Perché le aziende di AI acquistano libri per distruggerli? «Conoscenze analogiche libere da contaminazioni digitali» di Roberto Cosentino

Lingue poco rappresentate, conoscenze mai digitalizzate, una libreria digitale comune. Ecco i vantaggi che potrebbero portare gli acquisti massivi di libri

(Fonte: https://www.corriere.it/ 14 settembre 2026)

Che le aziende di intelligenza artificiale stiano acquistando **container pieni di libri** per l'addestramento dei propri modelli, **ormai non è più un segreto**. Dopo la causa ad Anthropic che ha scoperchiato il vaso di Pandora che ha portato alla luce il famigerato **Project Panama**, l'indagine di 404 Media ha poi fatto emergere le dinamiche di Amazon per le medesime motivazioni.

Ormai è noto che il colosso Usa acquisti numerosi volumi per inoltrarli presso i centri per l'addestramento dell'AI, dove vengono scansionati e distrutti. E a fugare ogni dubbio sulla natura dei centri, persino il logo: **un dinosauro che si avventa su di un libro aperto**. Abbiamo provato a contattare Amazon per porgere ulteriori domande. Nella nostra indagine, che ci ha portato a conoscere Mirko, il libraio del Centro Italia che ha raccontato di aver venduto libri a **Zoom Books**, abbiamo individuato nei forum riservati ai venditori di prodotti su Amazon **una serie di librerie di tutta Europa** in cerca di risposte sugli acquisti massivi di libri.

Una delle nostre domande poste alla compagnia riguardava l'ipotesi che vi siano altre aziende che sfruttano i sistemi di proprietà di Amazon (tra cui AbeBooks o Amazon stesso) per l'acquisto di libri per addestrare l'AI. Oltre alla risposta che il colosso statunitense ha già fornito ai colleghi di *404 Media*, ovvero «Amazon acquista libri attraverso canali commerciali per contribuire allo sviluppo e

al miglioramento dei prodotti e dei servizi utilizzati dai nostri clienti» ha aggiunto un laconico «Amazon preferisce non commentare ulteriormente». Sappiamo dunque che le società si stanno muovendo per l'acquisto massivo di libri e la rispettiva distruzione. **Ma perché questo avviene**?

Facciamo un passo indietro

Bisogna chiarire prima un aspetto della vicenda. **L'addestramento dell'AI tramite libri non è di certo una novità**. Alcune compagnie, come la stessa Amazon, acquista tomi **almeno dal 2024**.
La differenza con gli inizi è che le società facevano razzia dei libri piratati, ovvero in formato digitale, in violazione del copyright, benché si celavano dietro il diritto americano che disciplina il **fair use**, cioè la possibilità di utilizzare materiale protetto anche senza chiedere il permesso ai detentori dei diritti. Ma per rispondere alla domanda di cui prima, dobbiamo andare all'estate del 2025, quando il ricercatore **Simon Rowberry** nella sua ricerca **The value of books in the age of generative AI training data**, ha affermato che «I libri rappresentano una fonte liminale per gli LLM in quanto forniscono contenuti di ampio respiro, curati e redatti».
Questo perché i libri forniscono **un testo lungo, strutturato e curato a livello editoriale e qualitativo**, un materiale insomma di tutt'altra natura rispetto a quello reperito dal Web e, se antecedente all'approdo sul mercato degli LLM, del tutto privi della "contaminazione" **dai tic linguistici** (e dalle allucinazioni) presenti magari in volumi usciti dopo l'arrivo di ChatGpt e colleghi.
Oltre a questo, **grande importanza deriva dai metadati**: autore, editore, un anno di pubblicazione, un contesto. Permettono a un modello di seguire ragionamenti, storie e concetti per centinaia di pagine invece che per poche migliaia di caratteri.

