

25.Febbraio

Aigiarismo

*La cosa triste dell'intelligenza artificiale è che manca di artificio
e quindi di intelligenza.*

Alan Perlis

pioniere dell'informatica e dei primi linguaggi di programmazione

Il dato inquietante ed allarmante è che **ChatGPT** può produrre rapporti, saggi e manoscritti scientifici fluenti e apparentemente ben informati. Preoccupati per l'etica e l'accuratezza di tali contenuti, **Patrick Franzen**, direttore editoriale di *SPIE la società internazionale per l'ottica e la fotonica*



insieme ai dirigenti di altre riviste si stanno dando da fare per *proteggere* la letteratura accademica da una potenziale marea di manoscritti scritti in tutto o in parte da programmi per computer.

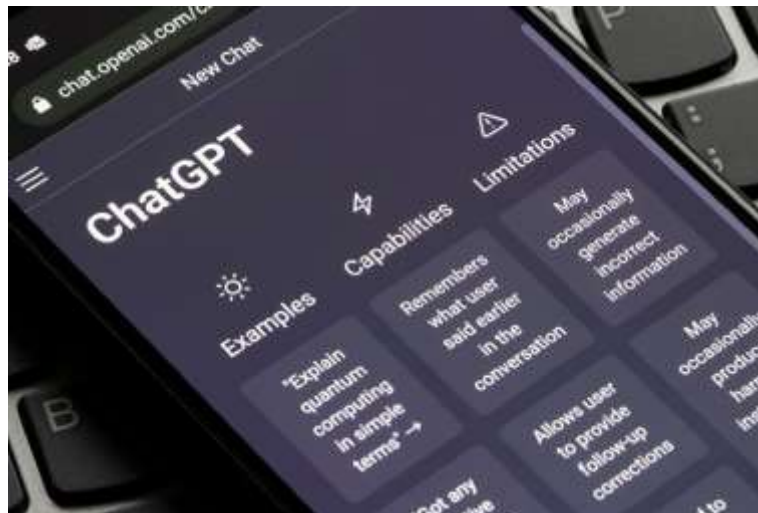
Anche se gli editori non hanno ancora formulato una *policy comune*, la maggior parte di quelli che hanno evitato un divieto assoluto del testo generato dall'intelligenza artificiale, chiedono agli autori di rivelare il loro uso degli strumenti automatizzati.

Per ora, editori e revisori hanno poche alternative, in quanto mancano di strumenti di applicazione.

Nessun software finora è in grado di rilevare costantemente il testo sintetico la maggior parte delle volte.

Quando lo strumento online **ChatGPT** è stato reso disponibile per uso pubblico gratuito, gli scienziati erano tra coloro che si sono accalcati per provarlo. (*Il creatore di ChatGPT, la società con sede negli Stati Uniti OpenAI, da allora ha limitato l'accesso agli abbonati.*)

Molti hanno segnalato la sua capacità senza precedenti e inquietante di creare testo dal suono plausibile, denso di dettagli *apparentemente* reali.



ChatGPT e i suoi “fratelli” , tra cui **Bard di Google**, presentato all'inizio di questo mese per utenti selezionati, e **Galactica di Meta**, che è stato brevemente disponibile per l'uso pubblico nel novembre 2022, sono algoritmi **AI** chiamati modelli di linguaggio di grandi dimensioni, addestrati su un vasto numero di campioni di testo estratti dal Internet. Il software identifica modelli e relazioni tra le parole, il che consente ai modelli di generare risposte pertinenti a domande e suggerimenti.

In alcuni casi, il testo risultante è **indistinguibile** da ciò che le persone scriverebbero come riportato nella **prestampa del 2022** :

Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers

In cui ricercatori che hanno letto gli abstracts di riviste mediche generati da **Chat GPT** non sono riusciti a identificare che uno su tre era stato scritto da una macchina

Ci si aspetta che gli sviluppatori di intelligenza artificiale creino versioni ancora più potenti, comprese quelle addestrate specificamente sulla letteratura scientifica, una prospettiva che ha inviato un'onda d'urto nel settore dell'editoria accademica.

