

INTELIGENCIA ARTIFICIAL >

E

Cuando el conocimiento de internet ya no es suficiente: la IA le pone el ojo a las librerías de segunda mano

Libreros de todo el mundo han detectado compras masivas de libros descatalogados por parte de empresas tecnológicas para entrenar sus modelos de inteligencia artificial



El librero Marçal Font, fotografiado en su casa.
GIANLUCA BATTISTA

SUSANA PÉREZ SOLER

21 JUL 2026 - 05:30CEST

Añadir EL PAÍS en Google

Marçal Font se sorprendió un día al detectar unos movimientos extraños en su librería Fènix de Badalona. “Me compraron cuatro libros por 35 euros con unos gastos de envío de 67 euros. Un minuto después [pedían otro libro](#) con 20 euros más de gastos de envío. Y así. No tenía ningún sentido”, recuerda. Fue entonces cuando contactó con su amigo Xavi Vinaixa, investigador y director técnico de una empresa tecnológica, para averiguar qué podía haber detrás y no tardaron en dar respuesta al misterio: una empresa estaba haciendo los pedidos para digitalizar los libros e incorporarlos a sistemas de inteligencia artificial.

Mediante un foro internacional se pusieron en contacto con libreros europeos a quienes [les ocurría lo mismo](#). Las compras se realizaban a través de ZoomBooks (empresa canadiense de compra-venta de libros de segunda mano) y los envíos se dirigían a operadores logísticos intermediarios como PrepFort, en Illinois, que ofrece [puestos de trabajo](#) para escanear, catalogar y registrar datos con precisión y rapidez. Tanto los libreros como los investigadores consultados sostienen que esta estructura dificulta identificar al destinatario final de los ejemplares, aunque todo apunta a que se trata de tecnológicas vinculadas al desarrollo de inteligencia artificial.

Las sospechas de los libreros encontraron eco en una investigación de *The Washington Post* sobre [Anthropic, la empresa creadora del modelo Claude](#). A partir de documentos judiciales relacionados con demandas por derechos de autor, el diario reveló la existencia de un programa interno conocido como [Project Panama](#), descrito por la propia compañía como un esfuerzo para digitalizar a gran escala libros impresos con el fin de incorporarlos [a sus sistemas de inteligencia artificial](#). Según la documentación analizada por el periódico estadounidense, el proceso se basa en el escaneo destructivo de ejemplares físicos.

Destruir libros para conseguir datos

La imagen recuerda inevitablemente [a Fahrenheit 451](#), aunque aquí los libros no se destruyen para censurar ideas, sino para extraer datos. “Desencuadernar un libro guillotinando el lomo para liberar sus páginas permite escanear y digitalizar el ejemplar de manera más rápida y barata que el escaneo conservador y cuidadoso utilizado por bibliotecas y archivos”, apunta Vinaixa, aunque implica la desaparición del ejemplar físico. Resulta paradójico que un proyecto creado para alimentarse de datos termine destruyendo datos por el camino.

A diferencia de [la novela de Bradbury](#), aquí nadie quiere eliminar el contenido. Al contrario, lo que quieren es absorberlo. “Durante años, las IA se entrenaron con enormes cantidades de texto disponible en la web. Ahora muchas compañías se enfrentan a varios problemas porque ya han consumido buena parte del texto público de calidad disponible y necesitan datos más diversos, especializados y mejor escritos para intentar evitar la homogeneización, los sesgos cognitivos, los errores y el vocabulario informal que predomina en internet”, añade Vinaixa.

Por eso ahora los libros son tan valiosos para las empresas tecnológicas: contienen conocimiento más elaborado, editado y estructurado que [gran parte del contenido de internet](#). “Compan libros especializados, de escasa salida comercial y gran valor documental: estudios locales, publicaciones técnicas, gastronomía tradicional o ensayos que a menudo solo interesan a investigadores”, apunta Font. Ahora están comprando libros con ISBN, lo que garantiza que gran parte de los libros adquiridos forman parte de un ecosistema relativamente bien catalogado y protegido por sistemas de depósito legal.

¿Pero qué ocurrirá si las empresas quieren seguir ampliando sus repositorios? ¿Dónde termina la demanda de las empresas tecnológicas por los datos y cuándo empieza a protegerse el contenido creado por la humanidad? “[Esta es mi mayor preocupación](#)”. Acabarán llegando a materiales mucho menos protegidos e identificados: fanzines, publicaciones locales, boletines vecinales o documentación de movimientos sociales que nunca entraron en los circuitos bibliográficos convencionales”, añade Font.