La ricerca di Creative Commons

Ad aggiungere un'ulteriore spiegazione a questo interesse da parte delle aziende di AI **è il rapporto pubblicato lo scorso aprile Public Interest Corpus** pubblicato su Authors Alliance, un'organizzazione la cui missione è **promuovere gli interessi degli autori** che desiderano «servire il bene pubblico» attraverso la condivisione dei propri lavori. Secondo il rapporto, i libri sono ormai riconosciuti come *training data* di alta qualità per i sistemi di intelligenza artificiale, non solo per le informazioni contenute, **ma anche perché contengono conoscenza in più discipline, periodi e scritti in diverse lingue**.
Ma un'altra risposta a tutto questo può arrivare direttamente da un paper del 2024 intitolato *Towards a Books Data Commons for AI Training*, stilato in collaborazione con **Creative Commons**, l'organizzazione internazionale no-profit che da oltre vent'anni lavora sul rapporto tra diritto d'autore, accesso alla conoscenza e riutilizzo dei contenuti. È la stessa realtà che ha creato le licenze Creative Commons, ormai diventate uno degli standard del web aperto. Il documento nasce da una serie di workshop organizzati con Open Future e Proteus Strategies, con il coinvolgimento di realtà come **Internet Archive**, **HathiTrust Digital Library**, Cornell University, UC Berkeley, UC Davis, Public Knowledge, Library Futures e NYU Engelberg Center. Dentro ci sono quindi alcuni dei

soggetti che, da anni, **lavorano proprio sui grandi archivi digitali**, sulla scansione dei libri, **sulla conservazione della conoscenza** e sui limiti del copyright.

Si parte da Books3 e si arriva a Books Data Common

Il punto di partenza è **Books3**, un dataset composto da oltre **170 mila libri** trasformati in testo, molti dei quali ancora protetti da copyright, poi confluito in **The Pile**, una delle più importanti raccolte utilizzate per l'addestramento dei modelli linguistici. Books3 è diventato rapidamente uno dei simboli della battaglia tra autori e aziende di intelligenza artificiale: il corpus è stato associato al training di modelli sviluppati da aziende e ha alimentato richieste di rimozione, contestazioni e cause contro alcune delle principali società del settore. Da qui nasce un problema: i libri sono così importanti per allenare un'AI, ma come si può procurarsene a sufficienza senza attingere da archivi piratati?

È da questa domanda che nasce l'idea di un **Books Data Commons**, una sorta di gigantesca biblioteca digitale pensata anche per l'addestramento dell'AI. Un archivio condiviso, costruito secondo regole precise, dal quale università, ricercatori e aziende possano attingere senza dover ogni volta ricominciare da zero. **Ed è qui che il libro fisico torna improvvisamente al centro della storia**. Perché avere milioni di ebook in circolazione non significa (solo) avere milioni di libri già utilizzabili per allenare un modello. **Molti sono protetti da vincoli contrattuali, altri ancora non esistono proprio in una versione digitale adatta a essere elaborata**. In tutti questi casi bisogna tornare alla carta: prendere il volume, scansionarlo, trasformare le pagine in testo con sistemi adeguati, correggere gli errori, ripulire il file, ricostruire la struttura e prepararlo perché possa essere usato nel training.

Il vantaggio dei libri fisici

C'è poi da considerare la lunga storia di questo patrimonio analogico. **La Rete racconta soprattutto gli ultimi decenni**. Le biblioteche e le librerie, invece, **custodiscono secoli di conoscenza**, **testi mai digitalizzati, lingue meno rappresentate**, manuali, saggi, narrativa, pubblicazioni scientifiche, opere locali, materiale fuori catalogo.

Ogni libro che non esiste già nei grandi dataset può rappresentare delle nuove informazioni. A quel punto anche il valore di un vecchio volume cambia completamente. **Per un libraio può essere un libro invendibile, come nel caso di Mirko e molti dei suoi colleghi**. Magari certi volumi sono fermi in magazzino da anni, a prendere polvere, con un valore commerciale vicino allo zero. Per chi costruisce un corpus di addestramento può invece essere un pezzo di testo che manca, una lingua poco rappresentata, **una conoscenza mai finita online** o un'opera che nessun altro ha ancora digitalizzato.

Il paper di Creative Commons insiste anche su un altro aspetto: digitalizzare libri su larga scala costa moltissimo. Google, ricorda il documento, **ha già scansionato circa 40 milioni di volumi**, grazie anche agli accordi con grandi biblioteche. La digitalizzazione dei primi 25 milioni sarebbe

costata circa **400 milioni di dollari**. Replicare oggi una biblioteca di quelle dimensioni **richiederebbe risorse enormi e anni di lavoro**. Comprare volumi rari e introvabili, in lingue meno diffuse, più che una stranezza può diventare quindi una risorsa preziosa.