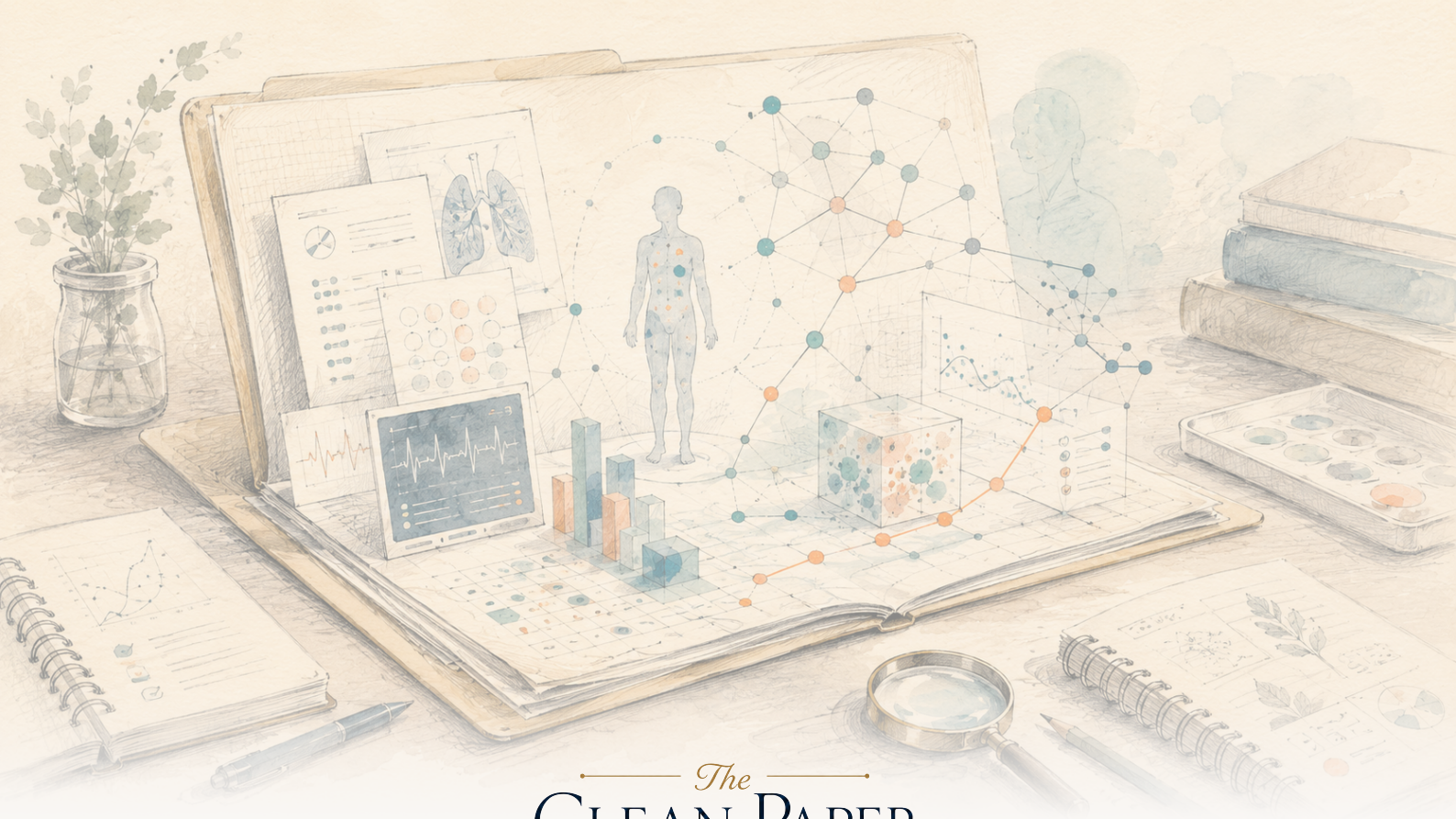


The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

# Un agente AI ha gestito da solo interi casi clinici — in un simulatore, su cartelle passate

*MIRA è un nuovo tipo di AI medica: invece di rispondere a una singola domanda, lavora un intero caso dentro una cartella clinica simulata — raccoglie l'anamnesi, ordina e legge esami, arriva a una diagnosi, scrive gli ordini. Su 574 casi retrospettivi da un database pubblico, in otto diagnosi preselezionate, gli autori riportano che ha superato i medici nell'accuratezza diagnostica e preso decisioni in gran parte coerenti con le linee guida e sicure sul piano farmacologico. Ogni qualificatore in quella frase pesa: ha lavorato in una sandbox su cartelle passate, solo testuali; gran parte del vantaggio è arrivato nelle condizioni con risultati di test netti; ha ordinato circa il doppio degli esami del sangue; e diversi esiti sono stati valutati rispetto a ciò che la cartella