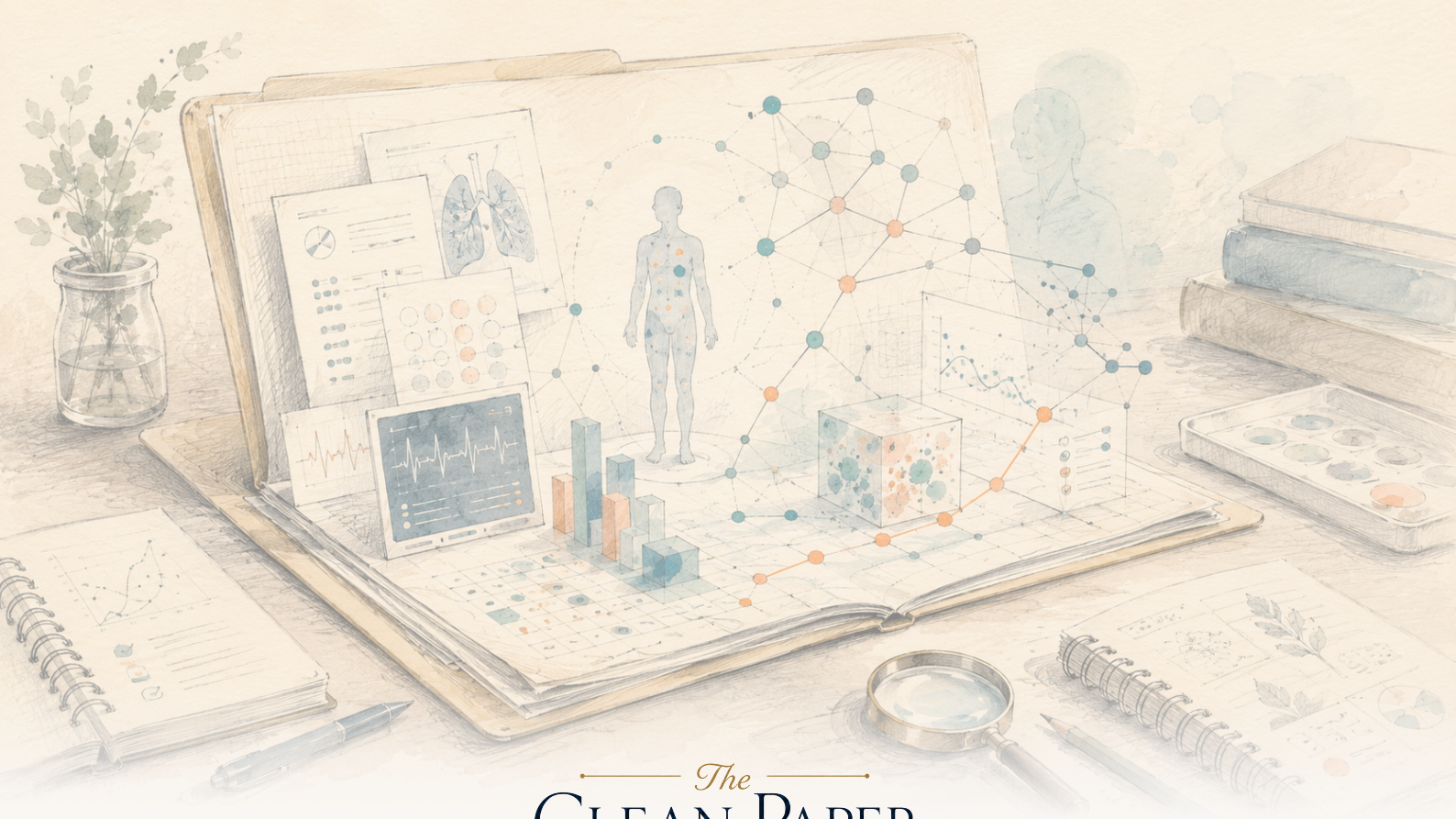

The Clean Paper

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

Un agent AI a lucrat singur cazuri întregi de pacienți — într-un simulator, pe dosare vechi

