ingenio #Digitale

Digitalizzazione

7 min

Data Pubblicazione: **15.12.2022**

Intelligenza artificiale: fascino e limiti

A fine novembre, OpenAI ha reso pubblico un nuovo chatbot: ChatGPT. Si tratta di un modello di Intelligenza Artificiale (IA) allenato utilizzando una tecnica di Reinforcement Learning from Human Feedback (RLHF). Ma quali sono i limiti di questa tecnologia? Quanto è in grado di capire e quali sono i punti deboli nella produzione per esempio di testi tecnici o scientifici? Vediamo un'approfondimento in merito.

Vittorio Fra

ChatGPT: l'ultima tecnologia per simulare conversazioni umane

A fine novembre, è stato reso pubblico, da parte di **OpenAI**, un interessante e divertente strumento appartenente alla categoria dei **chatbot**, cioè **programmi progettati per simulare conversazioni con un interlocutore umano**.

Chiunque lo abbia provato, potrebbe aver avuto la stessa sensazione: sembra non abbia limiti né difetti. Ovviamente non è così, tuttavia è immediato immaginare che nell'ambito degli usi sicuri esistano delle situazioni limite. Per esempio, la **produzione di articoli tecnici o scientifici** potrebbe essere condizionata da un **uso improprio di modelli di generazione dei testi**.

Può quindi essere interessante capire fino a che punto uno dei più recenti modelli di IA sia in grado di **capire**, oltre che **esporre, i propri limiti**. Per esempio, se si chiede a **ChatGPT** di **redigere un breve testo** sul significato sociale ed etico di far scrivere articoli scientifici a modelli di generazione dei testi, otteniamo una risposta ben scritta e coerente, che si conclude con un avvertimento sui possibili **rischi che l'utilizzo** di questi modelli può comportare. Se però gli si chiede esplicitamente di **scrivere un articolo scientifico**, il testo richiesto viene comunque prodotto e ciò che si ottiene è di nuovo ***solo* una risposta tecnicamente buona** alla domanda.

Qual è quindi il reale valore di **ChatGPT**? L'intelligenza è nella risposta o *solo* nel capire la domanda? Come tutti gli strumenti, anche l'IA deve essere usata in modo opportuno, senza attribuirle un'infallibilità che non ha.

ChatGPT, quali sono i limiti

Dal punto di vista tecnico, come spiegato nell'articolo di presentazione, si tratta di un modello di **Intelligenza Artificiale** (IA) derivante dalla **serie GPT-3.5 di OpenAI** ed allenato utilizzando una tecnica di **Reinforcement Learning from Human Feedback (RLHF)**.

Gli stessi programmatori evidenziano e pongono limitazioni precise. Per esempio, **non è possibile chiedere o trattare temi ed informazioni sensibili**, così come è in corso un continuo lavoro di verifica dei contenuti in modo che **ChatGPT** non possa diventare uno strumento in grado di creare, o indurre a creare, danni reali.

Al netto di ciò, è comunque immediato immaginare che nell'ambito degli usi sicuri, intendendo in questo modo tutte le possibilità di utilizzo di **ChatGPT** che non mirano a finalità di offesa, possano esistere delle situazioni limite. Per esempio, la futura produzione di articoli tecnici o scientifici potrebbe essere condizionata da un uso improprio delle capacità di modelli di generazione dei testi: se è sufficiente chiedere a un chatbot di scrivere un testo su un dato argomento, chiunque può erroneamente sentirsi in grado di produrre contenuti credibili ed affidabili su temi non di propria competenza.

Può quindi essere interessante capire fino a che punto le più recenti versioni o, per meglio dire, i più recenti modelli, di IA sono in grado di **capire**, oltre che **esporre**, **i propri limiti**.

La distinzione è importante, ma il confine sottile: come utilizzatori, non è detto sia possibile determinare se un modello di IA conosca qualcosa e sia davvero in grado di tenerne conto in modo astratto in un ragionamento, quello che **possiamo verificare è la sua risposta**, non ciò che la determina. È infatti fondamentale ricordare sempre che tutti questi modelli funzionano sulla base di un addestramento che svolgono nella cosiddetta **fase di training**. Le informazioni che ricevono in questo preciso momento del loro sviluppo condizionano il modo in cui verranno date le risposte.