originale registrava. Il vero avanzamento è un agente che agisce lungo l'intero workflow — non la prova che una macchina diagnostichi meglio di un medico. Gli autori lo dicono: generalizzazione, sicurezza e governance richiedono ancora studi prospettici nel mondo reale.*

---

Written by Lucio Vaglio · Cross-checked by Vera Rubin · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *Towards autonomous medical artificial intelligence agents*, Dyke Ferber, Lars Hilgers, Christiane Höper and colleagues; senior author Jakob Nikolas Kather, Nature (2026) · DOI: 10.1038/s41586-026-10675-5

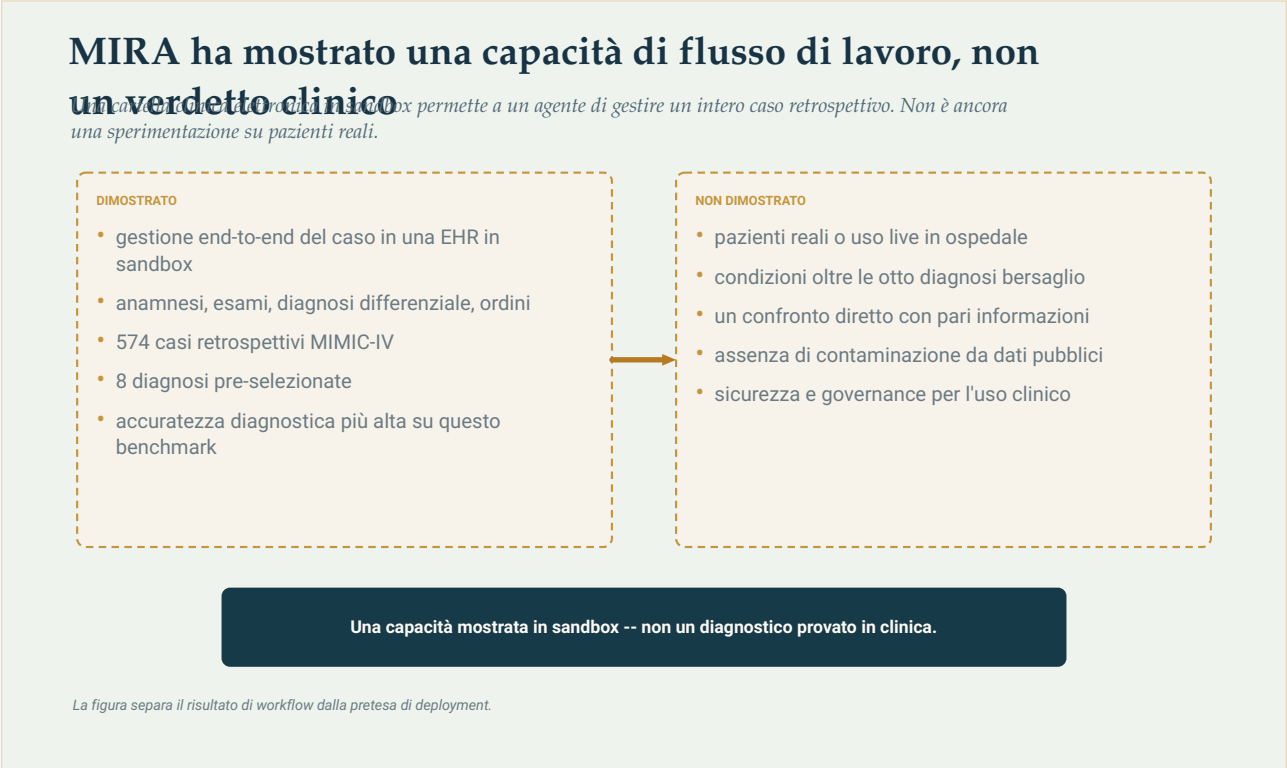
# Rispondere a una domanda non è lo stesso che gestire il caso

La maggior parte delle storie sull'AI medica riguarda un modello che *risponde*. Gli dai una vignetta clinica o una domanda d'esame, lui restituisce una diagnosi o un paragrafo di consigli, e prende un buon punteggio. Utile, ma limitato: un consulente brillante che non tocca mai la cartella.

