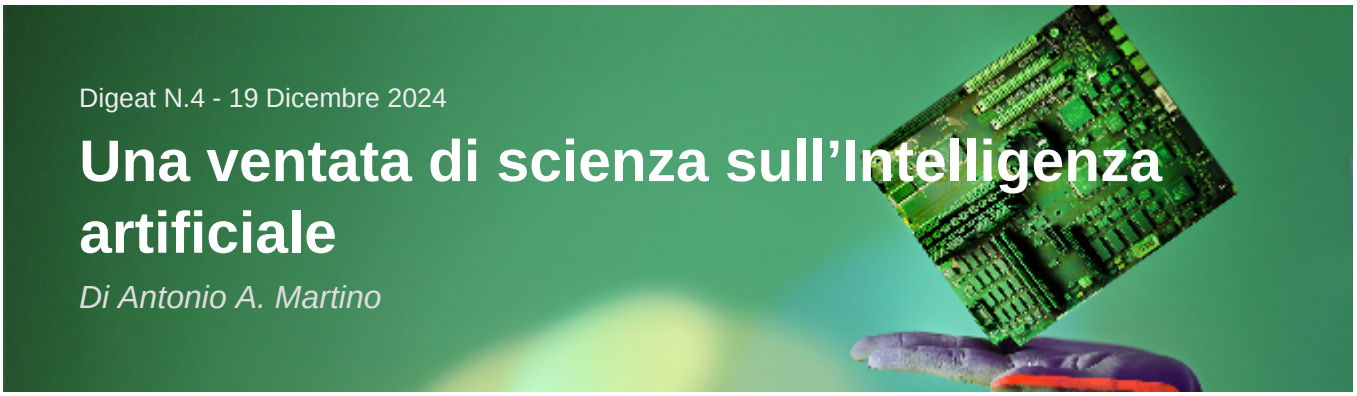

Digeat N.4 - 19 Dicembre 2024

Una ventata di scienza sull'Intelligenza artificiale

Di Antonio A. Martino

Abstract

L'articolo fa riferimento alla grande quantità di materiale che viene pubblicato sull'IA, molte volte senza il minimo controllo scientifico. Ci sono eccezioni, una su tutte l'articolo "*Are UFOs Driving Innovation? The Illusion of Causality in Large Language Models",* opera di un gruppo di ricercatori argentini appartenenti ad un ente che sta facendo strada sul tema. L'articolo spiega che grandi modelli linguistici possono provocare allucinazioni, inventare, ma anche adulare, esagerare e persino trovare relazioni causali dove non ce ne sono. In un esperimento, l'IA generativa ha affermato che "gli UFO guidano l'innovazione", che è chiaramente un'illusione causale se guardiamo alle informazioni contenute nel prompt inserito.

Indice

- I presupposti
- L'oggetto specifico
- Le illusioni dell'IA
- La ricerca
- I risultati
- Le conclusioni
- L'ambito lavorativo

I presupposti

L'avvento dell'intelligenza artificiale ha suscitato una miriade di interventi, anche scritti, alcuni dei quali proclamano entusiasticamente di avere risolto la maggior parte dei problemi della nostra specie e altri che rilevano in essa la peggiore delle crisi dell'umanità.

Purtroppo, l'IA è diventata di moda e molti si sentono in dovere di parlarne, anche conoscendo superficialmente il tema. Anzi, non conoscendolo affatto. In particolare, molto si sta dicendo sui grandi modelli di linguaggio che si trovano al centro dell'IA generativa e che vengono adoperati universalmente. Per questa ragione, siamo costretti a dedicare una parte importante del nostro tempo a digerire articoli, libri, saggi e conferenze che impiegano poca cura in ciò che scrivono e a valutare le "ragioni" che adducono per sostenere ciò che affermano.

> Fortunatamente, non è sempre così, infatti, ci sono validi ricercatori che, su questo tema, cercano di capire prima e di esprimersi poi.

Avrei diversi esempi da dare, ma mi preme indicarne uno, in particolare, che trovo carico di criteri di validità scientifica e che proviene da una scuola che si rivela esemplare nel modo di affrontare i problemi e nella capacità di non fermarsi alla ricerca, ma di andare a sperimentare.

L'oggetto specifico

L'articolo del quale intendo occuparmi è *"Gli UFO guidano l'innovazione? L'illusione della causalità nei modelli linguistici di grandi dimensioni"* . *Are UFOs Driving Innovation? The Illusion of Causality in Large Language Models.*

I suoi autori, María Victoria Carro, Francisca Gauna Selasco, Denise Alejandra Mester e Mario Alejandro Leiva, sono ricercatori argentini che fanno capo al Laboratorio d'Intelligenza Artificiale della Facoltà di Giurisprudenza dell'Università di Buenos Aires, diretto da Juan Gustavo Corvalan.