Finora, gli scienziati riferiscono di aver “giocato” con **ChatGPT** per esplorarne le capacità e alcuni hanno elencato **ChatGPT** come coautore di manoscritti. Gli esperti di editoria temono che un uso così limitato possa trasformarsi in un picco di manoscritti contenenti parti sostanziali di testo scritto dall'IA.

La principale preoccupazione per i gestori di riviste è *l'accuratezza*. Se il software non è stato esposto a *dati di addestramento* sufficienti per generare una risposta corretta, spesso produrrà una risposta anomala.

A novembre, **Meta** ha rimosso l'interfaccia pubblica di **Galactica**, il suo modello di linguaggio di grandi dimensioni specifico per gli scienziati, pochi giorni dopo la sua presentazione: gli utenti avevano identificato una miriade di errori fattuali nel testo generato.

E uno studio [preliminare del 2022](#):

Improving alignment of dialogue agents via targeted human judgements

uno chatbot in cerca di normalizzazioni sviluppato da una consociata di Google ha rilevato che fino al **20%** delle risposte conteneva errori.

Gli editori di riviste si preoccupano anche dell'etica, suggerendo che gli autori che usano i *generatori di testo* a volte presentano i risultati come se li scrivessero loro, una trasgressione che altri hanno soprannominato "**aigiarismo**".

Le recenti politiche di molte riviste richiedono che gli autori rivelino l'uso di strumenti di generazione di testo e vietino l'elenco di un modello linguistico di grandi dimensioni come **ChatGPT** come coautore, per sottolineare la responsabilità dell'autore umano nel garantire l'accuratezza del testo.

Nature e di tutte le riviste **Sprinter Nature** e della rete **JAMA** ha denunciato come l'intelligenza artificiale (AI) sta acquisendo la capacità di generare un linguaggio fluente, sfornando frasi sempre più difficili da distinguere dal testo scritto dalle persone. L'anno scorso, **Nature** ha riferito che alcuni scienziati stavano già utilizzando i chatbot come assistenti di ricerca, per aiutare a organizzare il loro pensiero, generare feedback sul loro lavoro, assistere con la scrittura del codice e riassumere la letteratura di ricerca ([Nature 611, 192–193; 2022](#)).

Ma il rilascio del chatbot AI **ChatGPT** a novembre ha portato le capacità di tali strumenti, noti come modelli di linguaggio di grandi dimensioni (LLM), a un pubblico di massa.

I suoi sviluppatori, OpenAI a San Francisco, in California, hanno reso il chatbot gratuito e facilmente accessibile per le persone che non hanno competenze tecniche.

Attualmente milioni di persone lo stanno usando ([vedi allegato 1](#))

e il risultato è stato un'esplosione di esperimenti di scrittura *divertenti* e talvolta "spaventosi" che hanno messo il turbo alla crescente eccitazione e costernazione per questi strumenti

ChatGPT può scrivere saggi presentabili per studenti riassumere documenti di ricerca, rispondere a domande abbastanza bene da superare esami medici e generare utili codici informatici.

Ha prodotto *abstract* di ricerca abbastanza buoni che gli scienziati hanno trovato difficile individuare che un computer li avesse scritti.

Preoccupante per la società, potrebbe anche rendere più facile la produzione di *spam*, *ransomware* e altri *output* dannosi.

Sebbene **OpenAI** abbia cercato di mettere barriere di sicurezza su ciò che farà il chatbot, gli utenti stanno già trovando dei modi per aggirarli.

La grande preoccupazione nella comunità di ricerca è che *studenti e scienziati possano ingannevolmente far passare il testo scritto da LLM come proprio, o utilizzare LLM in modo semplicistico (come condurre una revisione della letteratura incompleta) e produrre un lavoro inaffidabile.*

Diversi preprint e articoli pubblicati Hanno già accreditato ChatGPT con “paternità formale”

E' giunto il momento che ricercatori ed editori stabiliscano regole di base sull'uso etico degli LLM. *Nature*, insieme a tutte le riviste *Springer Nature*, ha formulato i seguenti due principi:

-Nessuno strumento LLM sarà accettato come autore accreditato su un documento di ricerca. *Questo perché qualsiasi attribuzione di paternità porta con sé la responsabilità per il lavoro e gli strumenti di intelligenza artificiale non possono assumersi tale responsabilità.*