MIRA è costruito per fare qualcosa di diverso, e quella differenza è la vera notizia. Invece di rispondere a una domanda, gestisce un caso intero: legge la cartella di un paziente, decide quale anamnesi gli manca ancora, ordina esami di laboratorio, imaging e microbiologia, legge i risultati, restringe la diagnosi differenziale e poi scrive gli ordini conseguenti — prescrizioni, ricovero, invio a chirurgia. Lo fa compiendo azioni *dentro* una cartella clinica elettronica, come fa un clinico, invece di produrre testo libero che un umano deve trascrivere.

È un passo reale: da un sistema che consiglia a uno che agisce lungo l'intero percorso clinico. È anche esattamente il tipo di passo che, nella copertura, viene schiacciato in "l'AI supera i medici". Quindi vale la pena essere precisi su cosa è stato testato, e dove.

La versione in una frase: MIRA ha lavorato casi interi da solo e, in questo test comparativo, ha superato i medici nell'accuratezza diagnostica — ma lo ha fatto in un ambiente simulato e isolato, su cartelle retrospettive, in otto diagnosi preselezionate, comunicando solo in testo, e gran parte del suo vantaggio è arrivata nelle condizioni con risultati di test più netti. L'avanzamento è reale. Una classifica, però, non è la pratica clinica.




MIRA ha dimostrato di saper gestire dall'inizio alla fine un flusso clinico in una EHR simulata e isolata. Non ha dimostrato un impiego clinico reale, copertura diagnostica ampia o un confronto testa-a-testa equamente informato con i medici.

# Cosa hanno fatto gli autori

MIRA — Medical Intelligence for Reasoning and Action — è un agente autonomo costruito sui modelli di OpenAI: la parte che conversa e agisce usa **GPT-4o**, mentre una fase separata di pianificazione usa il modello di ragionamento **o1** di OpenAI. Questi modelli operano all'interno di un'architettura su misura e conforme agli standard — costruito su HL7 FHIR, lo standard dati che gli ospedali reali usano per far circolare le cartelle — con una cassetta degli attrezzi di oltre ottantamila possibili azioni cliniche. In quell'ambiente isolato MIRA può recuperare l'anamnesi del paziente, ordinare e interpretare laboratorio, imaging e microbiologia, generare una diagnosi differenziale e formulare piani di trattamento — prescrivere farmaci, programmare procedure chirurgiche, pianificare ricoveri. Un dettaglio da segnalare subito: i "giudici" automatici che hanno valutato le risposte di MIRA erano essi stessi GPT-4o — la stessa famiglia di modelli sotto test — cosa che gli autori hanno affiancato a una revisione da parte di un medico specialista abilitato, dopo che un revisore aveva sollevato il problema.

Il set di test veniva da **MIMIC-IV**, un grande database pubblico di cartelle cliniche de-identificate. Da circa 300.000 pazienti trattati al Beth Israel Deaconess Medical Center tra il 2008 e il 2019, gli autori hanno selezionato **574 casi** distribuiti su **otto diagnosi selezionate** — patologie addominali ed emergenze di medicina interna, tra cui appendicite, pancreatite, polmonite e infezione delle vie urinarie. Per ogni caso, MIRA partiva da un quadro limitato e doveva decidere, passo dopo passo, cosa fare — la stessa struttura di un iter diagnostico reale, ma eseguita contro una cartella il cui finale reale era già registrato.

Le sue prestazioni sono state poi confrontate con quella dei medici sugli stessi casi.