MIRA este un nou tip de AI medicală: în loc să răspundă la o singură întrebare, lucrează un caz întreg într-un dosar medical simulat — ia anamneza, comandă și citește teste, ajunge la un diagnostic, scrie ordinele. Pe 574 de cazuri retrospective dintr-o bază de date publică, în opt diagnostice preselectate, autorii raportează că a depășit medicii la acuratețea diagnostică și a luat decizii în mare parte conforme cu ghidurile și sigure medicamentos. Fiecare calificare din acea frază contează: a rulat într-o sandbox pe dosare trecute, doar în text; mare parte din avantaj a venit în afecțiunile cu rezultate de test clare; a cerut cam de două ori mai multe analize de sânge; iar mai multe rezultate au fost evaluate față de ce consemnase fișa originală. Avansul real este un agent care acționează de-a lungul întregului workflow — nu dovada că o mașină diagnostichează acum mai bine decât un medic. Autorii spun asta: generalizarea, siguranța și guvernanța au încă nevoie de studii prospective în lumea reală.

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

A răspunde la o întrebare nu este același lucru cu a gestiona cazul

Cele mai multe povești despre AI medicală sunt despre un model care *răspunde*. Îi dai un caz clinic sau o întrebare de examen, iar el oferă un diagnostic ori un paragraf de recomandări și obține un scor bun. Este util, dar limitat: seamănă cu un consultant isteț care nu lucrează niciodată direct în dosarul pacientului.

MIRA este construit să facă altceva, iar aici se află noutatea reală. În loc să răspundă la o singură întrebare, gestionează un caz întreg: citește dosarul pacientului, decide ce informații din anamneză îi mai lipsesc, solicită analize de laborator, investigații imagistice și microbiologice, interpretează rezultatele, restrânge diagnosticul diferențial și apoi formulează acțiunile următoare — prescripții, internare sau trimitere către chirurgie. Face toate acestea acționând *în interiorul* unui dosar medical electronic, asemenea unui clinician, nu doar generând text pe care un om trebuie să-l transcrie.

Este un pas real: de la un sistem care oferă sfaturi la unul care acționează de-a lungul întregului flux de lucru. Este și exact genul de progres pe care relatările din presă îl pot reduce la „AI depășește medicii". De aceea merită precizat exact ce a fost testat și în ce condiții.

Pe scurt: MIRA a gestionat singur cazuri complete și, pe acest test de referință, a depășit medicii la acuratețea diagnosticului. Dar a făcut-o într-un mediu simulat și izolat, pe dosare retrospective, pentru opt diagnostice alese în prealabil și folosind numai text; în plus, o mare parte din avantaj a apărut în afecțiunile cu rezultate de test foarte clare. Progresul este real. Clasamentul nu este însă o clinică.

MIRA a arătat o capacitate de flux de lucru, nu un verdict clinic

Un test de referință într-un sandbox permite unui agent să parcurgă un caz retrospectiv întreg. Nu este încă un studiu pe pacienți reali.

DEMONSTRAT

- gestionare end-to-end a cazului într-un EHR sandboxed
- anamneză, teste, diagnostic diferențial, ordine
- 574 de cazuri retrospective MIMIC-IV
- 8 diagnostice preselectate
- acuratețe diagnostică mai mare pe acest benchmark

NEDEMONSTRAT

- pacienți reali sau implementare live în spital
- afecțiuni dincolo de cele opt ținte
- o comparație directă cu informații egale
- absența contaminării cu date publice
- siguranță și guvernanță pentru uz clinic

O capacitate arătată în sandbox -- nu un diagnostician dovedit în clinică.

Figura separă rezultatul de workflow de afirmația despre implementare.

