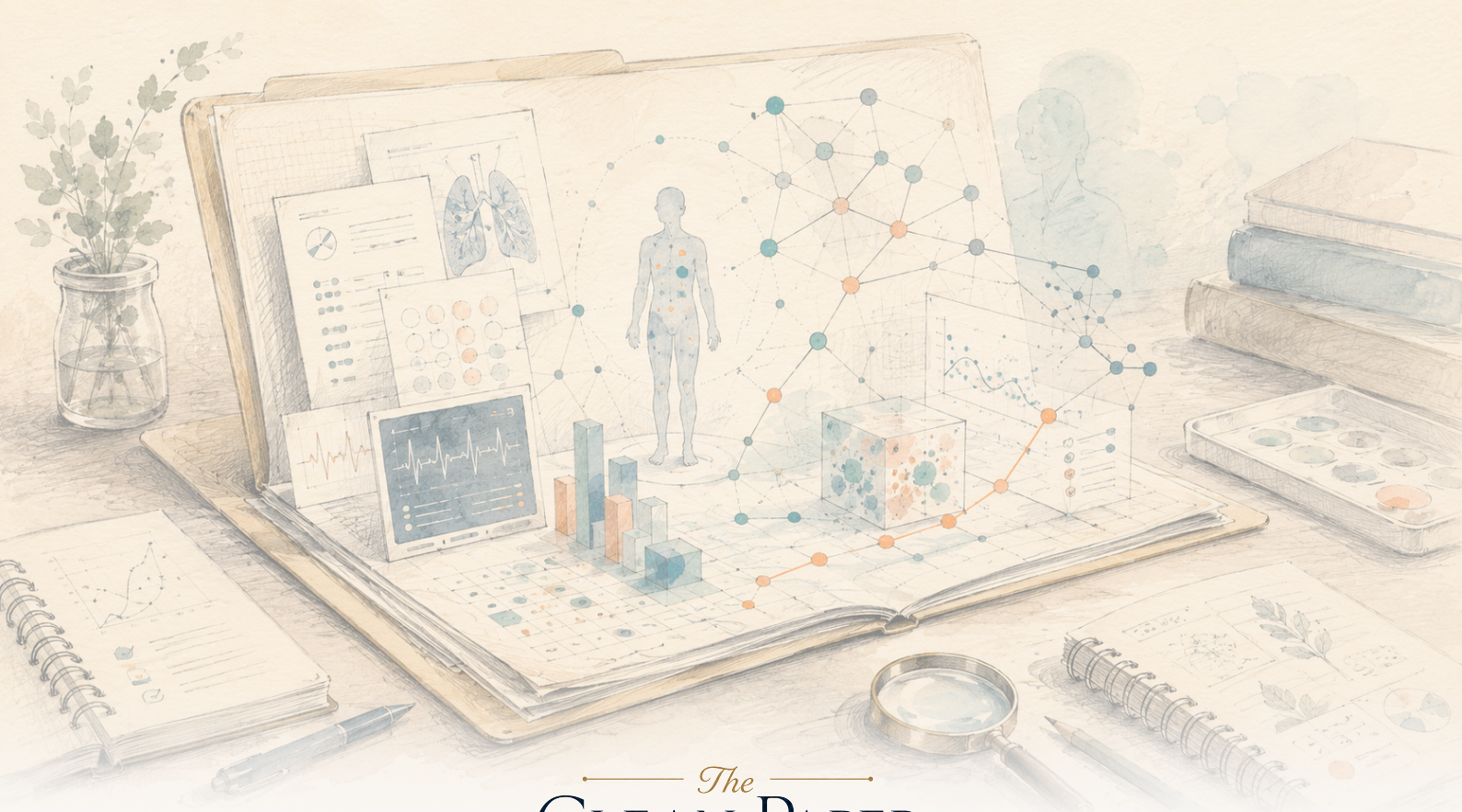


The Clean Paper

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

AI agent samostalno je obrađivao cijele slučajeve pacijenata — u simulatoru, na starim kartonima

MIRA je nova vrsta medicinske vještačke inteligencije: umjesto da odgovori na jedno pitanje, obrađuje cijeli slučaj unutar simuliranog bolničkog kartona — uzima anamnezu, naručuje i čita pretrage, dolazi do dijagnoze i piše naloge. Na 574 retrospektivna slučaja iz javne baze, kroz osam unaprijed odabranih dijagnoza, autori navode da je nadmašila ljekare u dijagnostičkoj tačnosti i donosila odluke koje su uglavnom bile usklađene sa smjernicama i sigurne u pogledu lijekova. Svaka ograda u toj rečenici važna je: radila je u sandboxu na starim kartonima i samo u tekstu; veliki dio prednosti postigla je kod stanja s jasnim rezultatima pretraga; naručivala je približno dvostruko više krvnih pretraga od ljekara; a nekoliko ishoda ocjenjivano je prema onome što je zabilježeno u izvornom kartonu. Stvarni napredak je agent koji djeluje kroz cijeli tok rada — a ne dokaz da mašina sada postavlja dijagnoze bolje od ljekara. Autori to i sami kažu: generalizacija, sigurnost i upravljanje i dalje zahtijevaju prospektivne studije u stvarnom svijetu.