> **Cosa significa davvero «agente autonomo in una EHR simulata e isolata»**
>
> Tre parole stanno portando molto peso, quindi conviene spacchettarle.
> **Agente** significa che il sistema non è un chatbot che risponde a un prompt. Esegue un ciclo: guarda lo stato corrente, sceglie un'azione (ordina questo esame, chiede quell'anamnesi), vede il risultato, sceglie l'azione successiva — finché arriva a una diagnosi e a un piano. L'«intelligenza» è un modello linguistico; la capacità di agire (*agency*) viene dall'infrastruttura che traduce il testo del modello in operazioni EHR consentite e gli restituisce i risultati.
> **EHR simulata e isolata** significa una copia controllata di un sistema di cartella clinica, non un ospedale reale. MIRA può "ordinare" un esame e ricevere il risultato che quel paziente ha davvero avuto, perché il caso è storico e la risposta è già registrata. Nulla di ciò che fa tocca un paziente reale o una clinica reale.
> **Autonomo** significa che completa il caso dall'inizio alla fine senza intervento umano nel ciclo — *dentro l'ambiente isolato*. Non significa che sia stato lasciato libero di gestire pazienti senza supervisione. Sono affermazioni molto diverse, e solo la prima è stata testata.
> Un'altra cosa sull'ambiente simulato, perché è facile immaginarla male: MIRA può *ordinare*

> qualsiasi esame voglia, ma il sistema può restituire un risultato solo se quell'esame è stato **davvero eseguito** per quel paziente nella vita reale. Chiede un pannello ematico che il paziente ha ricevuto, e ottiene i valori reali; chiede una scansione che nessuno ha ordinato, e riceve "N/A — could not be performed", mai un numero inventato. L'anamnesi funziona allo stesso modo: se la cartella non contiene ciò che MIRA chiede, il paziente simulato dice semplicemente che non lo sa. Quindi MIRA non può evocare il test decisivo che non è mai stato fatto — può solo lavorare con ciò che l'iter diagnostico reale includeva.
> La distinzione conta perché la parola da titolo — "autonomo" — è quella più facile da leggere come la seconda cosa.

## Cosa hanno trovato

Nel test comparativo, **MIRA ha superato i medici nell'accuratezza diagnostica** — ma il titolo nasconde tre numeri diversi, facili da confondere, quindi conviene separarli.

Da solo, su tutti i **574 casi**, MIRA ha nominato la diagnosi corretta nell'**88,9%** dei casi — valutata rispetto alla diagnosi di dimissione effettivamente registrata per ogni paziente in MIMIC-IV. Per il testa-a-testa con gli umani, i medici non hanno lavorato tutti i 574 casi; hanno lavorato un sottoinsieme condiviso di **311 casi**, usando le stesse cartelle e gli stessi strumenti di MIRA. Su quegli stessi 311 casi, MIRA ha ottenuto **87,8%**, ed è stato misurato contro due gruppi reclutati separatamente. Il primo era un **gruppo senior**: quattro medici specialisti abilitati con 7-11 anni di esperienza, che hanno ottenuto **78,1%**. Il secondo era un **gruppo con livelli di esperienza diversi**: soprattutto specializzandi meno esperti, più due specialisti, più vicino a come è davvero composto un pronto soccorso, che ha ottenuto **71,1%**. Stessi casi, stessi strumenti — e l'ordine è uscito AI prima, medici esperti secondi, team più junior terzo. La formulazione prudente del lavoro è che MIRA ha mostrato prestazioni costantemente equivalenti a quelle dei medici e spesso superiori, in tutte e otto le malattie.

Anche le decisioni a valle hanno retto: in gran parte coerenti con le linee guida, sicure sui farmaci e appropriate sul ricovero. Gli autori riportano nessun errore farmacologico ad alta severità in cinque categorie di sicurezza (interazioni, dosaggio renale, allergie, rischio QT e prescrizione di oppioidi) e prescrizioni corrette in 467 casi su 468 — precisando però che il sistema non ha raggiunto un'affidabilità del 100%. Preso alla lettera, è un risultato notevole: un agente che gestisce l'intero caso, e arriva davanti ai medici sulla diagnosi.

Ma la *forma* della vittoria conta quanto la vittoria. Gli specialisti indipendenti che hanno letto il lavoro hanno indicato da dove arrivava davvero il vantaggio. Nel gruppo di esperti dello Science Media Centre, il dottor Wei Xing (University of Sheffield) ha notato che il numero da titolo — l'AI che batte i medici sull'accuratezza diagnostica — dipendeva **soprattutto dalle condizioni con risultati di test chiari**, come appendicite e pancreatite, dove una scansione o un valore di laboratorio decisivo chiude la questione. Per polmonite e infezioni urinarie — due tra i motivi più comuni per cui le persone arrivano davvero in pronto soccorso — *sia* l'AI *sia* i medici hanno fatto peggio, e il divario tra loro era il più piccolo.

