2. Settembre

Non chiamatele allucinazioni: sono semplicemente stronzate

(parte prima: le perplessità)

La perplessità è l'inizio della conoscenza.
Khalil Gibran

grandi modelli linguistici stanno diventando sempre più efficaci nel sostenere conversazioni convincenti. Il modello linguistico più importante è ChatGPT di OpenAI, quindi è quello su cui ci concentreremo; tuttavia, ciò che diciamo si applica ad altri chatbot AI basati su reti neurali, tra cui **il chatbot Bard di Google, Claude (claude.ai) di AnthropicAI e LLaMa di Meta.**

Nonostante siano semplicemente dei software complicati, questi modelli sono sorprendentemente simili a quelli umani quando discutono di un'ampia varietà di argomenti.

Chiunque può andare all'interfaccia web di **OpenAI** e chiedere una risma di testo; in genere, produce un testo che è indistinguibile da quello di un normale parlante o scrittore inglese. La varietà, la lunghezza e la somiglianza con il testo generato dall'uomo di cui è capace GPT-4 hanno convinto molti commentatori a pensare che questo chatbot abbia finalmente decifrato il tutto: che questa sia una vera intelligenza artificiale (in contrapposizione a una mera intelligenza nominale), un passo più vicino a una mente simile a quella umana ospitata in un cervello di silicio.

Tuttavia, i grandi modelli linguistici e altri modelli di intelligenza artificiale come **ChatGPT** fanno molto meno di quanto faccia il cervello umano e non è chiaro se facciano ciò che fanno nello stesso modo in cui lo facciamo noi.

La differenza più ovvia tra un LLM e una **mente umana** riguarda gli *obiettivi* del sistema.

Gli esseri umani hanno una varietà di obiettivi e comportamenti, la maggior parte dei quali sono extra-linguistici: abbiamo desideri fisici di base, per cose come cibo e sostentamento; abbiamo obiettivi e relazioni sociali; abbiamo progetti; e creiamo oggetti fisici.

I grandi modelli linguistici mirano semplicemente a replicare il parlato o la scrittura umana. Ciò significa che il loro obiettivo primario, nella misura in cui ne hanno uno, è produrre un testo simile a quello umano. Lo fanno stimando la probabilità che una particolare parola appaia dopo, dato il testo che è venuto prima.

La macchina fa questo costruendo un ***modello statistico massiccio***, basato su grandi quantità di testo, per lo più tratto da Internet. Ciò avviene con relativamente poco input da parte di ricercatori umani o dei progettisti del sistema; piuttosto, il modello è progettato costruendo un gran numero di nodi, che agiscono come funzioni di probabilità per una parola di apparire in un testo dato il suo contesto e il testo che l'ha preceduta. Invece di inserire queste funzioni di probabilità manualmente, i ricercatori alimentano il sistema con grandi quantità di testo e lo addestrano facendogli fare previsioni sulla parola successiva su questi dati di addestramento. Quindi gli danno un ***feedback positivo o negativo*** a seconda che preveda correttamente. Dato abbastanza testo, la macchina può costruire da sola un modello statistico che fornisce la probabilità della parola successiva in un blocco di testo.

Questo modello associa a ogni parola ***un vettore*** che la colloca in uno spazio astratto ad alta dimensione, vicino ad altre parole che si verificano in contesti simili e lontano da quelle che non lo fanno. Quando produce testo, esamina la stringa di parole precedente e costruisce un vettore diverso, localizzando l'ambiente circostante la parola, il suo contesto, vicino a quelli che si verificano nel contesto di parole simili. Possiamo pensare a questi euristicamente come rappresentanti il significato della parola e il contenuto del suo contesto. Ma poiché questi spazi sono costruiti utilizzando l'apprendimento automatico tramite ripetute analisi statistiche di grandi quantità di testo, **non possiamo sapere** quali tipi di somiglianza sono rappresentati dalle dimensioni di questo spazio vettoriale ad alta dimensione.

Quindi non sappiamo quanto siano simili
a ciò che pensiamo come significato o contesto.

