Cómo Silicon Valley construyó la IA: comprando, escaneando y desechando millones de libros.

Tiempo de lectura: 3 min.

Anónimo

En los últimos años, las grandes empresas de tecnología y startups de Silicon Valley han librado una competencia frenética por adquirir conjuntos masivos de datos textuales, especialmente libros, como materia prima para entrenar sus modelos de inteligencia artificial (IA).

Según documentos judiciales revelados en demandas por derechos de autor, compañías como Anthropic, Meta, Google y OpenAI emprendieron acciones a gran escala para obtener millones de títulos físicos y digitales con el objetivo de mejorar la capacidad de sus modelos de lenguaje para “entender” y “escribir bien”.

Un caso emblemático es el llamado Project Panama de Anthropic, descrito en documentos internos como un esfuerzo por comprar y escanear “todos los libros del mundo”. La compañía gastó decenas de millones de dólares comprando grandes lotes de libros, a menudo en lotes de decenas de miles, y contrató servicios profesionales para desencuadernar y escanear las páginas a gran velocidad. Después del escaneo, muchas de estas copias físicas fueron recicladas o

descartadas, lo que ha generado preocupación entre autores y defensores del patrimonio cultural por la eliminación física de obras impresas.

Los detalles de Project Panama, inéditos hasta ahora, salieron a la luz en más de 4.000 páginas dee documentos incluidos en una demanda por derechos de autor interpuesta por escritores contra Anthropic. La empresa, valorada por sus inversores en unos 183.000 millones de dólares, aceptó pagar 1.500 millones de dólares para cerrar el litigio en agosto. Sin embargo, la decisión de un juez federal de hacer públicos numerosos documentos del caso permitió conocer con mayor profundidad la intensidad con la que Anthropic persiguió la obtención de libros.

Estos nuevos archivos, junto con otros presentados en demandas similares contra empresas de inteligencia artificial, revelan hasta qué punto compañías tecnológicas como Anthropic, Meta, Google u OpenAI llegaron a extremos notables para reunir enormes volúmenes de datos con los que “entrenar” sus sistemas. En esa carrera acelerada, los libros fueron considerados un botín esencial. Así lo reflejan los registros judiciales: en enero de 2023, uno de los cofundadores de Anthropic sostenía que entrenar modelos con libros permitiría enseñarles “a escribir bien”, en lugar de limitarse a reproducir un “lenguaje de baja calidad propio de internet”. En un correo interno de Meta fechado en 2024, el acceso a grandes bibliotecas digitales se calificaba directamente como “imprescindible” para competir con otros actores del sector.

Sin embargo, los documentos sugieren que las empresas no consideraron viable solicitar autorización directa a autores y editoriales. En su lugar, según las acusaciones recogidas en los autos, Anthropic, Meta y otras compañías recurrieron a métodos de adquisición masiva sin conocimiento de los creadores, incluida la descarga de copias pirateadas.

Estos esfuerzos reflejan las tensiones legales y éticas detrás del entrenamiento de IA con datos culturales. Muchos autores y editoriales han emprendido demandas alegando que la adquisición y uso masivo de sus obras para entrenar modelos de IA se hizo sin permiso y constituye una violación de derechos de autor. A su vez, las empresas tecnológicas han argumentado que el uso es “transformador” y, en algunos fallos judiciales, se ha considerado legal bajo la doctrina de fair use (“uso justo”). No obstante, los documentos judiciales también han expuesto que algunas empresas, incluyendo Meta, consideraron o incluso utilizaron descargas masivas desde bibliotecas pirata en línea como LibGen para obtener copias digitales de libros

sin pagar por ellos, lo que ha intensificado las críticas sobre prácticas poco transparentes.

En el caso de Meta, varios empleados expresaron internamente su inquietud ante la posibilidad de infringir la ley de derechos de autor al descargar millones de libros sin permiso. Aun así, un correo electrónico de diciembre de 2023 indicaba que la práctica había sido aprobada tras una “escalada a MZ”, en aparente referencia al consejero delegado Mark Zuckerberg. Meta declinó hacer comentarios al respecto.

Además de las cuestiones legales, expertos y críticos han señalado preocupaciones más amplias sobre el impacto cultural y social de estas prácticas. La destrucción física de libros tras su digitalización plantea preguntas sobre la preservación del patrimonio literario y el valor intrínseco de las obras impresas como registros culturales. Del mismo modo, la dependencia de datos extraídos de fuentes no autorizadas subraya la necesidad de un marco ético y regulador más robusto en torno al uso de contenidos creativos para construir inteligencias artificiales avanzadas.

Basado en :

Schaffer, Aaron; Oremus, Will y Tiku, Nitasha. “How Silicon Valley Built AI: Buying, Scanning & Discarding Millions of Books”, MSN (Washington Post), 27 de enero de 2026. https://www.msn.com/en-us/technology/artificial-intelligence/how-silicon-valley-built-ai-buying-scanning-and-discarding-millions-of-books/ar-AA1V4aZv

ver PDF
Copied to clipboard