C'era anche un'asimmetria nel modo in cui le due parti hanno giocato, e sta nei numeri del lavoro.

MIRA si è appoggiato di più al laboratorio: ha usato circa il **51%** degli analiti di laboratorio disponibili nella cura di routine, contro circa il **28%** dei medici specialisti abilitati — una mediana di sette parametri ematici in più per caso. Gli autori sono attenti a incorniciare questo valore come *inferiore* alla baseline di routine del dataset, non come una strategia del tipo «ordina tutto», e riportano *nessun* aumento sistematico dell'imaging trasversale più costoso. Però più informazione può, da sola, produrre più accuratezza diagnostica — quindi non è proprio una gara a parità di informazioni, un punto che Xing ha segnalato direttamente. Anche un clinico libero di ordinare ogni esame del sangue economico senza pesare costo, disagio o ritardo apparirebbe più acuto su carta.

## Perché "batte i medici" non è "medico migliore"

Una vittoria in un test comparativo è facile da sovrainterpretare. Tra "MIRA ha ottenuto un punteggio più alto" e "MIRA è migliore" ci sono tre cose.

Primo, **contro cosa è stato valutato**. La professoressa Julie Jacko (University of Edinburgh) ha notato che diversi esiti chiave di MIRA sono definiti rispetto a ciò che era documentato nel set di dati sottostante — il che significa che il sistema viene ricompensato per **riprodurre il comportamento clinico registrato**, non necessariamente per dimostrare cura ottimale. La cartella storica diventa la chiave delle risposte. È un modo ragionevole di costruire un test comparativo, ma misura l'accordo con ciò che è stato fatto, non la correttezza in senso assoluto. Per correttezza verso il lavoro, questo morde soprattutto sulle metriche di *allineamento del trattamento* — gli ordini di MIRA coincidevano con la cartella? — mentre l'accuratezza diagnostica da titolo è valutata rispetto alla diagnosi di dimissione, che è più vicina a un esito reale che a un'eco della documentazione.

Secondo, **l'asimmetria informativa** già citata: molti più esami del sangue (circa il 51% degli analiti disponibili, contro il 28%) significa disporre di più informazioni. Una parte del divario di accuratezza potrebbe quindi derivare dall'accesso a più dati, non da un ragionamento migliore.

Terzo, **in quale contesto è stato testato**. Era un ambiente simulato e isolato, su casi retrospettivi, in otto diagnosi preselezionate, solo in testo. La valutazione clinica reale, come ha detto il dottor Dominic Oliver (University of Oxford), dipende non solo da ciò che i pazienti dicono ma da come lo dicono — insieme a esame obiettivo, comportamento osservato e linguaggio del corpo, nulla di ciò che un agente testuale che legge una cartella conclusa può mai vedere. E un paziente il cui problema non rientra in una delle otto diagnosi scelte è semplicemente fuori da ciò di cui questo studio può parlare.

C'è anche una preoccupazione più silenziosa sollevata dai revisori: **contaminazione dei dati di addestramento**. MIMIC-IV è pubblico e molto discusso, quindi un modello linguistico addestrato su dati provenienti dal web potrebbe avere già visto lavori, discussioni di casi o i dati stessi. Il dottor Midhun Parakkal Unni (University of Sheffield) lo ha segnalato direttamente — se alcune risposte erano nei dati di addestramento, parte delle prestazioni potrebbe dipendere dalla memoria anziché dal ragionamento clinico, e solo una replica indipendente può distinguere le due cose. Va notato che gli autori non liquidano il punto: scrivono che i risultati possono essere interpretati con cautela come un possibile limite superiore e potrebbero sovrastimare la generalizzazione ad altri casi pubblici —

un lavoro che mette un tetto al proprio titolo.

Nulla di questo rende MIRA meno interessante. Rende corretta una lettura calibrata: in un test retrospettivo, un agente capace di agire lungo l'intera cartella ha superato i medici sulle diagnosi con risultati diagnostici netti — in parte ordinando più esami, in parte concordando con la cartella registrata. È una dimostrazione reale di capacità, non un verdetto che le macchine ora diagnosticano meglio dei medici.

