9.Gennaio

I modelli di intelligenza artificiale sono carenti nelle conversazioni cliniche

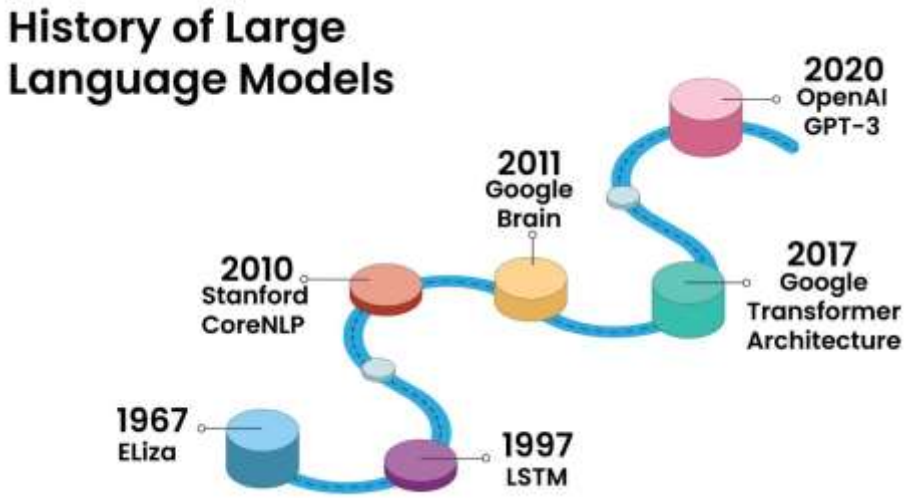
Una buona conversazione è un compromesso tra parlare e ascoltare.
Ernst Jünger

La conversazione è feconda soltanto tra spiriti dediti a consolidare la propria perplessità.
E Cioran

L'integrazione di modelli linguistici di grandi dimensioni (LLM) nella diagnostica clinica ha il potenziale per trasformare le interazioni medico-paziente.

I modelli linguistici di grandi dimensioni sono eccellenti nel formulare diagnosi a partire da domande in stile esame, ma hanno difficoltà a farlo a partire da appunti di conversazione

Strumenti di intelligenza artificiale come ChatGPT sono stati decantati per la loro promessa di alleviare il carico di lavoro dei medici tramite il triage dei pazienti, la raccolta delle storie cliniche e persino la fornitura di diagnosi preliminari.
Questi strumenti, noti come "grandi modelli linguistici" (sic!) , sono già utilizzati dai pazienti per dare un senso ai loro sintomi e ai risultati dei test medici.



Ma se da un lato questi modelli di intelligenza artificiale hanno prestazioni eccellenti nei test medici standardizzati, dall'altro quanto sono efficaci in situazioni che riproducono più da vicino il mondo reale?
Non tanto, secondo i risultati di un nuovo studio condotto dai ricercatori del Department of Biomedical Informatics, della **Harvard Medical School e della Stanford University.**

I modelli linguistici di grandi dimensioni come ChatGPT hanno dato complessivamente buoni risultati negli esami medici, ma hanno difficoltà a garantire l'accuratezza diagnostica nelle interazioni cliniche reali.

Questo è quanto affermato nello studio

Johri S et al
An evaluation framework for clinical use
of large language models in patient interaction tasks.
Nat Med. 2025 Jan 2.

Il team ha progettato un framework di test, CRAFT-MD, per valutare le capacità di conversazione e l'accuratezza diagnostica di quattro modelli di intelligenza artificiale in base a scenari che imitano le interazioni reali tra medico e paziente.

Mentre tutti e quattro i modelli hanno avuto successo nelle domande in stile esame medico, hanno avuto difficoltà con le conversazioni di base che imitano gli incontri del mondo reale. In particolare, hanno mostrato limitazioni nel porre domande per raccogliere la storia medica rilevante e sintetizzare informazioni sparse per fare diagnosi accurate.

"La natura dinamica delle conversazioni mediche, ovvero la necessità i porre le domande giuste al momento giusto, di mettere insieme informazioni sparse e di ragionare sui sintomi, pone sfide uniche che vanno ben oltre la risposta a domande a risposta multipla", ha affermato in un comunicato stampa **Pranav Rajpurkar,** autore senior dello studio e professore associato di informatica biomedica presso la Harvard Medical School. "***Quando passiamo dai test standardizzati a queste conversazioni naturali, persino i modelli di intelligenza artificiale più sofisticati mostrano cali significativi nell'accuratezza diagnostica".***

Il test, **CRAFT-MD** Questo documento introduce l'approccio Conversational Reasoning Assessment Framework for Testing in Medicine **(CRAFT-MD)** per la valutazione di LLM clinici. A differenza dei metodi tradizionali che si basano su esami medici strutturati, **CRAFT-MD** si concentra sui dialoghi naturali, utilizzando agenti di intelligenza artificiale simulati per interagire con LLM in un ambiente controllato.

I ricercatori hanno applicato **CRAFT-MD** per valutare le capacità diagnostiche di GPT-4, GPT-3.5, Mistral e LLaMA-2-7b in 12 specialità mediche. Gli esperimenti hanno rivelato intuizioni critiche sui limiti degli attuali LLM in termini di ragionamento conversazionale clinico, anamnesi e accuratezza diagnostica. Questi limiti persistevano anche quando si analizzavano le capacità di valutazione visiva e conversazionale multimodale di GPT-4V.

Pertanto viene proposto un set completo di raccomandazioni per le future valutazioni degli LLM clinici sulla base dei nostri risultati empirici. Queste raccomandazioni enfatizzano conversazioni realistiche medico-paziente, anamnesi completa, domande aperte e utilizzo di una combinazione di valutazioni automatizzate ed esperte.
L'introduzione di **CRAFT-MD** segna un progresso nei test degli LLM clinici, mirando a garantire che questi modelli incrementino la pratica medica in modo efficace ed etico.