Il modello prende quindi questi due vettori e produce un insieme di probabilità per la parola successiva; seleziona e colloca una di quelle più probabili, anche se non sempre la più probabile. Consentire al modello di scegliere casualmente tra le parole più probabili produce un testo più creativo e simile a quello umano; il parametro che controlla questo è chiamato "**temperatura" del modello** e aumentare la temperatura del modello lo fa sembrare più creativo e più incline a produrre falsità. Il sistema ripete quindi il processo finché non ha una risposta riconoscibile e dall'aspetto completo a qualsiasi richiesta gli sia stata data.

Dato questo processo, non sorprende che gli LLM abbiano **un problema con la verità**. Il loro obiettivo è fornire una risposta apparentemente normale a un prompt, non trasmettere informazioni utili al loro interlocutore.

Esempi di ciò sono già numerosi, ad esempio, un avvocato ha recentemente preparato il suo briefing usando ChatGPT e ha scoperto con suo disappunto che la maggior parte dei casi citati non erano reali come ha affermato il giudice P. Kevin Castel, ChatGPT ha prodotto un testo pieno di "false decisioni giudiziarie, con false citazioni e false citazioni interne".

Allo stesso modo, quando i ricercatori di informatica hanno testato la capacità di ChatGPT di assistere nella scrittura accademica, hanno scoperto che era in grado di produrre un testo sorprendentemente completo e talvolta persino accurato su argomenti biologici, dati i giusti prompt. Ma quando è stato chiesto di produrre prove a sostegno delle sue affermazioni, "ha fornito cinque riferimenti risalenti ai primi anni del 2000. Nessuno dei titoli dei documenti forniti esisteva e tutti gli ID PubMed (PMID) forniti erano di documenti diversi e non correlati"

Alkaissi, H., & McFarlane, S. I., (2023, February 19). Artificial hallucinations in ChatGPT: Implications in scientific writing. Cureus, 15(2), e35179.

Questi errori possono "fare una valanga": quando al modello linguistico viene chiesto di fornire prove o una spiegazione più approfondita di un'affermazione falsa, raramente si controlla; invece produce con sicurezza affermazioni più false ma che suonano normali.

Il problema di accuratezza per LLM e altri AI generativi è spesso definito il problema dell'"allucinazione dell'IA": il chatbot sembra allucinare fonti e fatti che non esistono. Queste imprecisioni sono definite "allucinazioni" sia in contesti tecnici che popolari

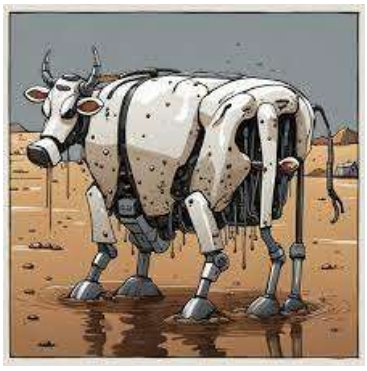
Questi errori sono piuttosto minori se l'unico scopo di un chatbot è imitare il linguaggio o la comunicazione umana. Ma le aziende che progettano e utilizzano questi bot hanno piani più grandiosi: i chatbot potrebbero sostituire le ricerche di Google o Bing con un'interfaccia conversazionale più intuitiva, o assistere medici o terapisti in contesti medici In questi casi, l'accuratezza è importante e gli errori rappresentano un problema serio.

Un tentativo di soluzione è quello di collegare il chatbot a una sorta di database, motore di ricerca o programma di calcolo in grado di rispondere alle domande a cui l'LLM sbaglia Sfortunatamente, anche questo non funziona molto bene. Ad esempio, quando ChatGPT è collegato a Wolfram Alpha, un potente software matematico, migliora moderatamente nel rispondere a semplici domande matematiche. Ma continua regolarmente a sbagliare, soprattutto per le domande che richiedono un pensiero in più fasi E quando sono collegati a motori di ricerca o altri database, è abbastanza probabile che i modelli forniscano informazioni false a meno che non ricevano istruzioni molto specifiche, e anche in quel caso le cose non sono perfette OpenAI ha in programma di correggere questo problema addestrando il modello a svolgere un ragionamento passo dopo passo (, ma ciò richiede molte risorse e c'è motivo di dubitare che risolverà completamente il problema; non è nemmeno chiaro se il risultato sarà un modello linguistico di grandi dimensioni, piuttosto che una forma più ampia di intelligenza artificiale.