-In secondo luogo, i ricercatori che utilizzano gli strumenti LLM dovrebbero documentare questo utilizzo nelle sezioni dei metodi o dei ringraziamenti. *Se un documento non include queste sezioni, l'introduzione o un'altra sezione appropriata può essere utilizzata per documentare l'uso del LLM.*

Sulla stessa linea politica il **Committee on publication Ethics** e la **World Association of Medical Editors**

La famiglia di riviste *Science* ha annunciato il mese scorso un divieto totale del testo generato

Le riviste potrebbero allentare la politica in futuro a seconda di ciò che la comunità scientifica decide sia un uso accettabile dei generatori di testo, afferma il redattore capo **Holden Thorp**. "È molto più facile allentare i nostri criteri che restringerli".

Alcuni funzionari editoriali stanno ancora elaborando i dettagli, ad esempio quando potrebbero chiedere al personale della rivista, ai redattori o ai revisori di esaminare o verificare i fatti del testo generato divulgato dagli autori, compiti che si aggiungerebbero a quello che spesso è già un pesante carico di lavoro volontario.

Sabina Alam, *direttrice dell'etica e dell'integrità editoriale. del gruppo editoriale Taylor & Francis* fornisce supporto e guida a più di 2.500 riviste che coprono le discipline scientifiche, tecniche e mediche nonché le scienze umane e sociali.



Taylor & Francis eBooks

ha comunicato che la prossima regola richiederà probabilmente la divulgazione di tale testo, a volte potrebbero chiedere agli autori di specificare quali parti del loro manoscritto sono state scritte da un computer, afferma

La ricerca di documenti da includere in una revisione sistematica può essere un uso legittimo se il ricercatore segue i metodi appropriati nel decidere quali documenti includere, ad esempio, dice, mentre tagliarlo e incollarlo in una prospettiva o in un pezzo di opinione "non va bene perché è non la tua prospettiva." La politica probabilmente si evolverà man mano che l'industria editoriale acquisisce maggiore esperienza lavorando con tali manoscritti", afferma. "Stiamo vedendo questo come un approccio graduale. Sono davvero i primi giorni.

I gestori delle riviste affermano anche che sperano di monitorare la nuova tecnologia utilizzando più tecnologia: rilevatori automatici in grado di contrassegnare il testo sintetico. Ma non è facile, afferma **Domenic Rosati**, *ricercatore senior la cui ricerca si concentra sullo sviluppo di metodi di elaborazione del linguaggio naturale* per aiutare nella comprensione, valutazione e sintesi di documenti scientifici e artefatti linguistici presso la **sciteAI** un'azienda che sviluppa software per assistere gli editori scientifici



I rivelatori attuali lasciano molto a desiderare.

OpenAI ha svelato il suo "classificatore" il mese scorso, che classifica il testo inviato su una scala da "probabile" che sia stato scritto da un computer a "molto improbabile". Il classificatore si stava allenando utilizzando campioni accoppiati di scrittura umana e testo generato dal computer da 34 algoritmi di cinque diverse società, inclusa la stessa **OpenAI**.

Ma **OpenAI** ammette diverse limitazioni. Lo strumento, ancora in fase di sviluppo, applica correttamente l'etichetta "probabile" solo il 26% delle volte. Le persone possono ingannarlo modificando il testo generato dal computer. E potrebbe non identificare in modo coerente il testo sintetico su argomenti che non sono stati inclusi nei dati di addestramento. Gli informatici affermano che queste limitazioni si applicano in genere anche ad altri rilevatori.

Soluzioni migliori potrebbero essere all'orizzonte.

OpenAI ha dichiarato nel dicembre 2022 che sta lavorando a modi per "filigranare" il testo generato. Programmerebbe i suoi modelli per inserire parole, ortografia e punteggiatura in un ordine rivelatore per creare un codice segreto rilevabile dai motori di ricerca.

E il mese scorso un team della Stanford University ha pubblicato una pre stampa:

Detect GPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature

che a differenza di altri rilevatori non richiede formazione.

Questo algoritmo esamina il testo creando più variazioni casuali e interrogando un generatore di testo per classificare le versioni che preferisce.

