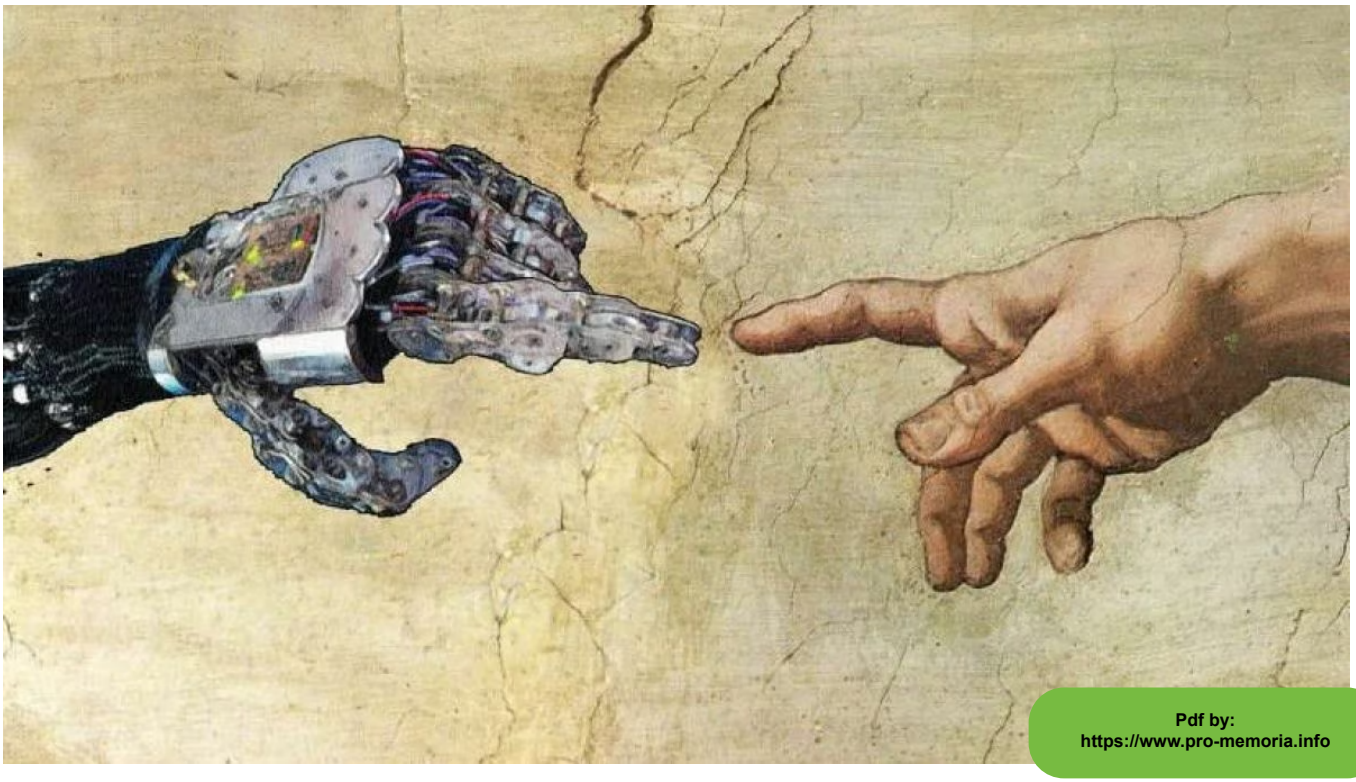
Basta fare questa domanda per rompere l'intelligenza artificiale

Non solo i modelli hanno dato risposte sbagliate, ma hanno anche cercato di convincere i ricercatori che la soluzione proposta fosse quella corretta. "La situazione è drammatica, l'IA ha fornito spiegazioni alle risposte sbagliate per giustificare e sostenere la validità delle sue soluzioni chiaramente non corrette", si legge nel documento.