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

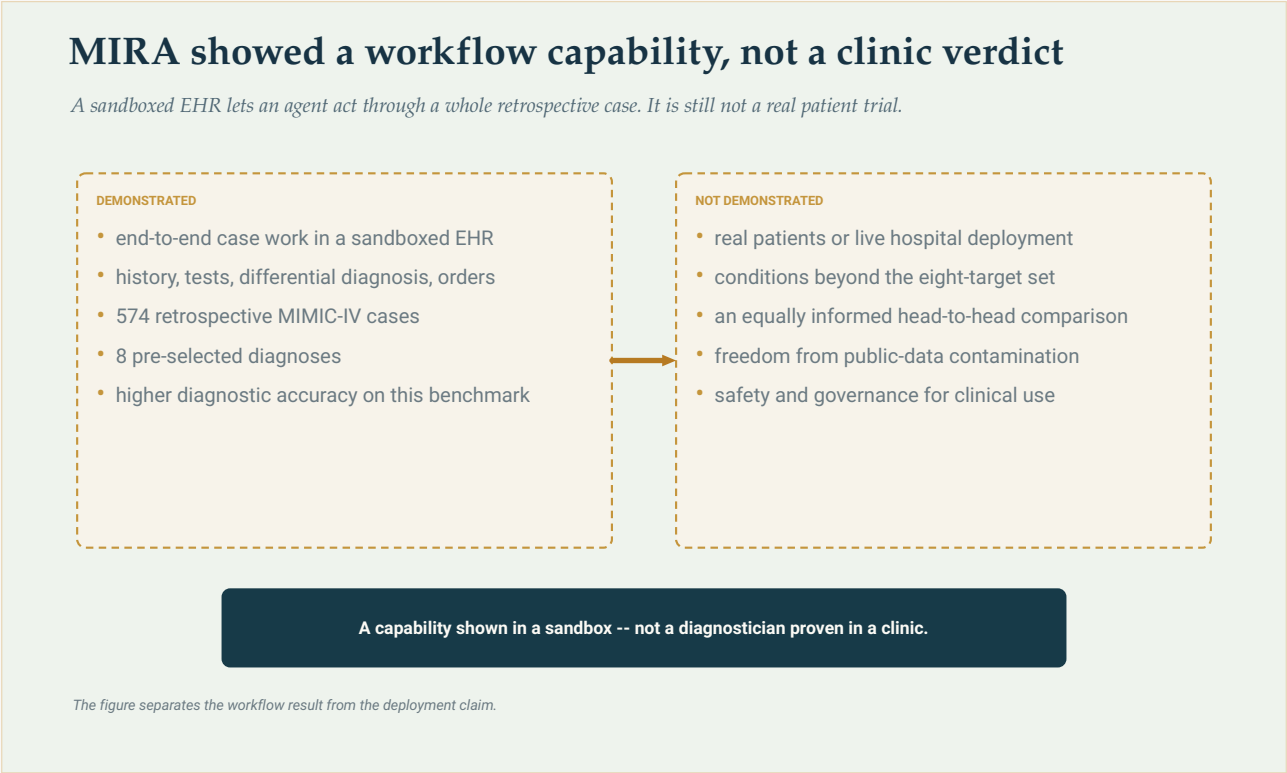
Odgovoriti na pitanje nije isto što i voditi cijeli slučaj

Većina priča o medicinskoj vještačkoj inteligenciji govori o modelu koji *odgovara*. Date mu kliničku vinjetu ili ispitno pitanje, a on vrati dijagnozu ili odlomak savjeta i postigne dobar rezultat. Korisno, ali usko: pametan konzultant koji nikada ne dodirne karton.

MIRA je napravljena za nešto drugo, a upravo je ta razlika stvarna novost. Umjesto odgovora na jedno pitanje, obrađuje cijeli slučaj: čita pacijentov karton, odlučuje koje podatke iz anamneze još treba, naručuje laboratorijske, slikovne i mikrobiološke pretrage, čita rezultate, sužava diferencijalnu dijagnozu i zatim piše naloge koji slijede — recepte, prijam, upućivanje na operaciju. To čini izvodeći radnje *unutar* elektroničkog zdravstvenog kartona, poput kliničara, umjesto da samo proizvodi slobodan tekst koji bi čovjek morao prepisati.

To je stvaran korak: od sistema koji savjetuje prema sistemu koji djeluje kroz cijeli tok rada. I upravo je to vrsta koraka koja se u medijskom izvještavanju lako spljošti u „AI nadmašuje ljekare". Zato vrijedi precizno reći što je testirano i gdje.

Verzija u jednoj rečenici: MIRA je samostalno obrađivala cijele slučajeve i na ovom benchmarku nadmašila ljekare u dijagnostičkoj tačnosti — ali učinila je to u sandboxu, na retrospektivnim kartonima, kroz osam unaprijed odabranih dijagnoza, komunicirajući samo tekstom, a velik dio prednosti ostvarila je kod stanja s najjasnijim rezultatima pretraga. Napredak je stvaran. Rezultat na skali nije ambulanta.