La misura in cui il generatore studiato dal team, sviluppato da **OpenAI** e simile a **ChatGPT**, preferisce il testo originale rispetto alle versioni modificate è costantemente diverso per il testo scritto dall'uomo rispetto a quello generato dall'IA, consentendo a **DetectGPT** di prevedere la probabilità che provenga un campione una macchina particolare.

Ma **DetectGPT** ha bisogno di più sviluppo prima che gli editori di riviste possano fare affidamento esclusivamente sui suoi risultati per prendere decisioni sui manoscritti. Eric Mitchell, uno studente di dottorato che ha guidato il team di Stanford.

La società **TurnItIn** che commercializza un rilevatore di plagio ampiamente utilizzato, ha dichiarato la scorsa settimana che prevede di lanciare un rilevatore di testo sintetico già ad aprile. **TurnItIn** afferma che lo strumento, addestrato alla scrittura accademica,

Un'ulteriore sfida computazionale è valutare l'accuratezza fattuale del testo generato dal robot e la qualità dei suoi riassunti, *afferma Rosati*. La sua azienda sta lavorando a un controllo automatico che analizzerebbe la letteratura scientifica esistente per determinare se una particolare citazione in un manoscritto presenta effettivamente la scoperta che il manoscritto afferma di fare, per esempio. Quello strumento potrebbe rivelare riferimenti fabbricati da una macchina così come quelli irrilevanti o errati forniti dagli esseri umani.

Alcuni redattori vedono promesse e pericoli nell'emergere di **ChatGPT** e dei suoi parenti. I programmatori stanno sviluppando e perfezionando software che creano ipotesi, scrivono codice per computer e analizzano dati, ad esempio, per rendere il lavoro dei ricercatori più efficiente e produttivo.

Allegato: 1

A proposito di ChatGPT

ChatGpt è un esempio di nuova intelligenza artificiale, detta 'generativa' poiché è in grado di scrivere testi - o addirittura produrre immagini - così come farebbe un essere umano.

Come fa ChatGpt a scrivere come un uomo?

Attraverso i cosiddetti LLM o modelli linguistici (**vedi allegato 2**)

che nel campo della ricerca sull'IA indicano i modelli linguistici di grandi dimensioni, vale a dire algoritmi di deep learning addestrati generalmente su milioni di testi. In questo modo l'IA è capace di predire quale parola ha più possibilità che segua un'altra, dato un determinato contesto.

È gratuita?

Sì, ChatGpt è aperta al pubblico e si può usare gratuitamente. Esiste anche un piano di abbonamento, chiamato ChatGpt Plus, al prezzo di 20 dollari al mese.

Questa versione 'premium' è disponibile anche in Italia e garantisce alcuni vantaggi: una maggiore velocità di generazione del testo, per esempio, e un accesso assicurato e illimitato. Chi usa ChatGpt free, infatti, potrebbe avere difficoltà ad accedere quando i server sono particolarmente sotto pressione. E comunque, quando si effettuano troppe richieste in un'ora, la versione gratuita smette di funzionare per un po' di tempo.

Come si accede?

Bisogna andare [sul sito di Open AI](#), la startup di San Francisco che ha sviluppato questa IA, e cliccare in alto, al centro, la parola "Try", vale a dire "Provala".

A questo punto bisogna registrarsi. Fornendo una mail oppure scegliendo di proseguire con un account Microsoft o Google di cui siamo già in possesso.

Dopo la registrazione, si atterra sulla finestra di dialogo e si può iniziare a dialogare con l'IA: basterà fare una domanda o scrivere qualsiasi tipo di pensiero nel box in fondo alla pagina.

Nella colonna di sinistra vengono archiviate le nostre chat con l'IA. Se passate l'indicatore del mouse su ognuna di queste, vi accorgete che si possono anche rinominare oppure cancellare. È un'area importante, poiché se ben organizzata ci permette di accedere a ogni conversazione per recuperare informazioni preziose, oppure per riprendere il dialogo nel punto in cui l'abbiamo lasciato.

Quali sono i limiti di ChatGpt?

Il più importante riguarda l'allucinazione dei dati, vale a dire la capacità dell'IA di scrivere testi coerenti, plausibili e tuttavia con informazioni totalmente false. Insomma: la prima regola di ChatGpt è quella di non fidarsi mai, ciecamente, di ChatGpt.

