La caverna de ChatGPT

Diego Germán González | 16/02/2023

Dicen que para novedades los clásicos. **Una alegoría escrita cuatro siglos años de nuestra era resulta ideal para entender cuáles son los límites de las nuevas aplicaciones de Inteligencia Artificial.** me refiero a la «Caverna de ChatGPT» que no es ni más ni menos que una adaptación de la famosa alegoría de la caverna de Platón

No tengo nada que objetar al uso de las herramientas de inteligencia artificial. De hecho, encuentro que facilitan mucho el trabajo. Pero, siempre y cuando **sea usada por personas que tengan conocimientos suficientes como para evaluar su trabajo.**

Por poner un ejemplo; uno puede pedirle a ChatGPT que escriba un plugin de WordPress, pero si se carece de conocimientos sobre PHP ese plugin puede producir graves problemas de seguridad.

Índice

La alegoría de la caverna

Platón fue un filósofo griego que vivió entre los siglos V y IV antes de Cristo. Expresaba sus pensamientos en la forma de mitos y alegorías. **La más conocida de ellas fue la de la caverna.**

Publicada en *La República*, la alegoría imagina a **un grupo de personas encadenadas en una caverna, detrás de ellos tienen un fuego que arroja sombras sobre la pared que tienen frente a ellos. Las sombras son lo único que ven e imaginan que son lo único que existe ignorando lo que hay más allá.**

Cuando uno de los prisioneros es liberado puede ver el mundo como realmente es y se da cuenta de lo limitado de sus experiencias en la caverna.

Según los estudiosos de Platón, esta alegoría pone de relieve que todos vivimos nuestras vidas basándonos en nuestras propias informaciones y experiencias. Informaciones y experiencias equivalentes a las sombras de la caverna. Al igual que los prisioneros, existe la verdadera realidad y está más allá de nuestra comprensión.

La caverna de ChatGPT

ChatGPT y sus competidores tienen tanto admiradores como detractores. Pero, nadie había dado una explicación técnica sobre sus fallas hasta un artículo publicado en New Yorker por el escritor de ciencia ficción Ted Chang

Para explicar las fallas de los modelos de lenguaje Chang hace una analogía con lo que sucede con imágenes y archivos de audio.

La grabación y reproducción de un archivo digital requiere de dos pasos: el primero es la **codificación, momento en el cual el archivo se convierte a un formato más compacto, y a continuación la decodificación, que es el proceso inverso**. El proceso de conversión se denomina sin pérdida (el archivo restaurado es igual que el original) o con pérdida (Parte de la información se perdió para siempre). La compresión con pérdida se aplica en archivos de imágenes, video o audio y la mayoría de las veces no es perceptible. Cuando lo es se denomina artefacto de compresión». Los artefactos de compresión se presentan en la forma de partes borrosas en imágenes o sonido metálico en audio.

Chang usa la analogía de un JPG borroso de la web para referirse a los modelos de lenguaje. Y, esta es bastante exacta. **Ambos comprimen la información conservando solo «Lo importante». L**os modelos de lenguaje generan, a partir de grandes cantidades de datos de texto, una representación compacta de los patrones y relaciones entre palabras y frases.

A partir de ella se genera texto nuevo tratando en lo posible que sea similar en contenido y significado al texto original. El problema es cuando en la web no hay información suficiente para que se pueda generar un texto nuevo. Esto se traduce en que ChatGPT pueda escribir un ensayo de nivel universitario, pero no hacer sencillas operaciones de 5 dígitos.

Chang concluye que:

> Incluso si es posible restringir que los grandes modelos de lenguaje participen en la creación, ¿deberíamos usarlos para generar contenido web? **Esto tendría sentido solo si nuestro objetivo es reempaquetar la información que ya está disponible en la Web.** Algunas empresas existen para hacer precisamente eso; generalmente las llamamos fábricas de contenido. Quizás la borrosidad de los modelos de lenguaje les sea útil, como una forma de evitar la infracción de derechos de autor. Sin embargo, en términos generales, diría que cualquier cosa que sea buena para las fábricas de contenido no es buena para las personas que buscan información. **El aumento de este tipo de reenvasado es lo que nos dificulta encontrar lo que estamos buscando en línea en este momento**; cuanto más se publica en la Web el texto generado por grandes modelos de lenguaje, más se convierte la Web en una versión más borrosa de sí misma.

Y, como los prisioneros de la caverna, nuestra experiencia sería mucho más pequeña de lo que la realidad nos ofrece.