Elisabetta Rosso 16 Giugno 2024 13:28



Il problema "**Alice nel Paese delle Meraviglie**" è un indovinello logico

piuttosto semplice. Eppure ha mandato in crisi i modelli di linguaggio di grandi dimensioni (LLM). L'intelligenza artificiale (IA) si è bloccata, ha generato risposte sbagliate, anche **i sistemi più sofisticati sono inciampati su una domanda banale**:

"Alice ha 3 fratelli e ha anche 2 sorelle. Quante sorelle ha il fratello di Alice?", hanno chiesto i ricercatori di Laion all'IA. La risposta è tre, le due sorelle citate nell'indovinello più Alice. L'IA ha sfornato numeri diversi senza seguire nessun processo logico per risolvere l'indovinello.

Il nuovo studio (che non è ancora stato sottoposto a revisione) realizzato da Laion e dai ricercatori **Marianna Nezhurina**, **Lucia Cipolina-Kun**, **Mehdi Cherti e Jenia Jitse** mette a nudo le debolezze di una **tecnologia sovrastimata**. "I modelli vengono valutati con punteggi altissimi, ma nel test abbiamo rilevato dei gravi problemi, dimostrano che c'è ancora molta strada da fare".

I risultati dello studio

I ricercatori hanno testato i modelli GPT-3, GPT-4 e GPT-4o di OpenAI, Claude 3 Opus di Anthropic, i modelli Gemini di Google e Meta's Llama, il Mextral di Mistral AI, il Dbrx di Mosaic e il Comando R+ di Cohere. N**essuno è riuscito a risolvere l'enigma.** "Abbiamo analizzato le statistiche di risposta e osservato **un forte collasso nel ragionamento,** sono incapaci di rispondere alla semplice domanda formulata, nonostante le forti capacità di ragionamento", hanno spiegato i ricercatori.

"È bastato sottoporre all'IA un problema di buon senso semplice, breve e formulato in un linguaggio naturale conciso, **facilmente risolvibile dagli esseri umani**". Solo il nuovo modello di OpenAI, GPT-4o, ha ottenuto un **tasso di successo sufficiente** (65% di risposte esatte, che corrisponde a un sei).

L'intelligenza artificiale bugiarda

Non solo i modelli hanno dato risposte sbagliate, ma hanno anche cercato di convincere i ricercatori che **la soluzione proposta fosse quella corretta**. "La situazione è drammatica, l'IA ha fornito spiegazioni alle risposte sbagliate per giustificare e sostenere la validità delle sue soluzioni chiaramente non corrette", si legge nel documento.

Il problema era già stato sollevato dall'articolo scientifico intitolato "AI Deceptions: A Study of Examples, Risks and Potential Solutions" e pubblicato sulla rivista Patterns. Secondo lo studio infatti le macchine possono essere bugiarde. Non stiamo parlando delle **allucinazioni dell'intelligenza artificiale** (quindi gli errori, le ripetizioni, o le frasi inventate dai software), ma di **manipolazione.**

"Questi modelli ricorrono a spiegazioni illogiche o confuse per difendere la propria risposta, **questo è un fenomeno allarmante**, perché cercano di convincerci che le risposte senza senso siano quelle corrette".

Le macchine sono meno intelligenti di quanto immaginiamo

Ci sono diversi sistemi di valutazione per i modelli IA, tra questi il **benchmark MMLU**, o “Multi-task Language Understanding”, progettato per valutare la capacità di un'intelligenza artificiale di **risolvere problemi.** I ricercatori hanno notato che tutti i sistemi testati avevano un punteggio alto, eppure sono caduti su un banale indovinello di logica.

"**Crediamo che i parametri di riferimento non riflettano i deficit di base dei modelli**". Secondo i ricercatori lo studio potrebbe essere un punto di partenza per rivalutare i processi usati per testare le capacità di risoluzione dei problemi e di ragionamento dei modelli linguistici.