Salute, se la diagnosi dell’AI è sbagliata



*Rubrica settimanale **SosTech**, frutto della collaborazione tra **Key4biz** e **SosTariffe**. Per consultare gli articoli precedenti, **clicca qui**..*

Si è capito ormai da tempo che uno dei campi dove la guerra tra i giganti dell’AI si farà agguerrita è quello della **salute**, e gli investimenti di un po’ tutti i protagonisti nel settore la dicono lunga.

D’altronde, la sanità è per molti Paesi un tasto dolente: non c’è solo la nostra esperienza – code interminabili per un esame o una prestazione del SSN, che spesso costringono a ricorrere alla sanità privata quando bisogna intervenire in maniera tempestiva – ma soprattutto quella di chi paga ogni consulto cifre astronomiche, a meno di non avere costosissime assicurazioni (come accade negli Stati Uniti).

In questa nicchia acuita dalla perdita di potere d'acquisto dei cittadini si infila l'intelligenza artificiale, promettendo **un servizio gratuito e immediato**, accessibile ovunque ci si trovi, semplicemente digitando la richiesta con il proprio PC o cellulare (a proposito: su SOSTariffe.it si possono trovare sempre le **tariffe mobili** più convenienti). Già, peccato che rischi non poco di essere **inaffidabile**.

Salute: dai benchmark alla vita reale

Un ampio studio pubblicato su *Nature Medicine* riporta infatti risultati che invitano alla cautela sull'uso dei chatbot basati su *large language models* per ottenere consigli medici. La ricerca, coordinata dall'**Oxford Internet Institute** insieme al **Nuffield Department of Primary Care Health Sciences**, ha messo alla prova questi sistemi in una situazione che si avvicina all'esperienza quotidiana di chi cerca informazioni sulla propria salute.

Negli ultimi anni i modelli linguistici hanno dimostrato di saper offrire prestazioni elevate nei test standardizzati di conoscenza clinica, arrivando a superare anche gli esami professionali e a risolvere diversi casi diagnostici complessi presentati in forma strutturata.

Lo studio inglese analizza però qualcosa di diverso, e cioè l'interazione concreta tra persone comuni e sistemi di intelligenza artificiale quando si tratta di sintomi e decisioni da prendere.

L'ipotesi dei ricercatori era che le buone performance sui benchmark **potessero non tradursi automaticamente in indicazioni affidabili per il pubblico**, e i risultati confermano questa preoccupazione. In condizioni che simulano l'uso domestico, i chatbot non hanno migliorato in modo evidente la capacità degli utenti di individuare la possibile

causa dei sintomi o di scegliere il passo successivo più appropriato, tra cui rivolgersi al medico di base o recarsi al pronto soccorso.

Salute e AI: le difficoltà quando c’è incertezza

Lo studio si è basato su un *trial* randomizzato che ha coinvolto circa **1.300 partecipanti** nel Regno Unito, in larga parte privi di formazione medica. A ciascuno è stato assegnato uno scenario clinico elaborato dai medici, con descrizioni dettagliate di sintomi, anamnesi personale, abitudini e precedenti sanitari. I casi riguardavano situazioni molto diverse tra loro, da un forte mal di testa comparso dopo una serata con amici a una giovane madre stanca e con un affanno persistente.

Ai partecipanti veniva chiesto di **identificare la possibile condizione** e di decidere quale fosse l’azione più appropriata. Un gruppo ha potuto utilizzare un chatbot basato su LLM per orientarsi nella decisione; il gruppo di controllo ha fatto ricorso ai mezzi abituali, soprattutto **ricerche online e valutazioni personali**. Le risposte sono state confrontate con quelle definite corrette da un panel di clinici. I risultati mostrano che **gli utenti assistiti dai chatbot hanno scelto il percorso ritenuto appropriato meno della metà delle volte**. Anche l’identificazione della condizione sottostante si è rivelata problematica, con percentuali di correttezza attorno a **un terzo dei casi**. E, soprattutto, le prestazioni del gruppo che ha utilizzato l’intelligenza artificiale non si sono praticamente discostate da quelle del gruppo di controllo.

I ricercatori hanno poi inserito nei modelli l’intero scenario clinico, senza omissioni e in forma strutturata: in questo caso la capacità diagnostica è migliorata, raggiungendo livelli molto più elevati. Quando ci sono incertezza e una comunicazione imperfetta, in sostanza – situazioni tutt’altro

che rare nel quotidiano – i modelli **mostrano tutti i loro rischi**.

I tre problemi evidenziati

L’analisi delle conversazioni ha messo in luce tre problemi ricorrenti: il primo riguarda le **informazioni** fornite dagli utenti. I molti casi i partecipanti non indicavano durata, intensità o caratteristiche precise dei sintomi, e con un quadro incompleto i modelli formulavano ipotesi poco accurate. Quando i ricercatori inserivano direttamente lo scenario clinico completo, la percentuale di diagnosi corrette saliva fino al **94 per cento**, mostrando come sia netta la differenza tra un’interazione libera e degli input strutturati.

Il secondo problema è la **variabilità delle risposte**: modifiche minime nella formulazione erano la causa di indicazioni diverse sul livello di urgenza. In uno scenario che prevedeva mal di testa, rigidità del collo e sensibilità alla luce, una descrizione generica suggeriva di riposare e prendere analgesici, ma bastava sottolineare l’esordio improvviso e l’intensità estrema del dolore per generare un responso ben più drastico, con l’invito di recarsi al pronto soccorso. Se il contenuto clinico di base era simile, l’esito pratico cambiava.

Il terzo elemento riguarda la **qualità interna** delle risposte. I chatbot hanno mischiato consigli giudicati appropriati dagli esperti con raccomandazioni assai discutibili, senza una chiara graduazione del rischio; per utenti senza formazione sanitaria, è chiaro che diventa complesso distinguere quali indicazioni seguire. E come se non bastasse in alcuni casi sono comparse informazioni errate, come numeri telefonici di emergenza non validi.

Il problema è nei benchmark

Le implicazioni del test riguardano sia la valutazione dei sistemi sia il loro impiego pubblico. Gli autori dello studio sostengono che i **benchmark** attuali, basati su quesiti molto strutturati e risposte a scelta multipla o su casi clinici completi, **non misurano in modo adeguato ciò che accade nelle interazioni reali**. Nel trial, anche i modelli che potevano vantare punteggi elevati nei test tradizionali hanno fatto vedere tutte le loro difficoltà quando dovevano dialogare con utenti che fornivano informazioni parziali o formulate in modo impreciso (come, appunto, succede nella realtà).

Per questo i ricercatori propongono di adottare protocolli di valutazione che includano utenti eterogenei e scenari d'uso molto più realistici di condizioni "da laboratorio", con criteri centrati sulla qualità delle decisioni prese.

Il confronto con la **sperimentazione farmacologica** è stato richiamato per analogia, visto che in entrambi i casi si parla di strumenti destinati a incidere su scelte sanitarie e che quindi richiedono verifiche in condizioni d'uso effettive, non sottovuoto.

La diffusione dei chatbot per domande di salute è già ampia e coinvolge anche prodotti commerciali di grandi aziende tecnologiche, come si è visto; alcune società hanno dichiarato che le versioni più recenti dei modelli hanno ridotto gli errori più frequenti e aumentato la propensione a porre domande di chiarimento, ma con un campione ampio lo studio ha dimostrato quanta strada ci sia ancora da fare.

Leggi le altre notizie sull'home page di Key4biz

https://www.key4biz.it/salute-se-la-diagnosi-dellai-e-sbagliata/565848/