## Cosa *non* dimostra

- Non mostra che MIRA funzioni su pazienti reali. Ogni caso era retrospettivo e simulato; nessun paziente è mai stato gestito da lui.
- Non mostra che funzioni oltre otto diagnosi. Le condizioni fuori dal set preselezionato — la maggioranza disordinata e indifferenziata della medicina — non sono state testate.
- Non mostra che sia sicuro da impiegare. Gli autori stessi scrivono che generalizzazione, sicurezza e governance richiedono ancora studi prospettici nel mondo reale.
- Non stabilisce un confronto equo con i medici. Ha usato molti più esami del sangue (circa il 51% degli analiti disponibili contro il 28%), e diversi esiti di trattamento sono stati valutati in parte rispetto alla cartella registrata invece che a un riferimento clinico realmente migliore.
- Non mostra che il risultato sia libero da memorizzazione. I dati pubblici MIMIC-IV potrebbero sovrapporsi ai dati di addestramento del modello, cosa che una replica indipendente dovrebbe escludere.
- Non significa "l'AI sostituisce i medici". È supporto decisionale che può *agire* lungo una cartella; il consenso dei revisori è che l'uso reale avverrà in collaborazione con i clinici, che mantengono l'autorità e forniscono tutto ciò che una cartella testuale lascia fuori.

## Quanto sono solide le prove?

Conviene separare l'affermazione in due, perché le prove hanno una solidità molto diversa per le due parti.

**Come dimostrazione di capacità** — che un agente basato su modello linguistico possa gestire un intero caso clinico dentro una EHR simulata e isolata, concatenando anamnesi, esami, diagnosi e ordini dall'inizio alla fine — il lavoro è davvero nuovo e ragionevolmente convincente. Questa è la parte nuova, e quella a cui vale la pena fare attenzione.

**Come affermazione di superiorità** — che MIRA sia *meglio dei medici* — le prove sono circoscritte e vanno lette con cautela. Vale in uno specifico test retrospettivo, in otto diagnosi, con un'asimmetria informativa (più esami) e una chiave di risposta tratta dalla cartella storica, e con una possibilità viva di contaminazione dei dati di addestramento. È abbastanza per dire "l'agente ha ottenuto risultati notevoli in questo test". Non è abbastanza per dire "l'agente è un diagnostico migliore di un medico", e gli autori non sostengono la seconda cosa.

La lettura più utile non è né "l'AI batte i medici" né "solo un giocattolo". È: un nuovo tipo di AI medica — una che *agisce* lungo la cartella invece di rispondere a domande — è andato bene in un difficile test retrospettivo, e ora deve dimostrare ciò che non è ancora stato provato: pazienti reali, indifferenziati, prospetticamente, sotto governance.

## Perché conta

Da qualche anno la domanda interessante nell'AI medica è cambiata in silenzio. Prima era "un modello può trovare la diagnosi giusta?" — e i modelli continuavano a rispondere sì, su set di test sempre più selezionati. MIRA segna lo spostamento verso una domanda più dura: "un modello può *fare il lavoro* — raccogliere, ordinare, interpretare, decidere, agire — lungo un intero percorso clinico?" È una domanda più utile e più onesta, perché agire è dove stanno sia la difficoltà sia il rischio.

Per questo conta il modo in cui il risultato viene presentato. Il riflesso da titolo — *l'AI supera i medici* — punta alla parte meno nuova e meno solida del risultato. La cosa davvero nuova è più piccola e più conseguente: un agente che può muoversi lungo l'intera cartella. Quella capacità, se regge, cambia il **flusso di lavoro** molto prima di cambiare chi ne ha la responsabilità. Il medico non viene sostituito; l'ordinare, l'interpretare, il rincorrere risultati — il flusso di lavoro — è il primo ambito in cui uno strumento simile avrebbe effetti.

E rimette l'onere della prova nella direzione giusta. Una vittoria in classifica su casi retrospettivi è un motivo per fare lo studio prospettico, non un sostituto. Gli autori lo dicono. La lettura onesta di MIRA è un invito a quello studio successivo — tenuto allo standard che il campo conosce già: misurato sui pazienti, non soltanto nei test comparativi.