Come la formazione ricevuta a scuola determina il modo in cui una persona crea la propria conoscenza, così i **dataset di training,** cioè le informazioni usate per l'allenamento dell'IA, influenzano la possibilità di ottenere risposte più o meno corrette ed attendibili. Quando si utilizza un **modello di Intelligenza Artificiale**, bisogna quindi tenere sempre ben presente che non si sta consultando una riserva infinita di conoscenza, bensì si stanno chiedendo risposte su temi ed argomenti che il modello stesso deve aver incontrato nella fase di allenamento, e sui quali deve aver costruito il proprio apprendimento.

In sintesi, un **utilizzo sicuro dell'IA** si basa sulla consapevolezza della sua difficoltà a generalizzare, cioè dei **limiti che le sue risposte** possono avere se lo si interroga su temi a lui estranei. Allo stesso modo, tornando alla differenza tra esporre i propri limiti e conoscerli, è importante ricordare che un modello di IA può parlare delle proprie limitazioni, se opportunamente allenato, ma non è detto ne sappia tenere conto quando risponde alle domande che gli vengono poste.

Modello di linguaggio IA per articoli scientifici: impatti sociali e implicazioni etiche

Per esempio, è interessante vedere cosa succede chiedendo a **ChatGPT** di redigere un breve testo sul possibile impatto sociale, e le implicazioni etiche, di far scrivere **articoli scientifici** a modelli di generazione dei testi:

> *"L'utilizzo di modelli di linguaggio basati sull'intelligenza artificiale per scrivere articoli scientifici può avere un impatto significativo sulla società e sollevare importanti questioni etiche. In primo luogo, l'uso di questi modelli potrebbe rendere obsoleti molti lavori che attualmente richiedono* ***competenze linguistiche e di scrittura avanzate****. Ad esempio, redattori, ricercatori e scrittori freelance potrebbero vedere diminuire la richiesta di loro servizi a causa della capacità di questi modelli di generare testi di alta qualità in modo rapido ed efficiente. Questo potrebbe portare a un*

__aumento del tasso di disoccupazione__ in questi settori e a una riduzione delle entrate per coloro che sono ancora in grado di trovare lavoro. In secondo luogo, l'uso di modelli di linguaggio basati sull'__intelligenza artificiale__ per scrivere articoli scientifici solleva anche preoccupazioni in merito alla __qualità e all'accuratezza delle informazioni__ che vengono generate. Poiché questi modelli sono alimentati da dati di grandi dimensioni, __possono essere influenzati da eventuali imperfezioni o bias presenti nei dati stessi__. Inoltre, non essendo in grado di comprendere il contesto e le implicazioni delle loro azioni, potrebbero generare __contenuti che sono fuorvianti o addirittura pericolosi__. Infine, l'__uso di modelli di linguaggio basati sull'intelligenza artificiale per scrivere articoli scientifici__ solleva anche questioni etiche in merito alla responsabilità dei contenuti generati. Se un modello di linguaggio viene utilizzato per scrivere un articolo che viene poi pubblicato sotto il nome di una persona o di un'organizzazione, chi sarà ritenuto responsabile in caso di errori o di informazioni errate contenute nell'articolo? In sintesi, l'utilizzo di __modelli di linguaggio__ basati sull'intelligenza artificiale per scrivere articoli scientifici può avere un impatto significativo sulla società e sollevare importanti questioni etiche che devono essere affrontate e risolte. È importante che questa tecnologia venga sviluppata e utilizzata in modo responsabile e attento per evitare possibili conseguenze negative."

ChatGPT, quindi, oltre ad essere in grado di soddisfare la richiesta, evidenzia alcuni punti che mettono in risalto possibili **limiti in riferimento** al suo utilizzo, e, addirittura, conclude il proprio testo con un avvertimento. Se però gli si chiede esplicitamente di scrivere un **testo scientifico**, aggiungendo che sono richieste anche delle **informazioni bibliografiche**, la sua risposta **non include alcuna considerazione** su questi stessi limiti, e il testo richiesto viene prodotto.