In sintesi:
Il problema qui non è che i grandi modelli linguistici allucinano, mentono o travisano il mondo in qualche modo. È che non sono progettati per rappresentare il mondo in alcun modo; al contrario, sono progettati per trasmettere linee di testo convincenti.

Quindi, quando viene loro fornito un database di qualche tipo, lo usano, in un modo o nell'altro, per rendere le loro risposte più convincenti. Ma non stanno in alcun modo tentando di trasmettere o trasmettere le informazioni nel database.

Peraltro niente nella progettazione dei modelli linguistici (il cui compito di addestramento è prevedere le parole in base al contesto) è in realtà progettato per gestire l'aritmetica, il ragionamento temporale, ecc. Nella misura in cui a volte ottengono la risposta giusta a tali domande è solo perché sono riusciti a sintetizzare stringhe rilevanti da ciò che era nei loro dati di addestramento. Non è coinvolto alcun ragionamento



Allo stesso modo, i modelli linguistici sono inclini a inventare cose perché non sono progettati per esprimere un insieme di informazioni sottostanti in linguaggio naturale; stanno solo manipolando la forma del linguaggio” Questi modelli non sono progettati per trasmettere informazioni, quindi non dovremmo sorprenderci troppo quando le loro affermazioni si rivelano false.

Difendiamo l'ARTICOLO 34 della Costituzione

Prima che la politica privatizzi sanità e facoltà di Medicina
su modelli aziendali statunitensi

L'istruzione inferiore, impartita per almeno otto anni, è obbligatoria e gratuita.
I capaci e meritevoli, anche se privi di mezzi, hanno diritto di raggiungere i gradi più alti degli studi.
La Repubblica rende effettivo questo diritto con borse di studio,
assegni alle famiglie ed altre provvidenze, che devono essere attribuite per concorso.
Articolo 34

Con un deficit previsto di almeno **86.000** medici negli Stati Uniti entro il 2036, il debito degli studenti di medicina sta attirando sempre più attenzione. L'anno scorso, gli studenti di medicina negli Stati Uniti si sono laureati con un debito medio di **206.924 dollari**, una cifra che per molti interessati a diventare medici incide sulla decisione di specializzarsi o meno.
Premesso che un anno di frequenza in una Facoltà di Medicina Statunitense ha un costo che oscilla dai **120.000 ai 600.000 dollari** a seconda delle Facoltà, gli studenti o le loro famiglie devono quasi sempre contrarre un debito.

I dati **dell'Association of American Medical Colleges** mostrano che il **70%** degli studenti della classe del 2023 si è laureato con un debito scolastico. Di questi, **l'84%** ha contratto un debito di oltre **$ 100.000.** Il **54%** si è laureato con almeno **$ 200.000** di debito.

In alcune facoltà , gli studenti contraggono più di **$270.000** di debiti. Una ripartizione delle scuole di medicina in cui i laureati contraggono il debito medio più alto e più basso è stato riportato pre cedentemente su ***baedeker***

Le donazioni di grandi donatori alle scuole di medicina restano relativamente rare e quindi attraggono l'interesse pubblico quando si verificano. Più di recente, **Bloomberg Philanthropies** ha annunciato una donazione di **1 miliardo di dollari** alla **Johns Hopkins University di Baltimora**, rinunciando alle tasse universitarie per gli studenti provenienti da famiglie che guadagnano meno di ***300.000 dollari*** all'anno.

Le consistenti donazioni alle scuole di medicina che consentono alle università di rinunciare alle tasse universitarie sono destinate sia a gettare una rete più ampia per i futuri medici, sia ad ampliare l'ampiezza delle loro considerazioni sulla specializzazione, hanno detto i leader a ***Becker's*** a luglio, rispondendo alla notizia della *donazione di Bloomberg*. Ad esempio, se gli studenti entrano alla facoltà di medicina sapendo che non avranno centinaia di migliaia di dollari di debiti per prestiti, potrebbero scegliere di dedicarsi all'assistenza primaria anziché a una specializzazione più remunerativa.