emitíamos, ahora lo hace también ChatGPT.

¿Es ChatGPT un virus salvaje?

Paul Kedrosky (Loizos, 2022) compara el lanzamiento de ChatGPT con la liberación de un virus que tiene muchos beneficios, pero también peligros y daños ocultos. Además, este virus tiene la capacidad de mejorarse cuando es corregido por las personas, dado que va absorbiendo todo lo que se le da, aprendiendo cada vez más, avanzando sector por sector. "Parece que podría comerse el mundo" (Loizos, 2022).

OpenAI tiene todo el poder

Un gran poder conlleva una gran responsabilidad, y OpenAI tiene mucho poder (Wu, 2023). Ésta es una de las primeras empresas de IA en sacudir verdaderamente al mundo no con uno, sino con múltiples modelos de IA generativos, incluidos Dall-E 2, GPT-3 y GPT-4.

OpenAI elige qué datos se utilizan para entrenar ChatGPT, pero esta información no está disponible para el público. Simplemente no conocemos los detalles sobre cómo se entrena el modelo, qué datos se usaron, de dónde provienen ni cuáles son los detalles de su arquitectura.

***

En resumen, mucho de lo que la herramienta responde ante las preguntas de los usuarios concuerda con la información disponible en internet, pero también ChatGPT inventa respuestas que parecen razonables y revuelve lo cierto con sus mentiras ("alucinaciones"). Si el usuario no conoce el tema, se "traga" la respuesta entera. Si lo conoce, puede decirle qué parte está mal, y así ChatGPT aprende más a partir de las personas que lo corrigen.

Gran uso de ChatGPT

GPT-3 está siendo usado para multitud de propósitos. Además, montadas sobre GPT-3, GPT-4 y parecidos, están apareciendo y se desarrollarán herramientas especializadas para áreas específicas, que agregarán más conocimiento en cada campo. Algunos ámbitos en los que ChatGPT ayuda son:

Educación

- Aprendizaje personalizado: personalizar y adaptar la experiencia de aprendizaje a las necesidades individuales de cada estudiante (K, s/f).

- Aprendizaje de idiomas: Duolingo Max, lecciones de lenguaje personalizadas.
- Asistencia en la escritura.
- Apoyo en la investigación.
- Preparación de exámenes.
- Asistentes de enseñanza virtuales.
- Gamificación.
- Accesibilidad.
- Simulaciones y conversaciones con personajes históricos.
- Desarrollo profesional.

Negocios

- Aprendizaje personalizado.
- Entrenamiento.
- Productividad: los integrantes del equipo obtienen la información que necesitan cuando requieren.
- Automatización: generación automática de preguntas, respuestas, evaluaciones y cuestionarios basados en conocimientos a partir del contenido existente.
- Experiencia: a través de un asistente virtual, en cualquier lugar y en cualquier momento.
- Herramientas para escribir mejor (Distel, 2022).
- Optimización de la cadena de suministro: analizando los patrones de ventas.