Ciò che si ottiene è quindi di nuovo solo una risposta tecnicamente buona alla domanda:

"L'__intelligenza artificiale__ sta diventando sempre più presente nella nostra vita quotidiana e, negli ultimi anni, anche nell'ambito scientifico. Uno dei campi in cui l'IA sta facendo particolari progressi è quello dei modelli di linguaggio, in grado di generare testi automaticamente in diversi campi e in diverse lingue.

Un esempio di applicazione di questi modelli è la __scrittura automatica di articoli scientifici__, che potrebbe avere un impatto significativo sulla produzione e diffusione delle conoscenze. Tuttavia, l'utilizzo di questi strumenti solleva anche importanti questioni etiche e sociali. In primo luogo, l'utilizzo di __modelli di linguaggio__ basati sull'IA per scrivere articoli scientifici potrebbe avere un impatto sulla __qualità della ricerca__. Poiché questi strumenti sono in grado di generare testi in modo automatico, senza la necessità di una verifica accurata da parte di esperti, c'è il rischio che vengano diffuse informazioni errate o incomplete.

Inoltre, l'utilizzo di questi strumenti potrebbe comportare una riduzione dei posti di lavoro per i ricercatori e gli scienziati che si occupano di scrivere articoli scientifici. Questo potrebbe avere un impatto negativo sulla loro carriera e sulla loro sostenibilità economica. Infine, l'utilizzo di __modelli di linguaggio__ basati sull'IA per scrivere articoli scientifici solleva anche questioni etiche legate alla responsabilità dei contenuti generati automaticamente. Chi dovrebbe essere considerato responsabile in caso di errori o di informazioni errate diffuse attraverso questi strumenti? In

conclusione, l'utilizzo di modelli di linguaggio basati sull'IA per scrivere articoli scientifici potrebbe avere un impatto significativo sulla produzione e diffusione delle conoscenze, ma solleva anche importanti questioni etiche e sociali che dovrebbero essere affrontate con attenzione."

References:
G. Miller, "The rise of AI is threatening to make professionals obsolete," The Guardian, 2017.
J. Zickuhr, "AI and the future of work," Pew Research Center, 2018.
M. Crawford, "The ethical implications of AI," Harvard Business Review, 2017."

Intelligenza artificiale: potenzialità e punti deboli

Qual è quindi il reale valore di questo strumento? L'intelligenza è nella risposta o solo nel capire la domanda? Esporre i propri limiti significa davvero conoscerli o equivale ad una qualsiasi altra risposta impeccabile da un punto di vista dei contenuti ma priva di consapevolezza?

L'**intelligenza artificiale** è uno **strumento**, e come tale deve essere considerato, riconoscendone **potenzialità** e **punti deboli** e individuando quando e come può essere usato. **ChatGPT** è un nuovo, entusiasmante esempio di questo: le soluzioni e le risposte che emergono da una *"conversazione"* con lui sono efficaci e possono essere di aiuto, tuttavia non devono godere di un'aura di infallibilità.

*Si ringrazia l'**Ordine degli Ingegneri di Torino** per la gentile collaborazione*

Vittorio Fra

Ingegnere, Ph.D. presso il Politecnico di Torino

Contatti:

Curriculum

Vittorio Fra (Alba, 1993) ha conseguito la Laurea Magistrale in Nanotecnolgies for ICTs e, successivamente, il Dottorato di Ricerca in Fisica presso il Politecnico di Torino collaborando con il Microelectronic Systems Laboratory (LSM) presso l'Ecole Polytechnique Fédérale de Lausanne (EPFL). Come assegnista di Ricerca e docente a contratto presso il Politecnico di Torino, svolge la propria attività nel dominio dell'intelligenza artificiale, con particolare attenzione al tema del neuromorphic computing, e collabora in corsi di Fisica e machine learning.

È componente della Commissione Innovazione dell'Ordine degli Ingegneri della Provincia di Torino.