In particolare questa IA sembra avere problemi con la matematica, ma anche con le dichiarazioni di persone reali, che spesso ChatGpt riprova come vere ma invece sono inventate.

Un altro limite è invece temporale: ChatGpt può esprimersi su persone, fatti e avvenimenti con una conoscenza che arriva a dicembre 2021. Tutto quello che è accaduto dopo, e che tuttora accade, lo ignora. Se le chiedete chi ha vinto l'ultima Coppa del Mondo di calcio, vi risponderà la Francia nel 2018.

Tra i numerosi limiti (**vedi allegato 3**) infine, c'è anche quello dei caratteri: non accetta domande più lunghe di 4096 caratteri. Anche le risposte non eccedono i 4096 caratteri.

Cosa si può fare con ChatGpt?

Praticamente tutto. Scrivere saggi, racconti, poesie, testi inediti di una canzone. Oppure semplicemente si può dialogare con l'IA alla ricerca di un'idea brillante, di informazioni su un luogo da visitare, di un podcast interessante da ascoltare (attenzione però, a volte inventa i link e andrete su una pagina che non esiste).

Ma si può anche chiedere all'IA di correggere un testo che abbiamo già scritto, oppure di rispondere a una mail che ci è stata inviata, o ancora - funzione preziosa - di controllare che il nostro scritto in inglese, o in una qualsiasi lingua che non sia la nostra, ha una forma corretta. Un'altra possibilità innovativa, che qualcuno potrà trovare utile, è riassumere il contenuto di un video YouTube senza doverlo guardare. Basterà copiare la trascrizione del video (disponibile per quasi tutti i video YouTube), incollarla su ChatGpt e chiederle di riassumere il testo.

Come si usa ChatGpt?

Può farlo chiunque, basta digitare una domanda o una frase con un linguaggio naturale. L'IA sarà in grado di processarlo, analizzarlo e di confezionare una risposta su misura, adeguata - nella maggior parte dei casi - a ciò che le è stato chiesto.

Quando riceviamo una risposta, si può fare una domanda ulteriore per approfondire, capire meglio. Oppure si può chiedere all'IA di riscrivere quel testo in modo diverso, magari in modo che sia comprensibile anche da un bambino di 10 anni. Le nostre domande, in gergo, si chiamano 'prompt'.



Che cos'è un prompt?

È l'insieme dei comandi (testuali) che un utente utilizza per chiedere a un'IA di scrivere un testo o creare un'immagine. Il prompt deve includere tutte le caratteristiche che guideranno l'IA a produrre il contenuto desiderato da un utente, proprio come questo l'aveva immaginato nella sua testa. I prompt sono i nuovi codici universali - al di là di ChatGpt - che guideranno le IA in futuro. Il bello è che non richiedono una preparazione particolare, oppure lo studio di manuali voluminosi: tutto ciò che serve è una buona dose di immaginazione.

Come è fatto un prompt?

Scrivere un prompt, in genere, vuol dire fare una domanda in modo discorsivo. E dunque, se per esempio siamo interessati a tecnologia e ambiente, possiamo chiedere all'IA di rispondere anche a

domande complesse come queste:

Quali sono gli effetti positivi e negativi dell'utilizzo di smartphone sulla vita delle persone?

oppure

Quali sono le principali cause del cambiamento climatico e quali sono le possibili soluzioni per combatterlo?

A ogni domanda si possono aggiungere delle istruzioni relative al tono o allo stile che l'IA deve usare, oppure relative alla lunghezza massima del testo che desideriamo.

Per esempio:

Quali sono le migliori strategie per motivare gli studenti a impegnarsi di più nella scuola?

Scrivi la risposta con un tono informale, come se fosse un docente a spiegarlo, usando al massimo 400 caratteri.

