ChatGPT: amico o nemico?

È stato difficile non accorgersi della tempesta che ha suscitato il sito Chat Generative Pre-trained Transformer (ChatGPT) negli ultimi mesi. Le testate giornalistiche e i social media sono stati tempestati di notizie sulla chatbot sviluppata da OpenAI (*la chatbot è un software che simula ed elabora le conversazioni umane scritte o parlate* ndr). In risposta a una richiesta scritta, ChatGPT è in grado di comporre e- mail, scrivere codici informatici e persino creare sceneggiature di film. I ricercatori hanno anche dimostrato che è in grado di superare esami di abilitazione alla professione medica. Ma l'entusiasmo è stato accompagnato da una serie di preoccupazioni etiche che potrebbero - e forse dovrebbero - limitarne l'adozione.

ChatGPT si basa su una versione raffinata del modello linguistico GPT-3.5. Il modello di base GPT-3 è stato addestrato su articoli, siti web, libri e conversazioni scritte ma un processo di perfezionamento (compresa l'ottimizzazione per il dialogo) consente a ChatGPT di rispondere alle richieste in modo colloquiale.

Nel campo dell'assistenza sanitaria, Sajan B Patel e Kyle Lam hanno illustrato la capacità di ChatGPT di generare un riassunto della dimissione del paziente a partire da una breve richiesta. L'automazione di questo processo potrebbe ridurre i ritardi nelle dimissioni dall'assistenza secondaria senza compromettere i dettagli, liberando tempo prezioso per i medici da investire nella cura dei pazienti e nella formazione continua. Uno studio separato ha anche testato la capacità di semplificare i referti radiologici: i referti generati sono stati ritenuti nel complesso corretti, completi e con un basso rischio percepito in quanto alla possibilità di creare danni per i pazienti. In entrambi i casi, però, si manifestavano erano evidenti. Nell'esempio di riepilogo della dimissione fornito da Patel e Lam, ChatGPT ha aggiunto al riepilogo informazioni extra che non erano incluse nelle loro richieste. Allo stesso modo, lo studio sui referti radiologici ha identificato errori potenzialmente dannosi, come la mancanza di reperti fondamentali per lo studio dell'immagine. Tali errori ci forniscono l'evidenza che, se tali metodi di applicazione dell'AI fossero implementati nella pratica clinica, sarebbe, comunque, necessario un controllo manuale degli output automatizzati.

I limiti di ChatGPT sono noti. Da OpenAI per sua stessa ammissione, l'output di ChatGPT può essere scorretto o parziale, ad esempio citando riferimenti ad articoli inesistenti o perpetuando stereotipi sessisti. Potrebbe anche rispondere dando istruzioni dannose che generano malware. OpenAI ha predisposto dei guardrail per minimizzare i rischi, ma gli utenti hanno trovato il modo di aggirarli e, poiché i risultati di ChatGPT potrebbero essere utilizzati per addestrare le future iterazioni del modello, questi errori potrebbero essere riciclati e amplificati. OpenAI ha chiesto agli utenti di segnalare le informazioni inappropriate

Le risposte di ChatGPT possono aiutare a migliorare il modello, ma ciò è stato criticato, in quanto spesso sono le persone colpite in modo sproporzionato dai pregiudizi dell'algoritmo (ad esempio quelle appartenenti a delle comunità marginali) a dover contribuire alla ricerca di soluzioni. Michael Liebrenz e colleghi sostengono che, sebbene ChatGPT possa servire a democratizzare la condivisione della conoscenza, in quanto è in grado di ricevere e produrre testi in più lingue (a vantaggio di chi non è madrelingua e pubblica in inglese), le imprecisioni nel testo generato potrebbero alimentare la diffusione della disinformazione.

Queste preoccupazioni hanno serie implicazioni riguardo all'integrità della documentazione scientifica, dato il rischio di introdurre nelle pubblicazioni non solo errori ma anche contenuti plagiati. Ciò potrebbe far sì che le future decisioni in materia di ricerca o di politica sanitaria vengano prese sulla base di informazioni false. Il mese scorso, la World Association of Medical Editors ha pubblicato le sue raccomandazioni sull'uso di ChatGPT e di altre chatbot nelle pubblicazioni accademiche, una delle quali prevede che i direttori delle riviste abbiano bisogno di nuovi strumenti per rilevare i contenuti generati o modificati dall'intelligenza artificiale (AI). In effetti, un rilevatore automatico di output dell'intelligenza artificiale ha dimostrato di saper distinguere meglio gli abstract di articoli di ricerca originali da quelli generati da ChatGPT rispetto a un rilevatore di plagio o a revisori umani, ma ha anche segnalato falsamente un abstract originale come "falso".

La tecnologia si evolve e le politiche editoriali devono forzatamente evolvere anch'essa. Elsevier ha introdotto una nuova politica sull'uso dell'IA e delle tecnologie assistite dall'IA nella scrittura scientifica, stabilendo che l'uso deve essere limitato al miglioramento della leggibilità e del linguaggio del lavoro, e deve essere dichiarato nel manoscritto; gli autori devono effettuare controlli manuali di qualsiasi prodotto generato dall'IA e questi strumenti non devono essere elencati o citati come autore o coautore, in quanto non possono assumersi le responsabilità che la paternità comporta (come la responsabilità del lavoro pubblicato).

L'uso diffuso della ChatGPT è apparentemente inevitabile, ma nella sua attuale versione l'uso incauto e non controllato potrebbe essere un nemico sia per la società che per l'editoria scientifica. È necessaria una maggiore attenzione e supervisione nell'addestramento dei modelli, così come un investimento in robusti rilevatori di output dell'IA. ChatGPT cambia le carte in tavola, ma non siamo ancora pronti a giocare.

Vedi l’articolo originale nel link per la bibliografia