Además, cada vez hay más litigios sobre *copyright*. En diciembre de 2023, [The New York Times demandó a OpenAI](#) y a Microsoft por utilizar su contenido para entrenar sistemas de inteligencia artificial generativa. El diario asegura que millones de artículos protegidos con derechos de autor se han utilizado para entrenar chatbots que ahora compiten con él. En abril de 2024, ocho periódicos estadounidenses se sumaron a la demanda.

El gran colapso

Si las IA terminan absorbiendo gran parte del corpus humano disponible, una posibilidad es que empiecen a entrenarse cada vez más con contenido generado por otras IA. Ese fenómeno ya tiene nombre en investigación: *model collapse* o colapso del modelo. La hipótesis es que, si una generación de modelos aprende de la producción de la generación anterior, el conocimiento puede degradarse progresivamente. Es una especie de fotocopia de una fotocopia de una fotocopia. “Por eso los libros descatalogados representan una reserva de contenido humano todavía no contaminado por la inteligencia artificial”, apunta Antonio Ortiz, analista y divulgador especializado en inteligencia artificial y ciencias tecnológicas.

“Ahora bien, si es cierto que la falta de datos es un problema dentro del sector, también lo es que no es algo tan central como hace dos años”, añade. “Los modelos más avanzados están mejorando gracias al aprendizaje por refuerzo (*reinforcement learning*). Antes la industria buscaba más volumen de datos y mayor computación, mientras que ahora apuesta por un entrenamiento especializado a partir de datos de calidad y problemas bien resueltos. Los datos siguen siendo importantes, pero [la carrera se está desplazando](#) hacia datos cada vez más cualificados”, añade.

La privatización del conocimiento

Los grandes avances científicos del siglo XX ocurrieron gracias a datos, [universidades e investigación pública](#), que se conservaba también en instituciones públicas. Pero algo parece estar cambiado: empieza a haber una traslación del descubrimiento, el desarrollo y la conservación científica hacia una cierta privatización.

“El principal riesgo es [una privatización del conocimiento](#). Lo lógico sería aprovechar primero el contenido humano para entrenar modelos de gran lenguaje públicos, como el europeo Mistral AI, porque el conocimiento tiene que seguir siendo un bien público, y respondiendo a lógicas de transparencia, reconocimiento de autoría y sostenibilidad”, sostiene Patricia Ventura, doctora en Comunicación, Medios y Cultura por la Universitat Autònoma de Barcelona.

“Toda la labor magnífica que se está haciendo para preservar el patrimonio hoy, la IA la hace infinitamente más rápido”, señala Font. “La cuestión no es si la IA está destruyendo el patrimonio, sino si las instituciones públicas serán capaces de identificar, [catalogar y conservar el patrimonio](#) antes de que las máquinas lo conviertan en datos privados”, añade. Para ello es importante entender qué está en juego en este momento.

Y no es la distopía planteada por Bradbury —no quieren quemar los libros para hacerlos desaparecer—, sino más bien absorber sus datos y privatizar el conocimiento. Porque, a diferencia del polémico proyecto iniciado por Google hace veinte años, [Google Books](#), esta vez no pretenden devolvernos el acceso a la fuente primaria digitalizada, sino permitirnos acceder a un conocimiento transformado: aquel que regurgita la máquina después de engullir el libro.

ÚLTIMAS NOTICIAS

05:45 | Burnham hereda una relación con la UE que apenas ha despegado tras el Brexit y nuevos roces con Trump

05:45 | La nueva financiación autonómica promueve una armonización entre comunidades para frenar las rebajas de impuestos

05:45 | Hablemos de salud

05:45 | Macroincendios en Guadalajara, Zaragoza, Almería: ¿nos tenemos que acostumbrar a esto?

LO MÁS VISTO

1. Lamine emula a Pelé y niega a Messi

2. El mal perder de Argentina: pelea tras la derrota y desplante a España durante la entrega de la Copa

3. Última hora de las celebraciones de los campeones del mundo - La selección se da un baño de masas a su vuelta a España en la celebración del título de campeones del mundo

4. Así esta siendo la celebración de España tras ganar el Mundial: actos, recorrido y horarios de la rúa por las calles de Madrid

5. El momento en el que Messi pide la tarjeta roja para Cucurella