Ed ecco la risposta che ha dato ChatGpt:

“Ciao! Ci sono molte strategie per motivare gli studenti a impegnarsi di più nella scuola. Innanzitutto, è importante creare un ambiente accogliente e positivo in classe. Inoltre, cercare di rendere le lezioni interessanti e coinvolgenti, ad esempio attraverso l'uso di attività pratiche e materiali didattici innovativi. Dare anche opportunità per la partecipazione attiva degli studenti, come discutere e condividere idee con i loro compagni di classe. Infine, sottolineare l'importanza dell'apprendimento e delle competenze acquisite, e come questi possono avere un impatto sulla loro vita futura”

In che lingua si può scrivere?

Si possono usare moltissime lingue, compreso l'italiano. Ma per stessa ammissione di ChatGpt i risultati migliori si ottengono scrivendo in inglese. Anche con l'italiano, tuttavia, si ottengono risposte sorprendenti. Il caso in cui si scelga di scrivere in inglese, si può sempre chiedere a ChatGpt di tradurre in italiano l'ultima risposta che ci ha fornito.

Si può fare “copia e incolla”?

ChatGpt non è ancora in grado, al momento, di sostituire un essere umano. Per i limiti che abbiamo visto, ma anche per una creatività che - sebbene stupisca - non è equivalente a quella di una persona in carne e ossa. Le sue risposte, inoltre, possono essere banali e superficiali. Per ottenere il meglio da ChatGpt, occorre fare delle domande mirate, ricche di dettagli, fantasiose già in partenza. In generale, i testi che si ottengono da ChatGpt vanno comunque lavorati, rimaneggiati, controllati, affinati. Chi è alla ricerca di un facile “copia e incolla” per i compiti scolastici, o per risolvere un problema sul lavoro, potrebbe restare deluso. ChatGpt resta, invece, un ottimo strumento che può aiutare e migliorare la nostra produttività, suggerendo spunti e punti di vista che magari non avevamo preso in considerazione.

ChatGpt e Bing sono la stessa cosa?

Sebbene la tecnologia alla base delle due chatbot sia la stessa, ChatGpt e [il nuovo Bing](#) - che usa l'intelligenza artificiale generativa per fornire le sue risposte - non sono la stessa cosa.

E non solo perché sono accessibili su piattaforme differenti. Microsoft, che in Open AI ha investito circa 10 miliardi di dollari, utilizza per Bing il modello di linguaggio di grandi dimensioni Gpt 3.5 (anche se qualcuno ipotizza sia più evoluto) che è alla base anche di ChatGpt. Ma su Bing, a differenza di ChatGpt, le risposte alle domande includono anche delle reference, vale a dire delle note che portano alla fonte utilizzata dall'IA per mettere insieme la sua risposta.

Questo non vuol dire che le risposte di Bing siano più accurate. Anche il motore di ricerca di Microsoft ha commesso errori grossolani nelle prime settimane in cui è stato accessibile al pubblico. L'accesso al nuovo Bing, infine, è più complicato: bisogna iscriversi a una lista di attesa

Esiste un modo per capire se un testo è stato scritto da un IA?

Sì, esiste ma al momento non sono molto affidabili. Uno strumento, in tal senso, è stato sviluppato proprio da Open AI: si chiama [Classifier](#) e fornisce risultati affidabili al 26%.

Un altro strumento [GptZero](#), promette di individuare i testi scritti da una IA ma non sempre ci riesce (stando ai test che abbiamo fatto).

Allegato 2

Modelli linguistici

Un modello linguistico è una distribuzione di probabilità su sequenze di parole. Data qualsiasi sequenza di parole di lunghezza m , un modello linguistico assegna una probabilità a tutta la sequenza. I modelli linguistici generano probabilità addestrandosi su corpora testuali in una o più lingue. Dato che le lingue possono essere utilizzate per esprimere un'infinita varietà di frasi valide (proprietà dell'infinito digitale), la modellazione linguistica affronta il problema di assegnare probabilità diverse da zero a sequenze linguisticamente valide che potrebbero non essere mai incontrate nei dati di addestramento. Diversi approcci di modellazione sono stati progettati per superare questo problema, come l'applicazione dell'ipotesi di Markov o l'utilizzo di architetture neurali come reti neurali ricorrenti o trasformatori. I modelli linguistici sono utili per una varietà di problemi nella linguistica computazionale; dalle applicazioni iniziali nel riconoscimento vocale per garantire che le sequenze di parole prive di senso (cioè a bassa probabilità) non siano previste, all'uso più ampio nella traduzione automatica (ad es.), tagging di parti del discorso, analisi, riconoscimento ottico dei caratteri, riconoscimento della grafia, induzione grammaticale, recupero di informazioni e altre applicazioni. I modelli linguistici vengono utilizzati nel recupero delle informazioni nel modello di verosimiglianza della query. Lì, un modello linguistico separato è associato a ciascun documento in una raccolta. I documenti sono classificati in base alla probabilità della query nel modello linguistico del documento. Comunemente, il modello di linguaggio unigramma viene utilizzato per questo scopo.

