LA NUOVA BUSSOLA QUOTIDIANA
Direttore Riccardo Cascioli
FATTI PER LA VERITÀ

tecnologia

Addestrare l'IA costa e Meta utilizza dati piratati

ATTUALITà 04_04_2025

Daniele Ciacci

L'ultimo report dell'*Atlantic* sui metodi di acquisizione dei dati per l'addestramento dei modelli di intelligenza artificiale solleva un interrogativo cruciale: quanto sono davvero sostenibili questi sistemi tecnologici che promettono di rivoluzionare il nostro modo di comunicare e lavorare?

Un articolo della famosa testata ha infatti rivelato come Meta abbia utilizzato Library Genesis

(LibGen), una vastissima biblioteca online di libri piratati, per addestrare il suo modello di intelligenza artificiale Llama 3. Anche il processo decisionale interno è stato chiaro: alla decisione, troppo onerosa, di acquistare libri e ricerche per allenare la propria IA, l'amministratore delegato Mark Zuckerberg ha invece approvato la possibilità di scaricare i dati da LibGen.
LibGen è una repository pirata di ebook e libri in diverse lingue e a disposizione di chiunque. Essa contiene 81 milioni di articoli di ricerca e oltre 7,5 milioni di opere, alcune anche di autori contemporanei come Sally Rooney e Jonathan Haidt, che hanno ancora garantiti i propri diritti d'autore.

In parole povere, Meta ha adottato consapevolmente contenuti piratati per i propri scopi commerciali. Pur sostenendo, Meta e OpenAI, che l'uso dei contenuti rientri nel "fair use", sono state comunque intentate diverse cause per violazione dei diritti d'autore, di cui ovviamente la prima interessata è LibGen stessa.

I Large Language Model (LLM), che apprendono il linguaggio umano e lo replicano tramite i moderni sistemi di IA, si trovano oggi di fronte a una doppia sfida economica. Da un lato, il fabbisogno energetico per l'addestramento e il funzionamento di questi sistemi è enormemente oneroso. Un singolo modello come GPT-3 può richiedere l'equivalente del consumo energetico annuale di centinaia di abitazioni, traducendosi in un'impronta di carbonio significativa.
Dall'altro lato, l'acquisizione dei contenuti necessari per addestrare questi sistemi sta diventando un terreno di scontro legale e morale, e le aziende tecnologiche sono disposte a percorrere strade eticamente discutibili pur di ottenere i dati necessari.

Il meccanismo è paradossale: per rendere l'intelligenza artificiale economicamente sostenibile, le aziende tecnologiche stanno sostenendo costi enormi in termini di risorse energetiche e legali. L'addestramento di un singolo modello può costare milioni di dollari, tra consumi elettrici, infrastrutture computazionali e potenziali contenziosi sui diritti d'autore.
Le soluzioni attuali sembrano essere scorciatoie che sollevano più problemi di quanti ne risolvano. La pirateria di contenuti scientifici e letterari non solo è illegale, ma decontestualizza la conoscenza, privando gli autori originali del riconoscimento del loro lavoro.

Pertanto, la vera sfida per l'intelligenza artificiale del futuro non sarà solo tecnologica, ma economica e ambientale: come possiamo sviluppare sistemi LLM che siano eticamente corretti nell'acquisizione dei contenuti, energeticamente efficienti ed economicamente sostenibili?

Le aziende tecnologiche dovranno necessariamente ripensare i loro modelli di sviluppo, investendo in fonti energetiche alternative (nucleari o rinnovabili), sistemi di acquisizione dati legali e trasparenti e modelli di addestramento più efficienti: l'innovazione tecnologica non può più prescindere da una visione sistemica che non consideri questi elementi.