Finanzas

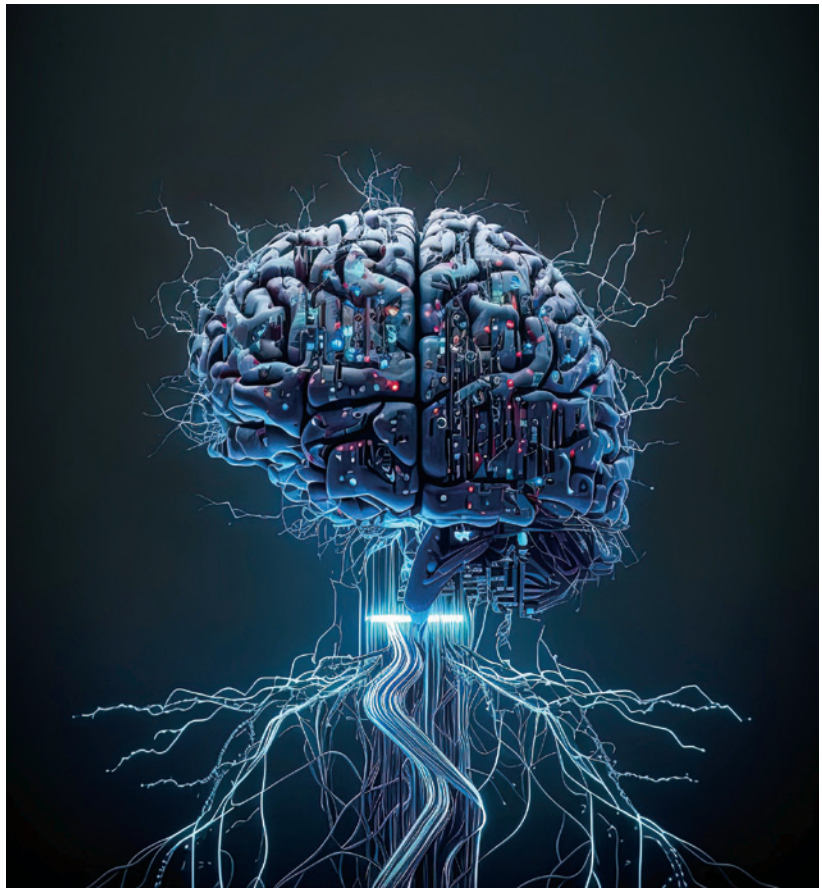
- Atención al cliente y experiencias (Martinez, 2023; Lamaj, s/f).
- Procesamiento de documentos.
- Asesoría financiera personalizada.
- Procesamiento de préstamos.
- Detección de fraudes.
- Análisis y predicción de inversiones.
- Análisis financiero.
- Cumplimiento y gestión de riesgos (Talerico, 2023).

Programación

- Entiende código: por ejemplo, Python (Anand, 2020).
- Diseña código: por ejemplo, plantillas web.
- Genera expresiones regulares: regex, etcétera.

Medicina

- Del diagnóstico al descubrimiento: en la investigación médica, dado lo avanzado del procesamiento de lenguaje natural de GPT-3, puede comprender e interpretar información médica, lo que hace posible analizar a un paciente y llevar a cabo tareas como: análisis de datos, generación de hipótesis, medicina personalizada, entre otras (Dev, 2023). La ventaja es tener una mayor precisión.
- Descubrimiento de medicamentos: GPT-3 puede analizar la información de la literatura científica y otras fuentes. Así, puede ayudar al descubrimiento de nuevos objetivos farmacológicos. Por ejemplo, al escanear grandes cantidades de información, GPT-3 puede identificar patrones y establecer conexiones que podrían no aparecer de inmediato para los investigadores humanos.
- Diseño de ensayos clínicos: GPT-3 puede ayudar a mejorar el diseño de ensayos clínicos.
- Desarrollo de fármacos: la capacidad de GPT-3 para predecir estructuras y propiedades moleculares se puede utilizar en el desarrollo de fármacos para generar compuestos químicos sintéticos, predecir la farmacocinética y los perfiles de toxicidad. Así, las aplicaciones de GPT-3 en el cuidado de la salud pueden ayudar a mejorar la formulación de fármacos y los regímenes de dosificación, así como a identificar los posibles efectos adversos antes de que comiencen los ensayos clínicos. Esto ayuda a mejorar la seguridad y la eficacia de los nuevos medicamentos.



En manos de grandes empresas tecnológicas (*Big Tech*)

Las empresas están ansiosas de explotar las "ventajas" de ChatGPT y GPT-4 para ofrecer productos y servicios especializados. Quizá exagerarán las bondades de sus productos y minimizarán sus problemas. A partir de esto, esperan tener una gran bonanza económica, no obstante los peligros ya mencionados.

Nos dice Melissa Heikkilä (2023) en *MIT Technology Review*:

> Si los reguladores no actúan ahora, el auge de la IA generativa concentrará aún más el poder de las *Big Tech*. Ése es el argumento central de un nuevo informe del instituto AI Now [un instituto de investigación sin fines de lucro]. Y tiene sentido. Para entender por qué, considere que el auge actual de la IA depende de dos cosas: grandes cantidades de datos y suficiente poder de cómputo para procesarlos. Ambos recursos sólo están realmente disponibles para las grandes empresas. Y aunque algunas de las aplicaciones más emocionantes, como el chatbot ChatGPT de Open IA y la IA de generación de imágenes Stable Diffusion de Stability.AI [así como el caso de Duolingo Max], son creadas por nuevas empresas [*startups*], dependen de acuerdos con las *Big Tech* que les dan acceso a sus vastos recursos informáticos y de datos.