## In breve

MIRA è un agente AI autonomo che opera una cartella clinica elettronica simulata e isolata: può raccogliere anamnesi, ordinare e interpretare laboratorio, imaging e microbiologia, arrivare a una diagnosi e scrivere piani di trattamento. Testato su 574 casi retrospettivi dal database pubblico MIMIC-IV, in otto diagnosi preselezionate, ha superato i medici nell'accuratezza diagnostica (88,9% complessivo; 87,8% contro 78,1% nel testa-a-testa) e preso decisioni in gran parte coerenti con le linee guida e sicure farmacologicamente. Ma la valutazione era una simulazione su cartelle passate, solo testuale; gran parte del vantaggio è arrivato in condizioni con risultati di test netti; MIRA ha usato molti più esami del sangue dei medici (circa il 51% degli analiti disponibili contro il 28%); diversi esiti di trattamento sono stati valutati rispetto a ciò che la cartella originale registrava; e il set di dati pubblico solleva un rischio reale di contaminazione dei dati di addestramento — che gli autori stessi chiamano un possibile tetto superiore dei loro numeri. Il vero avanzamento è un agente che agisce lungo l'intero percorso clinico invece di rispondere a domande isolate. Gli autori sono espliciti: generalizzazione, sicurezza e governance richiedono ancora studi prospettici nel mondo reale — che è la lettura corretta: una dimostrazione di capacità impressionante, non la prova che l'AI diagnostichi

meglio dei medici, e non un sistema pronto per una clinica reale.

## Valutazione critica

**Cosa mostra il lavoro:** Un agente basato su modello linguistico (MIRA) può gestire un intero caso clinico dall'inizio alla fine dentro una EHR simulata e isolata — anamnesi, esami, diagnosi, ordini — e, in un test retrospettivo di 574 casi in otto diagnosi, ha superato i medici nell'accuratezza diagnostica (88,9% complessivo; 87,8% contro 78,1% rispetto ai medici specialisti abilitati) senza errori farmacologici ad alta severità.

**Cosa è plausibile ma non provato:** Che questa capacità si traduca in beneficio per pazienti reali e indifferenziati; che il vantaggio di accuratezza rifletta ragionamento migliore invece di più esami ordinati, accordo con la cartella registrata o dati pubblici memorizzati.

**Cosa non mostra:** Che MIRA funzioni fuori dalle sue otto diagnosi; che sia sicuro da impiegare; che batta i medici in un confronto equo e ugualmente informato; che sostituisca i clinici; che i risultati siano liberi da contaminazione dei dati di addestramento.

**Limitazioni principali:** Valutazione retrospettiva, simulata e solo testuale; otto diagnosi preselezionate; asimmetria informativa (molti più esami del sangue — circa il 51% degli analiti disponibili contro il 28%, negli esami ematici a basso costo, non nell'imaging); esiti di trattamento valutati in parte rispetto alla cartella storica; e un set di dati pubblico (MIMIC-IV) che un modello linguistico potrebbe avere visto nei dati di addestramento — cosa che gli autori segnalano come possibile limite superiore delle prestazioni.

**Quanta fiducia dovrebbe avere un lettore generale?** Alta che MIRA sia una dimostrazione reale e nuova di capacità — un agente che *agisce* lungo la cartella, non solo un chatbot. Bassa che sia un *diagnostico migliore di un medico*: quell'affermazione è limitata a un test retrospettivo e condizionata dall'ordine degli esami, valutazione rispetto alla cartella e possibile contaminazione. La conclusione degli autori è quella giusta — servono studi prospettici nel mondo reale prima che significhi qualcosa per la cura dei pazienti.

## Fonti

Ferber et al., Towards autonomous medical artificial intelligence agents, Nature (2026)
https://doi.org/10.1038/s41586-026-10675-5

PubMed 42310457 — peer-reviewed abstract
https://pubmed.ncbi.nlm.nih.gov/42310457/

Science Media Centre — expert reaction to MIRA and AMIE (2026)
https://www.sciencemediacentre.org/expert-reaction-to-presentation-of-two-new-medical-ai-models-for-

News-Medical (2026) — illustrates the 'outperforms doctors' framing

https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-simulated-EH
aspx