Allegato 3

I limiti di Chat GP

ChatGPT a volte scrive risposte plausibili ma errate o prive di senso. Risolvere questo problema è impegnativo, in quanto:

- (1) durante l'addestramento RL, al momento non esiste alcuna fonte di verità;
- (2) addestrare il modello a essere più cauto fa sì che rifiuti le domande a cui può rispondere correttamente; e
- (3) l'addestramento supervisionato inganna il modello perché la risposta ideale dipende da ciò che sa il modello, piuttosto che da ciò che sa il dimostratore umano.

ChatGPT è sensibile alle modifiche al fraseggio di input o al tentativo di ripetere la stessa richiesta più volte. Ad esempio, data una frase di una domanda, il modello può affermare di non conoscere la risposta, ma data una leggera riformulazione, può rispondere correttamente.

Il modello è spesso eccessivamente prolisso e abusa di determinate frasi, ad esempio affermando che si tratta di un modello linguistico addestrato da OpenAI. Questi problemi derivano da pregiudizi nei dati di addestramento (i formatori preferiscono risposte più lunghe che sembrano più complete) e problemi ben noti di ottimizzazione eccessiva. Idealmente, il modello dovrebbe porre domande di chiarimento quando l'utente ha fornito una query ambigua. Invece, i nostri modelli attuali di solito indovinanano cosa intendeva l'utente. Sebbene il sistema cerca di rifiutare le richieste inappropriate, a volte risponde a istruzioni dannose o mostra un comportamento di parte

Un anno fa... Baedeker/Replay del 21 Febbraio 2022

Tranquilli! la sottovariante Omicron BA.2 non è poi così virulenta come si temeva

Sicurezza significa sapere ciò che sta per succedere. Sicurezza significa niente emozioni, niente rischi, niente sfide. Sicurezza significa niente crescita, e niente crescita vuol dire morte. A parte il fatto che la sicurezza è un mito. Finché siamo su questa terra, e il sistema dei preprint resta qual è, la sicurezza nella verità di un dato scientifico non l'avremo mai, come potrai vedere alla fine di questo report. Anche se i casi di COVID-19 stanno diminuendo in tutto il mondo, BA.2 continua a crescere in proporzione ai casi per i quali sono disponibili sequenze virali aggiornate: all'inizio di questo mese rappresentava il 21% dei casi sequenziati in tutto il mondo. Questa sottovariante è arrivata a dominare in Danimarca, India e altri paesi. In Sud Africa, è cresciuta dal 27% all'86% dei casi sequenziati tra il 4 e l'11 febbraio. Negli Stati Uniti la sua prevalenza è triplicata, al 3,6%, tra la fine di gennaio e il 5 febbraio. Entro la settimana, terminata il 19 febbraio, era del 3,8%. I dati epidemiologici pubblicati il mese scorso, stimano che BA.2 sia fino al 40% più contagioso di BA.1

A gettare benzina sul fuoco della infettività di BA.2 è stato un preprint di Kei Sato dell'Università di Tokyo, "Virological characteristics of SARS-CoV-2 BA.2 variant" che dimostrava come la variante BA.2 rendeva i criceti più malati rispetto a BA.1 e di conseguenza questa variante potrebbe essere più patogena negli esseri umani e più trasmissibile nella popolazione umana.