MIRA a demonstrat că poate parcurge cap-coadă un flux de lucru într-un dosar medical electronic izolat. Nu a demonstrat utilizare clinică reală, acoperire diagnostică largă sau o comparație directă și echitabilă cu medicii, în condiții de acces egal la informații.

Ce au făcut autorii

MIRA — Medical Intelligence for Reasoning and Action — este un agent autonom construit pe modele OpenAI: componenta care conversează și acționează folosește **GPT-4o**, iar o etapă separată de planificare folosește modelul de raționament **o1**. Sistemul este integrat într-un cadru personalizat, conform standardelor, construit pe HL7 FHIR — standardul folosit pentru schimbul de date medicale — și dispune de peste optzeci de mii de acțiuni clinice posibile. În mediul izolat, MIRA poate extrage istoricul pacientului, solicita și interpreta analize de laborator, imagistică și microbiologie, construi un diagnostic diferențial și formula planuri de tratament, inclusiv prescripții, proceduri chirurgicale și internări. Un detaliu important: evaluatorii automați care au punctat răspunsurile MIRA foloseau tot GPT-4o — aceeași familie de modele ca sistemul testat. După ce un evaluator al manuscrisului a ridicat această problemă, autorii au adăugat și verificarea de către un medic specialist certificat.

Setul de testare a provenit din **MIMIC-IV**, o bază publică mare de dosare medicale electronice de-identificate. Din aproximativ 300.000 de pacienți tratați la Beth Israel Deaconess Medical Center între 2008 și 2019, autorii au selectat **574 de cazuri** care acopereau **opt diagnostice țintă** — patologii abdominale și urgențe de medicină internă, printre ele apendicită, pancreatită, pneumonie și infecție urinară. Pentru fiecare caz, MIRA pornea de la informații limitate și trebuia să decidă, pas cu pas, ce să facă mai departe — o structură asemănătoare unei evaluări clinice reale, dar rulată pe un dosar al cărui deznodământ era deja înregistrat.

Performanța lui a fost apoi comparată cu a medicilor pe aceleași cazuri.

> **Ce înseamnă de fapt „agent autonom într-un dosar electronic izolat”**
>
> Trei termeni sunt esențiali aici și merită clarificați.
> **Agent** înseamnă că sistemul nu este un chatbot care răspunde la un prompt. Funcționează într-o buclă: examinează starea curentă, alege o acțiune — de exemplu solicită un test sau cere o informație din anamneză — primește rezultatul și decide pasul următor, până ajunge la un diagnostic și la un plan. „Inteligența” vine din modelul lingvistic; componenta de agent este infrastructura din jurul lui, care transformă ieșirea modelului în operațiuni permise în dosarul electronic și îi transmite rezultatele.
> **Dosar electronic izolat** înseamnă o copie simulată a unui sistem de dosar medical, nu un sistem activ într-un spital. MIRA poate „solicita” un test și primi rezultatul pe care pacientul l-a avut în realitate, deoarece dosarul este istoric, iar rezultatul este deja înregistrat. Nimic din ceea ce face nu afectează un pacient sau o clinică reală.
> **Autonom** înseamnă că duce cazul de la început la sfârșit fără intervenția unui om — *în interiorul mediului izolat*. Nu înseamnă că a fost lăsat să gestioneze nesupravegheat pacienți reali. Sunt

afirmații foarte diferite, iar numai prima a fost testată.
Mai este un detaliu important despre acest mediu izolat: MIRA poate *solicita* orice test, dar sistemul oferă un rezultat numai dacă acel test a fost **efectuat în realitate** pentru pacientul respectiv. Dacă solicită un set de analize pe care pacientul l-a făcut, primește valorile reale; dacă solicită o investigație care nu a fost efectuată, primește „N/A — could not be performed", nu o valoare inventată. La fel se întâmplă cu istoricul: dacă informația nu există în dosar, pacientul simulat răspunde că nu știe. MIRA nu poate, așadar, obține un test decisiv care nu a fost făcut niciodată; poate lucra doar cu informațiile disponibile în evaluarea clinică reală.
Distincția contează pentru că termenul de titlu — „autonom" — este cel mai probabil să fie citit ca a doua afirmație.

