Recenti Prog Med 2025; 116: 137-138

I modelli linguistici di grandi dimensioni possono rispondere a quesiti clinici?

EUGENIO SANTORO[1]

[1]Unità di Ricerca in sanità digitale e terapie digitali, Laboratorio di Metodologia della ricerca clinica, Dipartimento di Oncologia clinica, Istituto di Ricerche farmacologiche Mario Negri Irccs, Milano.

Pervenuto su invito il 28 febbraio 2025. Non sottoposto e revisione critica esterna alla redazione della rivista.

Riassunto. Con l'avanzamento dei modelli linguistici di grandi dimensioni (Llm) come ChatGPT, la loro applicazione in medicina sta crescendo, ma è cruciale che le risposte siano allineate alle linee guida internazionali. Recenti studi hanno dimostrato che gli Llm possono essere utili in ambito medico, rispondendo correttamente a quesiti sulla gestione e il trattamento di specifiche patologie. Tuttavia, l'accuratezza di questi modelli deve riguardare anche la completezza, la chiarezza e la contestualizzazione delle risposte. Accanto a queste caratteristiche devono essere garantite pertinenza, rilevanza e aggiornamento delle fonti usate dai Llm per rispondere alle domande. Sono inoltre necessari studi che indaghino sulla omogeneità delle risposte tra i vari Llm e tra le diverse lingue impiegate dagli stessi, e processi di addestramento che garantiscano una maggiore solidità soprattutto quando si trattano patologie rare o complesse. Sebbene i Llm possano supportare la formazione medica e il processo decisionale, la loro integrazione nella pratica clinica richiede ulteriori validazioni e confronti con le linee guida internazionali.

Parole chiave. ChatGPT, formazione medica, intelligenza artificiale, modelli linguistici di grandi dimensioni, processo decisionale.

Can large language models answer clinical questions?

Summary. With the advancement of large language models (LLMs) such as ChatGPT, their application in medicine is growing, but it is crucial that the responses are aligned with international guidelines. Recent studies have shown that LLMs can be useful in the medical field, providing correct answers to questions about the management and treatment of specific diseases. However, the accuracy of these models must also include readability and thoroughness of the answers and consistency with guidelines. In addition to these characteristics, relevance, pertinence, and up-to-date nature of the sources used by the LLMs to answer questions must be ensured. Furthermore, studies are needed to investigate the consistency of responses across different LLMs and languages used by them, as well as training processes that ensure greater reliability, especially when dealing with rare or complex diseases. Although LLMs can support medical education and decision-making, their integration into clinical practice requires further validation and comparison with international guidelines.

Key words. Artificial intelligence, ChatGPT, decision-making, large language models, medical education.

Con l'avanzamento delle tecnologie di intelligenza artificiale (IA), i modelli linguistici di grandi dimensioni (Llm) come ChatGPT, Gemini, Copilot, Perplexity e Claude sono diventati strumenti sempre più utilizzati nel campo della medicina. Questi modelli sono addestrati su enormi quantità di dati, tra cui articoli scientifici, libri, documenti medici e discussioni online, e sono in grado di rispondere a una vasta gamma di domande su temi complessi, inclusi quelli legati alla salute[1].

Tuttavia, quando si parla di applicazioni in ambito sanitario, è fondamentale che le risposte fornite siano non solo coerenti e chiare, ma anche allineate con le linee guida internazionali, che stabiliscono gli standard per la pratica clinica. La valutazione dell'accuratezza di questi modelli rispetto a tali linee guida è quindi cruciale per garantire che possano essere utilizzati in modo sicuro e responsabile da medici, studenti di medici, cittadini/pazienti in contesti clinici, formativi e informativi.

Studi recenti hanno mostrato che i Llm possono essere estremamente promettenti e perfettamente aderenti alle linee guida internazionali nel rispondere a quesiti medici come la profilassi dell'endocardite infettiva nelle procedure odontoiatriche[2], la sindrome dell'ovaio policistico[3], l'iperprolattinemia[4].

L'accuratezza di questi modelli non dovrebbe riferirsi solo alla correttezza delle risposte, ma anche alla loro completezza, chiarezza e capacità di contestualizzare le informazioni. Inoltre, poiché le linee guida possono cambiare frequentemente in risposta a nuove scoperte scientifiche o a modifiche delle pratiche cliniche, è importante che i Llm vengano costantemente aggiornati per riflettere tali cambiamenti e per garantire l'accesso alle fonti più recenti.

Uno studio del 2024 ha affrontato proprio questi aspetti[5]. I ricercatori hanno sottoposto alla versione 4 di ChatGPT un questionario composto di 23 domande basate sulle raccomandazioni delle linee guida del National institute for health and care excellence (Nice) sulla gestione dell'acne. Le risposte sono state valutate da 5 dermatologi e si sono dimostrate sufficientemente affidabili in termini di qualità, leggibili-

tà, completezza e coerenza con le linee guida. Tuttavia, è emerso che le fonti utilizzate da ChatGPT non erano completamente aggiornate, anche se erano pertinenti e rilevanti.