Melissa Heikkilä cita a Sarah Myres West, directora gerente del instituto AI Now: "Un par de empresas *Big Tech* están preparadas para consolidar su poder a través de la IA, en lugar de democratizarlo".

Un competidor que puede vencerlo

En un artículo publicado en marzo de 2023, investigadores del tema de IA de la Universidad de Stanford y el instituto canadiense para IA, nombrado MILA, propusieron una tecnología que podría ser mucho más eficiente que GPT-4 o *software* similar, para engullir grandes cantidades de datos y transformarlos en una respuesta. Se llama Hyena. Ya veremos si resulta vencedora (Ray, 2023).

Recomendaciones

A los usuarios de la herramienta:

- "No creas todo lo que te dijeren".
- Úsalo como un buen ayudante, pero verifica.

A los gobiernos:

- Establezcan regulaciones sobre la producción de noticias falsas o engañosas, así como de las herramientas que las producen.

En resumen, éste es un caso en el que la investigación en IA continuará en manos de las empresas *Big Tech*. El sector académico generará poca investigación de vanguardia, debido a los grandes recursos de cómputo que se necesitan para analizar los enormes volúmenes de datos de entrenamiento. La evidencia sugiere que la IA aplicada al análisis de textos seguirá en manos privadas.

Regulación

Se espera un gran *boom* y muchas exageraciones comerciales: todo el mundo apresurándose a sacar el mejor provecho de esta herramienta de la IA (Heikkilä, 2023). Afortunadamente, ahora tenemos una mejor comprensión de todas las formas catastróficas en que la IA puede fallar. Los reguladores de todo el mundo están prestando mucha atención. Por ejemplo, más de diez mil miembros del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) firmaron la solicitud "Pausar experimentos gigantes de IA" en la que expresan su preocupación y piden una moratoria de seis meses (Anderson, 2023).

Con respecto a cómo debe regularse, el informe de AI Now argumenta que la propuesta de la Casa Blanca de abordar la responsabilidad de la IA con medidas posteriores al lanzamiento de productos, como auditorías algorítmicas, no es suficiente para mitigar los posibles daños de la IA. Sarah Myres West señala: los reguladores deben actuar rápidamente; "debe haber consecuencias cuando [las empresas de tecnología] violan la ley" (Heikkilä, 2023). En resumen, los gobiernos deben establecer regulaciones sobre los entes y las empresas que generen noticias falsas o engañosas, o que desarrollen herramientas capaces de producirlas.

Conclusión

ChatGPT aprende sobre lo que sucede y sucedió en el mundo mientras va leyendo y "digiriendo" una gran cantidad de conocimientos, tomados de Wikipedia, internet, grandes bases de datos... Para que sus transformadores los sinteticen y generalicen, requieren mucho poder de procesamiento. Durante su

entrenamiento, hay personas ("tutores") que le corrigen las respuestas malas. Los seres humanos aprendemos de esta misma manera, excepto que la cantidad de información que procesamos y guardamos es mucho menor, y también lo es nuestra velocidad de procesamiento. No usamos (supongo) transformadores.

También nosotros hacemos deducciones probables, quizá incorrectas: si no conocemos cuántos hijos tuvo Benito Juárez, suponemos que fueron cinco, dado que este oaxaqueño vivió en el siglo XIX. Nosotros nos percatamos cuando "estamos en lo cierto" (porque obtuvimos el dato de una fuente que suponemos confiable) o si estamos haciendo una estimación estadística. ChatGPT no puede detectar eso, únicamente aplica sus generalizaciones y muestra el resultado, pero no se da cuenta de que está mal. Sin embargo, tal como sucede con nosotros, un tutor o un usuario lo puede corregir, y así errará cada vez menos. Nótese que mientras más personas lo corrijan, mejores serán sus respuestas, pero sus errores o mentiras serán más difíciles de detectar.