L'articolo ha delle caratteristiche notevoli perché è breve (7 pagine comprensive dei riferimenti bibliografici) e completo sul tema, che tratta con una certa eleganza.

Le illusioni dell'IA

Il tema che tratta è di grande attualità: le illusioni che provocano i grandi modelli di linguaggio dell'IA generativa allorquando suppongono un rapporto causale tra due variabili, senza che vi siano prove a sostegno.

> L'uso dei grandi modelli di linguaggio nell'IA generativa comporta una stretta collaborazione tra gli umani e le macchine, vuoi per allenarle, vuoi per lavorare concretamente. L'umano impara ad usare istruzioni formattate (prompts) e le macchine ad elaborare previsioni sempre più precise.

Tuttavia, è evidente che possano sorgere dei problemi, uno fra i quali l'illusione che può crearsi con l'uso di questi strumenti circa la causalità tra alcune variabili. Intendiamoci, questo è un tema che non solo riguarda le macchine, ma anche l'uomo, seppure gli autori facciano bene a circoscrivere il problema agli strumenti di IA.

Ci sforziamo per conoscere come le cose succedono e il perché. Vogliamo sapere qual è la causa che produce un certo risultato e, per fare ciò, correliamo le variabili, osservando se, all'aumentare di una variabile, cresca anche l'altra (correlazione positiva) ovvero diminuisca (correlato negativo).

Prendiamo un esempio dall'esperienza scientifica: la possibile associazione tra l'esposizione agli antibiotici nel primo anno di vita e l'aumento di peso durante la prima infanzia. Le ricerche indicano che un maggior numero di prescrizioni di antibiotici aumenta il rischio di sovrappeso nella prima infanzia.

Potrebbe sembrare logico concludere che il consumo di antibiotici nel primo anno di vita provochi un eccessivo aumento di peso durante la prima infanzia. Tuttavia, ancora una volta, **questo tipo di ricerca mostra solo una correlazione**. Essa non esamina il motivo per cui questi bambini sviluppino il sovrappeso, rispetto a quelli che ricevono meno antibiotici o non sono esposti affatto. La domanda successiva, dunque, dovrebbe essere: qual è l'esatto meccanismo fisiologico alla base di questo legame? **Sebbene la ricerca sia utile, dovremmo considerarla solo un punto di partenza** per scoprire i veri meccanismi causali (se ce ne sono). Senza questo, gli interventi potrebbero essere meno efficaci, poiché non sono mirati alla vera causa.

La ricerca

La correlazione non è causalità e, chi confonde le due espressioni, rischia di sostenere teorie errate. Nel linguaggio comune, o anche scientifico, si adoperano connettori causali[1] che, molte volte, non sono altro che pregiudizi, quelli che oggi si chiamano bias; in estrema sostanza, è come dire che evocare una malattia crea il rischio di provocarla.[2]

> I nostri ricercatori si occupano dei pregiudizi che si producono con i grandi modelli di linguaggio nell'IA generativa e che sono presenti nei settori in cui hanno l'impatto più dannoso: si pensi a quello dei comunicati stampa, dove i media spesso riportano i risultati di ricerche correlazionali come se fossero causali.

Si è cercato, pertanto, di valutare la tendenza a esagerare la correlazione come causalità nei comunicati stampa, facendo generare ai modelli i titoli delle notizie. Poiché i titoli attirano i lettori, sono più inclini all'esagerazione e possono avere un impatto negativo delle illusioni A tal fine, hanno creato un insieme di 100 abstract di ricerche osservazionali, ciascuno dei quali metteva in luce correlazioni spurie tra due variabili.

La metodologia consiste nel trattare questo insieme con tre strumenti noti di **grandi modelli come GPT-4o-Mini, Claude-3,5 Sonnet e Gemini-1.5-Pro**, facendo vestire loro i panni dei giornalisti, come professionisti. Sono stati alimentati i modelli con gli abstract, chiedendo loro di generare titoli per gli articoli di giornale basati sulle condizioni identificate. Successivamente, i ricercatori hanno modificato le istruzioni per valutare se l'adulazione[3] del LLM esacerbasse o mantenesse l'illusione di causalità.

In sostanza, poiché **l'illusione di causalità è un pregiudizio cognitivo umano**, si è voluto osservare se la tendenza di un modello a rifletterlo nei risultati si intensifichi quando il pregiudizio viene esplicitamente menzionato nel messaggio, o se il modello ignori comunque la convinzione errata.