Più interessante è il confronto in questi contesti delle prestazioni degli Llm più noti. Per esempio in un recente studio nel quale i ricercatori hanno sottoposto 100 domande riguardanti screening, diagnosi e trattamento del tumore della cervice (basate sulle principali linee guida internazionali) a 9 differenti Llm, le risposte fornite da 7 di questi si sono dimostrate sufficientemente accurate, con un grado di affidabilità più elevato per ChatGPT-4.0 e Claude 2.0[6].

Su questi temi si sono concentrati anche gli autori di uno degli articoli che appaiono su questo numero della rivista, interamente dedicato all'intelligenza artificiale, in cui hanno testato 3 Llm (ChatGPT 4.0, Gemini e Copilot) su 9 domande riguardanti lo screening del tumore al seno, basate sulle linee guida internazionali più recenti[7]. Le risposte, sia in italiano sia in inglese, hanno avuto punteggi simili, ma con alcune differenze tra i modelli. Le domande generali sull'imaging mammografico hanno ottenuto risposte accurate, mentre quelle più specifiche hanno mostrato qualche imprecisione. Un aspetto interessante emerso è che, soprattutto in italiano, le fonti utilizzate non erano sempre di enti scientifici accreditati, evidenziando un limite dei Llm nel fornire risposte mediche approfondite e aggiornate.

D'altra parte è noto in letteratura che questi modelli non sono esenti da errori, soprattutto quando trattano domande complesse o situazioni che richiedono una comprensione approfondita delle sfumature cliniche oppure quando riguardano patologie particolarmente rare, come dimostra un recente studio condotto sulla gestione e sul trattamento dei sarcomi[8].

C'è inoltre chi sostiene che particolare attenzione dovrebbe essere posta nella costruzione di prompt appropriati che possono migliorare l'accuratezza delle risposte alle domande mediche professionali. Un recente studio ha infatti confrontato diversi stili di prompt per valutare la coerenza delle risposte dei Llm riguardo alle linee guida per l'osteoartrite dell'American Academy of Orthopedic Surgeons, rilevando che prompt più elaborati e precisi offrono risposte più coerenti e di qualità superiore[9].

Esistono quindi sufficienti evidenze per considerare i Llm come supporto alla pratica clinica? Siamo ancora lontani, ma meno lontani rispetto a pochi anni/mesi fa. La tecnologia avanza, così come il modo di "pensare" degli Llm. A oggi possiamo concludere sostenendo che i Llm potrebbero essere utili nella formazione medica e nel processo decisionale, ma la loro integrazione nella pratica clinica richiede ulteriori validazioni e un continuo confronto con le linee guida internazionali. Tutto ciò in attesa che siano sviluppati Llm addestrati con articoli, linee guida, studi clinici rispetto a una singola area medica o a una specifica patologia e che siano adeguatamente studiati dal punto di vista della formazione medica, della diagnostica e della evidence-based medicine[10].

Conflitto di interessi: l'autore dichiara l'assenza di conflitto di interessi.

Bibliografia

1. Bettoli V, Naldi L, Santoro E, et al. ChatGPT and acne: accuracy and reliability of the information provided – The AI-check study. J Eur Acad Dermatol Venereol 2024; doi: 10.1111/jdv.20324.
2. Rewthamrongsris P, Burapacheep J, Trachoo V, Porntaveetus T. Accuracy of Large Language Models for infective endocarditis prophylaxis in dental procedures. Int Dent J 2025; 75: 206-12.
3. Gunesli I, Aksun S, Fathelbab J, Yildiz BO. Comparative evaluation of ChatGPT-4, ChatGPT-3.5 and Google Gemini on PCOS assessment and management based on recommendations from the 2023 guideline. Endocrine 2024; doi: 10.1007/s12020-024-04121-7.
4. Şenoymak MC, Erbatur NH, Şenoymak İ, Fırat SN. The role of artificial intelligence in endocrine management: assessing ChatGPT's responses to prolactinoma queries. J Pers Med 2024; 14: 330.
5. Naldi L, Bettoli V, Santoro E, et al. Application of ChatGPT as a content generation tool in continuing medical education: acne as a test topic. Dermatol Reports 2024; doi: 10.4081/dr.2024.10138.
6. Kuerbanjiang W, Peng S, Jiamaliding Y, Yi Y. Performance evaluation of Large Language Models in cervical cancer management based on a standardized questionnaire: comparative study. J Med Internet Res 2025; 27: e63626.
7. Signorini M, Fontani S, Minichetti P, Teggi S, Barusco A, Favat M. Valutazione dell'accuratezza di modelli linguistici di grandi dimensioni nel rispondere a domande sullo screening mammografico in italiano e inglese: uno studio basato sulle linee guida Eusobi. Recenti Prog Med 2025; 116: 162-7.
8. Valentini M, Szkandera J, Smolle MA, Scheipl S, Leithner A, Andreou D. Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients? Front Public Health 2024; 12: 1303319.
9. Wang L, Chen X, Deng X, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024; 7: 41.
10. Zheng C, Ye H, Guo J, et al. Development and evaluation of a large language model of ophthalmology in Chinese. Br J Ophthalmol 2024; 108: 1390-7.

Indirizzo per la corrispondenza:
Eugenio Santoro
E-mail: eugenio.santoro@marionegri.it