Ce au găsit

Pe testul de referință, **MIRA a depășit medicii la acuratețea diagnosticului** — dar această formulare comprimă trei cifre diferite, care merită separate.

Singur, pe toate cele **574 de cazuri**, MIRA a numit diagnosticul corect în **88,9%** din cazuri — punctat față de diagnosticul de externare efectiv înregistrat pentru fiecare pacient în MIMIC-IV. Pentru comparația directă cu oamenii, medicii nu au lucrat toate cele 574 de cazuri; au lucrat un subset comun de **311 cazuri**, folosind aceleași dosare și instrumente pe care le avea MIRA. Pe aceleași 311 cazuri, MIRA a obținut **87,8%** și a fost măsurat față de două grupuri recrutate separat. Primul era un **grup de medici cu experiență**: patru medici specialiști certificați, cu 7–11 ani de experiență, care au obținut **78,1%**. Al doilea era un **grup cu niveluri diferite de experiență**: în principal rezidenți aflați la începutul formării, plus doi specialiști, o structură mai apropiată de personalul real al unui departament de urgență; acest grup a obținut **71,1%**. Aceleași cazuri, aceleași instrumente — iar ordinea a ieșit AI primul, medicii experimentați pe locul doi, echipa mai junior pe trei. Formularea prudentă a lucrării este că MIRA a fost „consistently equivalent to, and often exceeded" medicii, în toate cele opt boli.

Și deciziile ulterioare au fost, în mare parte, conforme cu ghidurile, adecvate pentru internare și sigure din punct de vedere medicamentos. Autorii raportează zero erori medicamentoase de severitate mare în cinci categorii de siguranță (interacțiuni medicamentoase, dozare renală, alergii, risc QT și prescriere de opioide) și prescripții corecte în 467 din 468 de cazuri — notând totuși că sistemul „did not achieve 100% reliability". Luat ca atare, este un rezultat izbitor: un agent care rulează întregul caz și iese înaintea medicilor la diagnostic.

Dar *felul* în care a fost obținut acest avantaj contează la fel de mult ca avantajul în sine. Specialiștii independenți care au citit lucrarea au arătat de unde venea de fapt avantajul. În panelul de experți al Science Media Centre, dr. Wei Xing (University of Sheffield) a notat că cifra de titlu — AI-ul care bate medicii la acuratețe diagnostică — era **condusă mai ales de afecțiunile cu rezultate de test clare**, precum apendicita și pancreatita, unde o scanare decisivă sau o valoare de laborator închide întrebarea. Pentru pneumonie și infecții urinare — două dintre cele mai comune motive pentru care oamenii ajung efectiv la urgențe — *și* AI-ul, *și* medicii au făcut cel mai prost, iar diferența dintre ei era cea mai mică.

A existat și o asimetrie între cele două părți, vizibilă în cifre. MIRA s-a bazat mai mult pe analize de laborator: a folosit aproximativ **51%** dintre analiții disponibili în îngrijirea de rutină, față de aproximativ **28%** pentru medicii specialiști — o mediană de șapte parametri sanguini suplimentari per caz. Autorii precizează că nivelul rămânea *sub* practica de rutină din setul de date și că MIRA nu solicita sistematic mai multă imagistică transversală, mai costisitoare. Totuși, accesul la mai multe informații poate crește de la sine acuratețea diagnosticului. Comparația nu este, așadar, perfect echivalentă în privința informațiilor disponibile, lucru semnalat direct de Xing. Un clinician liber să solicite toate analizele ieftine, fără să cântărească costul, disconfortul sau întârzierea, ar putea de asemenea să pară mai performant într-o astfel de evaluare.