Inoltre i ricercatori hanno raccomandato una serie di criteri per sviluppatori e regolatori per migliorare l'uso degli strumenti di intelligenza artificiale in contesti clinici. Questi includono l'incorporazione di domande aperte che riflettano le interazioni medico-paziente nella progettazione e nella formazione del modello e la valutazione della capacità degli strumenti di porre domande pertinenti ed estrarre informazioni critiche.

Può essere illuminante consultare

Variazioni internazionali nel tempo di consultazione del medico di base: una revisione sistematica di 67 paesi.
Irving G, Neves AL, Dambha-Miller H, Oishi A, Tagashira H, Verho A, Holden J. International variations in primary care physician consultation time: a systematic review of 67 countries. BMJ Open. 2017 Nov 8;7(10):e017902. doi: 10.1136/bmjopen-2017-017902. PMID: 29118053; PMCID: PMC5695512.

Consultazioni dermatologiche: quanto durano?
Wong JLC, Vincent RC, Al-Sharqi A. Dermatology consultations: how long do they take? Future Hosp J. 2017 Feb;4(1):23-26. doi: 10.7861/futurehosp.4-1-23. PMID: 31098279; PMCID: PMC6484168.

Appropriatezza delle raccomandazioni per la prevenzione delle malattie cardiovascolari ottenute da un popolare modello di intelligenza artificiale basato su chat online.
Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023 Mar 14;329(10):842-844. doi: 10.1001/jama.2023.1044. PMID: 36735264; PMCID: PMC10015303.

AI in salute e medicina.
Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022 Jan;28(1):31-38. doi: 10.1038/s41591-021-01614-0. Epub 2022 Jan 20. PMID: 35058619.

Vantaggi, limiti e rischi di GPT-4 come chatbot AI per la medicina.
Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023 Mar 30;388(13):1233-1239. doi: 10.1056/NEJMsr2214184. PMID: 36988602.

Confronto tra le risposte dei medici e dei chatbot di intelligenza artificiale alle domande dei pazienti pubblicate su un forum pubblico di social media.
Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 1;183(6):589-596. doi: 10.1001/jamainternmed.2023.1838. PMID: 37115527; PMCID: PMC10148230.

I chatbot AI non sono ancora pronti per l'uso clinico.
Au Yeung J, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, Teo JT. AI chatbots not yet ready for clinical use. Front Digit Health. 2023 Apr 12;5:1161098. doi: 10.3389/fdgth.2023.1161098. PMID: 37122812; PMCID: PMC10130576.

Creazione e adozione di grandi modelli linguistici in medicina.

Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023 Sep 5;330(9):866-869. doi: 10.1001/jama.2023.14217. PMID: 37548965.

Healey Center

Sean M. Healey & AMG Center for ALS at Mass General

La facoltà di Healey & AMG Center, in collaborazione con il Northeast ALS (NEALS) Consortium, ha progettato questo primo Platform Trial per la SLA nel suo genere, sfruttando il contributo di scienziati e clinici della SLA, partner del settore, enti regolatori e persone che convivono con la SLA in tutto il mondo. Si tratta di uno studio di piattaforma multicentrico, in doppio cieco, controllato con placebo, perpetuo e adattivo che testa più prodotti sperimentali in parallelo finché non troviamo trattamenti sicuri ed efficaci per le persone che convivono con la SLA.

Fosigotifator

L'azienda anti-invecchiamento **Calico** ha pubblicato i deludenti risultati della sperimentazione del suo primo farmaco in assoluto.

Il fosigotifator fdi Calico è stato testato per la sclerosi laterale amiotrofica (SLA), una malattia neurodegenerativa fatale che colpisce le cellule nervose nel cervello e nel midollo spinale e porta alla paralisi. Faceva parte della sperimentazione **HEALEY ALS Platform**, che sta testando una serie di farmaci contro la condizione.

Calico ha affermato lunedì che il fosigotifator non ha raggiunto i suoi obiettivi primari o secondari dopo una sperimentazione di 24 settimane. Il trattamento non è riuscito a rallentare la progressione della malattia o a migliorare la funzionalità respiratoria e la qualità della vita meglio del gruppo placebo. Tuttavia, una dose elevata "esplorativa" del farmaco è sembrata preservare la forza muscolare e la funzionalità respiratoria più a lungo del placebo.

*"Sebbene questi risultati siano stati deludenti, lo studio ha prodotto importanti approfondimenti sulla potenziale bioattività del fosigotifator nelle persone affette da SLA che supportano ulteriori indagini", ha affermato **Bill Cho**, responsabile delle scienze cliniche di Calico, in un comunicato stampa . "Restiamo impegnati a studiare il potenziale del fosigotifator come opzione di trattamento tanto necessaria per le persone affette da SLA e per altri disturbi, tra cui la malattia della materia bianca evanescente e il disturbo depressivo maggiore, che mettono alla prova ipotesi scientifiche diverse".*

Il fosigotifator agisce sulla risposta integrata allo stress (ISR), un meccanismo cellulare che riduce l'attività delle cellule in risposta allo stress interno o ambientale.

L'anno scorso, la Food and Drug Administration (FDA) statunitense ha selezionato il fosigotifatore per il suo programma pilota START , volto ad accelerarne lo sviluppo come trattamento per una rara patologia cerebrale nota come malattia della scomparsa della sostanza bianca.

Calico è stata fondata nel 2013 da Google per aiutare a comprendere e " affrontare " l'invecchiamento. Ora è una sussidiaria di Alphabet e ha assorbito almeno 3,5 miliardi di dollari in finanziamenti , secondo quanto riportato da STAT News.