L'INK SCRITTURA PROFESSIONALE

I Conversatori Scritti Non Umani sono fonti terziarie

di Giovanni Acerboni, 26 aprile 2024

"Ed elli avea del cul fatto trombetta" (Inf. XXI, v. 139).

Che cosa c'entra Dante con ChatGPT e gli altri Large Language Model generatori di testo, che sono Conversatori Scritti Non Umani - CSNU?

Rispondo con un'altra domanda: a chi è capitato che i CSNU abbiano usato "elli" o "avea" nelle loro risposte? Credo mai. Eppure nel dataset di addestramento di ChatGPT c'è la Divina Commedia (gliel'ho chiesto).

Il fatto è che la Divina Commedia è una fonte primaria e, come tale, statisticamente poco rilevanti nel dataset di addestramento.

Il dataset di addestramento

I CSNU elaborano le loro risposte sfruttando il dataset con cui sono stati addestrati. Per esempio, ChatGPT è stato originariamente addestrato con:

- informazioni pubblicate in Internet;
- informazioni di cui è stata acquisita la licenza d'uso.

Per quanto riguarda le prime, questo è lo schema (1):

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Una volta uscito sul mercato, ChatGPT ha utilizzato anche le informazioni fornite dagli utilizzatori. Risulta del tutto evidente che le fonti con cui è stato composto e viene aggiornato il dataset sono:

- in minoranza primarie = origine della conoscenza (opere letterarie, ricerche scientifiche, leggi ecc.);

L'INK SCRITTURA PROFESSIONALE S.a.s.
via Biella 24/A - 20143 Milano
telefono: ++39 02 36 53 48 71
fax: ++39 02 99 98 18 14
e-mail: info@scritturaprofessionale.it
web: www.scritturaprofessionale.it
Partita IVA e Codice Fiscale: 04305670962

- poi secondarie = elaborazione di fonti primarie (manuali universitari, libri di storia, siti istituzionali, cronaca giornalistica);
- prevalentemente terziarie = elaborazione di fonti secondarie (Common Crawl, Wikipedia, Social media, Siti web, altri articoli giornalistici ecc.).

I CSNU elaborano le risposte in base a un complesso meccanismo statistico che produce frasi sempre ben formate ma non sempre contenuti adeguati, pertinenti, veri (dipende molto anche da come viene composto il prompt).

Una delle cause di questi difetti è il rumore, l'interferenza tra fonti di rango diverso, oltre che di argomento diverso.

Per esempio, la difficoltà – e a volte l'impossibilità – di ottenere risposte adeguate a richieste molto specifiche e tecniche dipende dalla rarità – e a volte dalla mancanza – nel dataset di testi di livello specialistico. I quali, pur se presenti, non incidono statisticamente nell'elaborazione della risposta, che viene costruita piuttosto con informazioni di livello più generale, presenti nella stragrande maggioranza di fonti di rango inferiore.

Pertanto, **i CSNU sono per definizione fonti terziarie**, al massimo.

Per quanto possa apparire o essere originale e persino valida, la risposta di un CSNU non può per definizione uscire dal recinto culturale del suo dataset.

Il test della bibliografia

Dell'affidabilità di una fonte terziaria bisogna sempre dubitare. Ma non siamo più abituati a farlo. Tendiamo a fidarci. Il successo di quella fonte, decretato da tanti come noi, ci induce a fidarci.

Il test della bibliografia funziona piuttosto bene per accertarsi (o quasi) del rango di una fonte. Una buona fonte secondaria, come un manuale universitario, ha una bibliografia con tutte le fonti primarie essenziali e con una selezione di fonti secondarie di buon livello con le quali l'autore discute nel testo (cioè, non le cita solo per far bella figura).

Le fonti terziarie possono persino non avere la bibliografia, oppure averla, come Wikipedia o la manualistica farlocca, ma incompleta, incoerente, sconclusionata.

I CSNU, addirittura e spesso, inventano la bibliografia. Ciò accade, verosimilmente, sia per il meccanismo statistico che sta alla base della risposta, sia perché i CSNU non pensano, non hanno conoscenza. Pertanto, non possono esercitare alcun controllo critico sulla quantità sterminata delle informazioni di cui dispongono e, di conseguenza, nemmeno sulla qualità della risposta che offrono all'utilizzatore.

L'addestramento dei CSNU sistemici

Non credo – ma posso facilmente sbagliare – che i CSNU nelle versioni che l'AI Act definisce "generiche" verranno in futuro addestrati con dataset composti diversamente, proprio perché sono prodotti commerciali general purpose, destinati a un pubblico enorme e indifferenziato.

Piuttosto, le cose cambieranno con i CSNU "sistemici", cioè quelli addestrati con un dataset specifico per svolgere task specifici. Per esempio, un dataset di contratti o di policy ecc. potrebbe consentire al CSNU di elaborare una bozza di contratto o di policy molto più raffinata di quella che potrebbe elaborare un CSNU generico.

Anche nel caso dei CSNU sistemici possiamo però porre due questioni già viste:

- la prima: il rumore prodotto dal dataset con cui il CSNU è stato preaddestrato potrebbe ridurre la qualità delle risposte;
- la seconda: il dataset specifico costituisce in se stesso una fonte primaria o secondaria?

Seguendo l'esempio, un contratto, per quanto possa essere persino un pezzo unico, non origina da se stesso perché non può uscire dal recinto delle norme che regolano quella materia. Norme che il contratto potrebbe anche evitare di citare, ma non può evitare di rispettare, nella sostanza e nella forma linguistica.

Pertanto, un addestramento che abbia lo scopo di specializzare il CSNU all'elaborazione di bozze di contratti deve basarsi su un dataset composto anche dalle fonti primarie. A questo punto la risposta elaborata dal CSNU sarà molto più affidabile perché il recinto culturale è più ampio. È però ancora presto per valutare il risultato dei progetti di addestramento.

Il rango delle fonti è essenziale per l'affidabilità dell'informazione generata. Solo fonti affidabili rendono vera la regola della linguistica computazionale "bigger data, better data". Il machine learning fatto di quantità indiscriminate di dati (garbage in) produce risultati inaffidabili (garbage out).

Note

1) Tom Brown e altri, Language Models are Few-Shot Learners, 22 luglio 2020. Una sintesi in Dannis Layton, ChatGPT – Show me the Data Sources, 30 gennaio 2023.