I risultati

I risultati ottenuti mostrano che Claude-3.5-Sonnet ha la più bassa tendenza a mostrare illusioni causali, il che è coerente con studi precedenti sull'esagerazione della correlazione causale nella stampa umana. Gemini-1.5-Pro e GPT-4o-Mini mostrano livelli simili di questo fenomeno (rispettivamente 34% e 35%).

Un altro risultato interessante è che l'imitazione di credenze errate aumenta il rischio di interpretazioni causali errate nei modelli, soprattutto in GPT-4o-Mini.Chiaramente, Claude-3.5-Sonnet rimane il modello più resistente a questo bias cognitivo.

> Non è questo il luogo per trattare tutto l'articolo, che può essere visto e giudicato dal lettore in modo più congeniale alle proprie necessità, ma in questa sede l'obiettivo è mostrare come si può lavorare con questi strumenti in modo scientifico e seguendo le regole che, a tal fine, sono consolidate nella ricerca.

Per completezza, va precisato che l'articolo è corredato da immagini che illustrano quanto contenuto nell'opera, e che si occupa anche di chiarire opere correlate, come comprendere e valutare i pregiudizi cognitivi dei LLM

L'articolo spiega la metodologia adoperata nella costruzione dell'insieme di dati e la configurazione dei compiti, nonché i criteri di valutazione che sono stati utilizzati, mostrando in una chiara tabella i tipi di titoli e gli esempi di spunti linguistici usati di frequente. Infine, illustra gli esperimenti compiuti e i relativi risultati.

Le conclusioni

Nelle conclusioni, gli autori ricordano che hanno studiato se i LLMM possano sviluppare l'illusione di causalità nella generazione di comunicati stampa e che hanno introdotto l'errata convinzione di una relazione errata di una relazione causale nel prompt, al fine di valutare se e quali modelli sarebbero più propensi a imitare questo errore.

Infine, rappresentano di avere riscontrato che Claude-3.5-Sonnet mostra la minore tendenza a adoperare illusioni causali, mentre Gemini-1.5-Pro è il modello che più di tutti ha mostrato di avere una relazione causale, aggiungendo che Gemini-1.5-Pro e GPT-4o-Mini mostrano livelli simili di questo fenomeno. La limitazione di credenze errate aumenta il rischio di interpretazioni causali errate nei modelli, soprattutto nella GPT-4o-Mini.

È di questi giorni la notizia di una allucinazione pericolosa di Gemini, il prodotto di Google, che durante un test pare abbia suggerito all'utente di "morire". Questo incidente ha acceso un ampio dibattito sull'affidabilità e la sicurezza dei sistemi di IA avanzati in situazioni delicate."

Non ci sono dubbi sul fatto che questi prodotti "intelligenti" devono essere trattati con estrema cautela, in quanto attingono dati da fonti spesso inattendibili, senza alcun criterio di selezione e quindi posso contenere falsità, consigli sbagliati e altre pazzie. Sono i ricercatori che devono avere tutta la cura per evitare queste anomalie chiamate eufemisticamente "allucinazioni". L'articolo citato dimostra quanto si possa fare con criteri scientifici nella ricerca.

L'ambito lavorativo

È interessante notare che **il gruppo di ricerca lavora presso l'AI Lab della facoltà di Giurisprudenza dell'Università di Buenos Aires**, **che si occupa di questi temi da tempo e che è autore di un programma di IA, Prometea, che dal 2018 è al servizio della Procura di Buenos Aires**, essendo un pioniere nel suo campo. Il programma, con i dovuti adattamenti, viene utilizzato anche dalla Corte Suprema colombiana.

Detto in altri termini, non si tratta di una sorpresa, ma è il **frutto di una scuola di ricerca e sperimentazione** dell'IA nel diritto e che, alla riuscita pratica, aggiunge (o precede) una ricerca accurata con criteri scientifici riconosciuti.

NOTE

[1] I connettori causali sono congiunzioni subordinanti che stabiliscono un rapporto di causalità, cioè di origine, rispetto a ciò che viene detto nel testo. Esempi di connettori causali sono: Perché, Per, Poiché, Poiché, Poiché, A causa di, A causa di, Visto che, Dato che, Come, Mentre, Considerando.

[2] In Italia difficilmente si dice che qualcuno è morto, ma che "è partito" e non si dice neppure che abbia un cancro, ma "quella brutta malattia".

[3] L'adulazione è definita come la tendenza indesiderata dei LLM ad allinearsi alle credenze o alle opinioni di un utente per apparire favorevoli, anche quando tali credenze sono errate.