De ce „a bătut medicii" nu înseamnă „medic mai bun"

Un avantaj pe un test de referință este ușor de suprainterpretat. Între „MIRA a avut scor mai mare" și „MIRA este mai bun" stau trei lucruri.

Mai întâi, contează **criteriul după care a fost evaluat**. Profesoara Julie Jacko (University of Edinburgh) a notat că mai multe dintre rezultatele-cheie ale MIRA sunt definite relativ la ce era documentat în setul de date de bază — ceea ce înseamnă că sistemul este recompensat pentru **a reproduce comportamentul clinic înregistrat**, nu neapărat pentru a demonstra îngrijire optimă. Fișa istorică devine cheia de răspuns. Este o metodă rezonabilă de a construi un test de referință, dar măsoară acordul cu ce s-a făcut, nu corectitudinea într-un sens absolut. Ca să fim corecți cu lucrarea, asta mușcă cel mai tare din metricile de *aliniere a tratamentului* — ordinele MIRA se potrivesc cu fișa? — în timp ce acuratețea diagnosticului, evidențiată în titlu, este evaluată față de diagnosticul de externare, care este mai aproape de un rezultat real decât de un ecou al documentației.

În al doilea rând, există **asimetria informațională** deja menționată: mult mai multe analize de sânge (aproximativ 51% dintre analiții disponibili, față de 28%) înseamnă mai multe informații pe care se poate baza diagnosticul. O parte din diferența de acuratețe poate proveni din acest acces suplimentar, nu doar dintr-un raționament mai bun.

În al treilea rând, contează **mediul în care a fost testat**. A fost un sistem izolat, pe cazuri retrospective, limitat la opt diagnostice preselectate și la informații textuale. Evaluarea clinică reală, cum a spus dr. Dominic Oliver (University of Oxford), depinde nu doar de ce spun pacienții, ci de cum o spun — împreună cu examinarea fizică, comportamentul observat și limbajul corpului, niciuna dintre acestea nefiind vizibilă unui agent textual care citește un dosar încheiat. Iar un pacient a cărui problemă nu intră într-unul dintre cele opt diagnostice alese este pur și simplu în afara a ceea ce poate spune acest studiu.

Mai există și o îngrijorare mai tăcută ridicată de evaluatori: **contaminarea datelor**. MIMIC-IV este public și intens discutat, așa că un model lingvistic antrenat pe internetul deschis poate să fi văzut deja lucrări, discuții de caz sau datele în sine. Dr. Midhun Parakkal Unni (University of Sheffield) a semnalat direct această posibilitate: dacă unele răspunsuri se aflau în datele de antrenare, o parte din performanță ar putea reflecta memorare, nu raționament clinic, iar numai o replicare independentă poate separa cele două explicații. Notabil, autorii nu dau din mână peste punct: scriu că rezultatele

lor „could be cautiously interpreted as a possible upper bound” și „may overestimate generalization to other public cases” — o lucrare care pune un plafon propriului titlu.

Nimic din toate acestea nu face MIRA mai puțin interesant. Impune doar o interpretare calibrată: pe un test retrospectiv, un agent capabil să acționeze de-a lungul întregului dosar i-a depășit pe medici mai ales la diagnosticele cu rezultate de test clare — parțial folosind mai multe analize și parțial prin concordanța cu dosarul înregistrat. Este o demonstrație reală de capacitate, nu un verdict că sistemele AI diagnostichează în general mai bine decât medicii.