Caveat: he tratado de citar todas las fuentes en las que me basé para hacer este trabajo.

Adolfo Guzmán Arenas
Centro de Investigación en Computación, Instituto Politécnico Nacional.
aguzman@ieee.org

Referencias específicas

Agomuoh, F. (27 de enero de 2023), "The 6 biggest problems with ChatGPT right now", *Digital Trends*. Disponible en: ‹https://tinyurl.com/ProblemasPrecison›, consultado el 15 de mayo de 2023.

Anand, A. (28 de septiembre de 2020), "Deep Learning Trends: top 20 best uses of GPT-3 by OpenAI", *Educative*. Disponible en: ‹https://tinyurl.com/UsoEduc›, consultado el 15 de mayo de 2023.

Anderson, M. (7 de abril de 2023), "'AI Pause' Open Letter Stokes Fear and Controversy", *IEEE Spectrum*. Disponible en: ‹https://tinyurl.com/CartaEspera›, consultado el 15 de mayo de 2023.

Dev, P. (16 de febrero de 2023), "The impact of GPT-3 in healthcare, pharma, medical research, and diagnosis", *Accubits Blog*. Disponible en: ‹https://tinyurl.com/UsoMedico›, consultado el 15 de mayo de 2023.

Distel, A. (26 de noviembre de 2022), "The 16 Best GPT-3 Tools To Help You Write Faster", *Jasper*. Disponible en: ‹https://www.jasper.ai/blog/gpt3-tools›, consultado el 15 de mayo de 2023.

Heikkilä, M. (19 de abril de 2023), "OpenAI's hunger for data is coming back to bite it", *MIT Technology Review*. Disponible en: ‹https://tinyurl.com/LeyProtegeDatos›, consultado el 15 de mayo de 2023.

Hetler, A. (2023), "Definition. ChatGPT", *TechTarget*. Disponible en: ‹https://tinyurl.com/AsiTrabaja›, consultado el 15 de mayo de 2023.

K, M. (s/f), "10 Practical Ways Educators Can Use GPT-3 for Enhanced Learning in and Beyond the Classroom", *Noodle Factory*. Disponible en: ‹https://tinyurl.com/UsoEducacion›, consultado el 15 de mayo de 2023.

Lamaj, D. (s/f), "The 3 Best GPT-3 Tools and How to Use Them in Marketing", *Publer*. Disponible en: ‹https://publer.io/blog/best-gpt3-tools/›, consultado el 15 de mayo de 2023.

Loizos, C. (9 de diciembre de 2022) "Is ChatGPT a 'virus that has been released into the wild'?", *TechCrunch*. Disponible en: ‹https://tinyurl.com/WildVirus›, consultado el 15 de mayo de 2023.

Martinez, C. (7 de marzo de 2023), "50 Ways to use Chat GPT-3 in Finance, FP&A and Investing. Part 4", *Medium*. Disponible en: ‹https://tinyurl.com/UsoFinanzas›, consultado el 15 de mayo de 2023.

Ortiz, S. (15 de marzo de 2023), "What is GPT-4? Here's everything you need to know", *ZDnet*. Disponible en: ‹https://www.zdnet.com/article/what-is-gpt-4-heres-everything-you-need-to-know/›, consultado el 15 de mayo de 2023.

Ray, T. (20 de abril de 2023), "This new technology could blow away GPT-4 and everything like it", *ZDnet*. Disponible en: ‹https://tinyurl.com/QuizaHyena›, consultado el 15 de mayo de 2023.

Talerico, A. (24 de febrero de 2023), "How Finance and Banking Professionals Can Use ChatGPT", *Corporate Finance Institute*. Disponible en: ‹https://tinyurl.com/UsoRiesgo›, consultado el 15 de mayo de 2023.

Wu, G. (6 de mayo de 2023), "8 Big Problems With OpenAI's ChatGPT", *Make Use Of*. Disponible en: ‹https://tinyurl.com/ProblemasGPT›, consultado el 15 de mayo de 2023.