



Recientemente, OpenAI ha anunciado la presentación de su última herramienta web, GPTBot, un rastreador online. El propósito de este bot es recolectar datos de acceso público con el fin de entrenar modelos de inteligencia artificial, asegurando que este proceso se llevará a cabo de manera transparente y responsable, según afirma la empresa.

De acuerdo a la documentación de lanzamiento proporcionada por OpenAI, el rastreador web realizará un filtrado para eliminar fuentes que requieran acceso mediante sistemas de pago, así como cualquier información de identificación personal (PII por sus siglas en inglés) o contenido que viole las regulaciones internas de la compañía. Según el creador de GPT, permitir la participación del bot ayudará a mejorar la precisión y capacidades de los sistemas de inteligencia artificial en el futuro.

Este paso revolucionario no solo se compromete a perfeccionar la precisión, habilidades y seguridad de los modelos de inteligencia artificial, sino que también provoca debates profundos sobre la ética de los datos, la propiedad y el uso en la era digital. Aunque OpenAI admite que obtiene información de la internet para entrenar sus enormes modelos de lenguaje como GPT-4, esta solución parece ser incompleta a la hora de abordar los dilemas éticos que surgen al utilizar datos provenientes de sitios web de terceros.

Restricciones de acceso a GPTBot pueden ser aplicadas

Reconociendo la diversidad inherente en los entornos digitales, OpenAI brinda a los administradores de sitios web la capacidad de determinar el nivel de interacción que GPTBot tendrá con sus sitios. Los webmasters pueden restringir completamente el acceso de GPTBot o definir los directorios que puede explorar mediante ajustes cuidadosos en los archivos robots.txt.

El lanzamiento de GPTBot ofrece a los administradores de sitios web y a los proveedores de contenido una nueva perspectiva, abriendo una ventana hacia la exploración de sus dominios digitales. Los webmasters pueden analizar las interacciones de GPTBot con sus sitios gracias a una documentación detallada, y pueden controlar el acceso mediante el uso del protocolo estándar robots.txt.



El control de acceso es una técnica sencilla que implica la inclusión de las siguientes directivas:

- User-agent: GPTBot Disallow: /

La siguiente estructura puede ser utilizada para un enfoque más preciso que permita un acceso selectivo:

- User-agent: GPTBot Allow: /directorio-1/ Disallow: /directorio-2/
- Equilibrio delicado: Consideraciones legales, éticas y de propiedad

Recientemente, OpenAI ha presentado una solicitud de marca registrada para 'GPT-5', sugiriendo que la empresa está entrenando la siguiente versión de GPT-4, que según varias fuentes, estará cerca de lograr la inteligencia artificial general (AGI), un objetivo que la compañía ha perseguido desde el principio. Sin lugar a dudas, GPTBot ayudará a la organización a recopilar más datos de internet para entrenar este nuevo modelo. Por otro lado, la empresa también ha dejado de utilizar su Clasificador de IA para identificar el texto producido por GPT.