Ce *nu* demonstrează

- Nu arată că MIRA funcționează pe pacienți reali. Fiecare caz a fost retrospectiv și simulat; niciun pacient nu a fost vreodată gestionat de el.
- Nu arată că funcționează dincolo de opt diagnostice. Afecțiunile din afara setului preselectat — majoritatea dezordonată și nediferențiată a medicinei — nu au fost testate.
- Nu arată că este sigur de implementat. Autorii înșiși scriu că generalizarea, siguranța și guvernanța au încă nevoie de studii prospective în lumea reală.
- Nu stabilește o comparație perfect echitabilă cu medicii. A folosit mult mai multe analize de sânge (aproximativ 51% dintre analiții disponibili față de 28%), iar mai multe rezultate privind tratamentul au fost evaluate parțial prin concordanța cu dosarul înregistrat, nu cu un standard clinic independent al celei mai bune îngrijiri.
- Nu arată că rezultatul este lipsit de posibilul efect al memorării. Datele publice MIMIC-IV se pot suprapune cu datele de antrenare ale modelului, posibilitate pe care o replicare independentă ar trebui să o excludă.
- Nu înseamnă „AI înlocuiește medicii”. Este suport decizional care poate *acționa* de-a lungul unui dosar; consensul evaluatorilor este că folosirea reală va fi în parteneriat cu clinicienii, care păstrează autoritatea și aduc tot ce un dosar textual lasă pe dinafară.

Cât de solide sunt dovezile?

Afirmația trebuie împărțită în două, deoarece forța dovezilor diferă mult între cele două componente.

Ca demonstrație de capacitate — faptul că un agent bazat pe un model lingvistic poate parcurge un caz clinic întreg într-un dosar electronic izolat, legând anamneza, testele, diagnosticul și acțiunile medicale de la început până la sfârșit — lucrarea este cu adevărat nouă și destul de convingătoare. Aceasta este partea care merită cea mai multă atenție.

Ca afirmație de superioritate — că MIRA este *mai bun decât medicii* — dovezile sunt limitate și trebuie interpretate cu grijă. Rezultatul se referă la un test retrospectiv specific, cu opt diagnostice, o asimetrie informațională (mai multe teste), criterii de evaluare derivate parțial din dosarul istoric și posibilitatea contaminării datelor de antrenare. Este suficient pentru a spune că „agentul a avut performanțe impresionante pe acest test”. Nu este suficient pentru a spune că „agentul este un diag-

nostician mai bun decât un medic”, iar autorii nu susțin această a doua afirmație.

Concluzia cea mai utilă nu este nici „AI bate medicii”, nici „este doar o jucărie”. Un nou tip de AI medicală — unul care *acționează* de-a lungul dosarului în loc să răspundă doar la întrebări — s-a descurcat bine pe un test retrospectiv dificil. Următorul pas este să demonstreze aceeași capacitate acolo unde nu a fost încă încercată: prospectiv, pe pacienți reali și neselectați, într-un cadru de guvernanță clinică.

De ce contează

De câțiva ani, întrebarea interesantă din AI medicală se schimbă treptat. La început era „poate un model să identifice diagnosticul?”, iar modelele răspundeau tot mai des afirmativ pe seturi de test atent construite. MIRA mută discuția spre o întrebare mai dificilă: „poate un model *să desfășoare efectiv activitatea* — să adune informații, să solicite teste, să interpreteze, să decidă și să acționeze — de-a lungul unui întreg flux de lucru?” Este o întrebare mai utilă și mai realistă, pentru că tocmai în acțiune apar atât dificultatea, cât și riscul.

De aceea contează felul în care este prezentat rezultatul. Reflexul de titlu — *AI depășește medicii* — pune accentul pe partea cea mai puțin nouă și mai puțin bine susținută. Noutatea reală este mai restrânsă, dar mai importantă: un agent care se poate deplasa prin întregul dosar și poate executa acțiuni. Dacă această capacitate se confirmă, ea va schimba **fluxul de lucru** cu mult înainte să schimbe cine poartă responsabilitatea. Medicul nu este înlocuit; solicitarea testelor, interpretarea și urmărirea rezultatelor sunt primele locuri în care un astfel de instrument ar putea interveni.

Și reașază standardul dovezilor în direcția corectă. Un rezultat bun într-un clasament pe cazuri retrospective este un motiv pentru a face un studiu prospectiv, nu un înlocuitor pentru acesta. Autorii spun explicit acest lucru. Interpretarea prudentă a MIRA este, așadar, o invitație la studiul următor — evaluat pe pacienți, nu doar pe teste de referință.