L'Organizzazione Mondiale della Sanità (OMS) ieri, 23 febbraio, ha dichiarato che la variante SARSCoV-2 BA.2 ritenuta una versione più contagiosa di Omicron, non rende le persone più malate dell'Omicron originale al punto tale da non meritare una designazione separata con una propria lettera greca. La dichiarazione dell'OMS è arrivata a seguito di nuovi preprint (non sottoposti a revisione paritaria) sulla propagazione di BA.2 provenienti da Sud Africa e Danimarca. L'OMS ha anche citato dati ancora riservati provenienti dal Regno Unito, che dovrebbe rilasciare un nuovo rapporto informativo su Omicron venerdì 25 (cioè oggi per chi legge) La dichiarazione di ieri, dai consulenti dell'OMS che seguono l'evoluzione di SARS-CoV-2, ha affermato che la combinazione dei risultati del "mondo reale" di tutti e tre i paesi non ha mostrato alcuna prova che la sottovariante, BA.2, causi una malattia più grave di BA.1. Tuttavia, in forma cautelativa, la variante BA.2 dovrebbe continuare a essere monitorata da altri paesi, ma dovrebbe rimanere classificata ancora come Omicron". Alcuni scienziati hanno sostenuto che le differenze tra le due varianti sono abbastanza significative che BA.2, che presenta circa 50 mutazioni che la distinguono dal primo ceppo pandemico di SARSCoV-2, di cui circa 30 condivise con BA.1, dovrebbero essere rubricata come una variante indipendente di preoccupazione. E' importante ricordare che BA.1 possiede anche ulteriori mutazioni che non divide con BA.2; il rapporto netto tra i due ceppi è di circa 40 mutazioni. Il gruppo di lavoro del OMS che ha stimato la gravità clinica dei casi di COVID-19 in Sud Africa, Regno Unito e Danimarca, afferma che dai dati disponibili non emerge alcuna differenza di gravità tra BA.2 e BA.1" e che l'infezione da BA.1 fornisce una forte protezione contro la reinfezione da BA.2, "almeno per il periodo limitato per il quale i dati sono disponibili" In Sud Africa, i ricercatori del National Institute for Communicable Diseases (NICD) hanno recentemente analizzato i dati di oltre 95.000 persone che sono state infettate da SARS-CoV-2 in Sud Africa tra il 1° dicembre 2021 e la fine di gennaio. I ricercatori hanno riferito in un preprint del 19 febbraio di non aver riscontrato un aumento del rischio di ricovero tra coloro i cui test diagnostici indicavano di avere BA.1 e quelli i cui test indicavano BA.2. Né hanno riscontrato un rischio più elevato di malattia grave da BA.2 rispetto a BA.1 in oltre 3000 dei ricoverati in ospedale. Ritengono tuttavia la necessità di dati aggiuntivi registrati da altri paesi in cui SARS-CoV2 non ha infettato una fascia così ampia della popolazione. Tali dati in realtà esistono e sono stati resi pubblici il 22 febbraio in un preprint dallo Statens Serum Institute l'agenzia per la salute pubblica danese, che ha analizzato le reinfezioni registrate nel paese dalla fine di novembre 2021 all'11 febbraio, un periodo di tempo durante il quale BA.2 è diventata dominante. Hanno scoperto che le persone erano raramente reinfezzate con BA.2 subito dopo un'infezione da BA.1 e che, quando si verificavano, queste reinfezioni erano lievi, con conseguente assenza di ricoveri o malattie gravi. L'89% delle persone reinfezzate non erano vaccinate e nessuna aveva più di 38 anni. Anche le differenze tra i sintomi dell'infezione iniziale e della seconda erano

"trascurabili. L'Agenzia per la sicurezza sanitaria del Regno Unito dovrebbe il 25 febbraio pubblicare ulteriori dati del "mondo reale" su BA.2. In Inghilterra, dove i nuovi casi giornalieri stanno diminuendo precipitosamente e anche i ricoveri sono in calo, BA.2 ha rappresentato il 13% dei casi sequenziati tra il 1° febbraio e il 7 febbraio.

In conclusione: in questa alternanza di preprint sembrerebbe che i dati sull'aumento della gravità della malattia da BA.2 nei criceti non siano coerenti con i recenti risultati epidemiologici osservabili nel "mondo reale". I modelli animali sono utili solo se riflettono la realtà e solo se i relativi preprint vengono validati, e confermati. Il compito a cui dobbiamo lavorare, non è di arrivare alla sicurezza, ma di arrivare a tollerare almeno l'insicurezza