Pe scurt

MIRA este un agent AI autonom care operează într-un dosar medical electronic izolat: poate lua istoricul, comanda și interpreta laborator, imagistică și microbiologie, ajunge la un diagnostic și scrie planuri de tratament. Testat pe 574 de cazuri retrospective din baza publică MIMIC-IV, în opt diagnostice preselectate, a depășit medicii la acuratețea diagnostică (88,9% total; 87,8% versus 78,1% în comparația directă) și a luat decizii în mare parte conforme cu ghidurile și sigure medicamentos. Dar evaluarea a fost o simulare pe dosare trecute, doar în text; mare parte din avantaj a venit în afecțiuni cu rezultate de test clare; MIRA a folosit mult mai multe analize de sânge decât medicii (aproximativ 51% dintre analiții disponibili față de 28%); mai multe rezultate de tratament au fost punctate față de ce consemnase fișa originală; iar setul de date public ridică un risc real de contaminare a datelor de antrenare — pe care autorii înșiși îl numesc un posibil plafon superior al numerelor lor. Avansul real este un agent care acționează de-a lungul întregului flux de lucru, nu doar răspunde la întrebări

izolate. Autorii sunt expliciți că generalizarea, siguranța și guvernanța încă cer studii prospective în lumea reală — ceea ce este lectura corectă: o demonstrație impresionantă de capacitate, nu dovada că AI diagnostichează mai bine decât medicii și nu un sistem gata pentru o clinică reală.

Verificare fără exagerări

Ce arată lucrarea: Un agent bazat pe un model lingvistic (MIRA) poate parcurge un caz clinic întreg, de la început la sfârșit, într-un dosar electronic izolat — anamneză, teste, diagnostic și acțiuni medicale — și, pe un test retrospectiv cu 574 de cazuri și opt diagnostice, a depășit medicii la acuratețea diagnosticului (88,9% în total; 87,8% față de 78,1% pentru medicii specialiști certificați), fără erori medicamentoase de severitate mare.

Ce este plauzibil, dar nedemonstrat: Că această capacitate se traduce în beneficiu pentru pacienți reali, nediferențiați; că avantajul de acuratețe reflectă raționament mai bun, nu mai multe teste comandate, acord cu fișa înregistrată sau date publice memorate.

Ce nu arată: Că MIRA funcționează în afara celor opt diagnostice; că este sigur de implementat; că bate medicii într-o comparație corectă, cu informații egale; că înlocuiește clinicienii; că rezultatele sunt lipsite de contaminarea datelor de antrenare.

Limitări principale: Evaluare retrospectivă, simulată și exclusiv textuală; opt diagnostice preselectate; asimetrie informațională (mult mai multe analize de sânge — aproximativ 51% dintre analiții disponibili față de 28%, în principal analize ieftine, nu imagistică); unele rezultate privind tratamentul evaluate parțial prin raportare la dosarul istoric; și un set public de evaluare (MIMIC-IV) pe care modelul lingvistic l-ar fi putut întâlni în datele de antrenare — posibilitate pe care autorii o descriu drept un motiv pentru care performanța ar putea reprezenta un plafon superior.

Câtă încredere ar trebui să aibă un cititor general? Mare că MIRA reprezintă o demonstrație nouă și reală de capacitate — un agent care *acționează* de-a lungul dosarului, nu doar un chatbot. Mică în afirmația că este un *diagnostician mai bun decât un medic*: această concluzie se limitează la un test retrospectiv și este complicată de numărul diferit de teste solicitate, de evaluarea parțială prin raportare la dosar și de posibilitatea contaminării datelor. Concluzia autorilor este cea potrivită: sunt necesare studii prospective în lumea reală înainte ca rezultatul să poată fi transpus în îngrijirea pacienților.

Surse

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