MIRA je pokazala sposobnost vođenja cijelog toka rada unutar sandboxiranog EHR-a. Nije pokazala kliničku upotrebu u stvarnom svijetu, široku dijagnostičku pokrivenost ni poštenu uporedbu s ljekarima koji raspolažu jednakom količinom informacija.

Original Aurora diagram — The Clean Paper CC BY 4.0

Šta su autori uradili

MIRA — Medical Intelligence for Reasoning and Action — autonomni je agent izgrađen na OpenAI-jevim modelima: dio koji razgovara i djeluje koristi **GPT-4o**, a zaseban korak planiranja koristi OpenAI-jev model za rasuđivanje **o1**. Smješteni su u prilagođeni okvir usklađen sa standardima — izgrađen na HL7 FHIR-u, standardu podataka kojim stvarne bolnice razmjenjuju kartone — s više od osamdeset hiljada mogućih kliničkih radnji. Unutar tog sandboxa MIRA može dohvatiti anamnezu, naručiti i protumačiti laboratorijske, slikovne i mikrobiološke pretrage, izraditi diferencijalnu dijagnozu i formulirati plan liječenja — propisati lijekove, zakazati hirurški zahvat, planirati prijam. Vrijedi odmah istaknuti jednu pojedinost: automatizirani „suci" koji su ocjenjivali MIRINE odgovore i sami su bili GPT-4o — ista porodica modela koja se testira — pa su autori, nakon što je recenzent upozorio na problem, dodali pregled specijalista s odgovarajućom certifikacijom.

Testni skup potjecao je iz **MIMIC-IV**, velike javno dostupne anonimizirane baze elektroničkih zdravstvenih kartona. Od otprilike 300,000 pacijenata liječenih u Beth Israel Deaconess Medical Centeru između 2008. i 2019., autori su odabrali **574 slučaja pacijenata** kroz **osam ciljnih dijagnoza** — abdominalne patologije i hitna stanja interne medicine, među njima apendicitis, pankreatitis, pneumoniju i infekciju mokraćnog sistema. Za svaki slučaj MIRA je počinjala s ograničenom slikom i korak po korak morala odlučiti što će dalje učiniti — istog oblika kao stvarna obrada, ali izvedena nad kartonom čiji je stvarni završetak već poznat.

Njena izvedba zatim je upoređena s ljekarima na istim slučajevima.

Šta zapravo znači „autonomni agent u sandboxiranom EHR-u"

Tri riječi ovdje nose mnogo značenja, pa ih vrijedi rastaviti.
Agent znači da sistem nije chatbot koji odgovara na prompt. Izvodi petlju: pogleda trenutno stanje, odabere radnju (naruči ovu pretragu, zatraži taj podatak iz anamneze), vidi rezultat, odabere sljedeću radnju — sve dok ne dođe do dijagnoze i plana. „Inteligencija" je jezični model; *agentnost* daje okolni softverski okvir koji tekst modela pretvara u dopuštene operacije nad EHR-om i rezultate vraća modelu.
Sandboxirani EHR znači simuliranu, odvojenu kopiju sistema medicinskih kartona, a ne živi bolnički sistem. MIRA može „naručiti" pretragu i dobiti rezultat koji je taj pacijent stvarno imao, jer je slučaj historijski i odgovor je već zabilježen. Ništa što učini ne dotiče stvarnog pacijenta ni stvarnu kliniku.
Autonomno znači da dovršava slučaj od početka do kraja bez čovjeka u petlji — *unutar sandboxa*. To **ne** znači da je puštena da bez nadzora vodi stvarne pacijente. To su vrlo različite tvrdnje, a testirana je samo prva.
Još jedna stvar o sandboxu, jer ga je lako pogrešno zamisliti: MIRA smije *naručiti* bilo koju pretragu, ali sistem može vratiti rezultat samo ako je ta pretraga **zaista bila izvedena** tom pacijentu u stvarnom životu. Zatraži li krvnu pretragu koju je pacijent dobio, vraćaju se

> stvarne vrijednosti; zatraži li snimku koju niko nije napravio, vraća se „N/A — could not be performed", nikada izmišljena brojka. Isto vrijedi za anamnezu: ako karton ne sadrži ono što MIRA pita, simulirani pacijent jednostavno kaže da ne zna. Dakle, MIRA ne može prizvati jednu odlučujuću pretragu koja nikada nije učinjena — može raditi samo s onime što je stvarna obrada slučaja zaista sadržala.
> Ta je razlika važna jer je riječ iz naslova — „autonomno" — upravo ona koju će se najlakše pogrešno pročitati kao drugo značenje.

Šta su pronašli

Na benchmarku je **MIRA nadmašila ljekare u dijagnostičkoj tačnosti** — ali naslov skriva tri različite brojke koje je lako pomiješati pa ih vrijedi razdvojiti.

Sama, na svih **574 slučaja**, MIRA je navela ispravnu dijagnozu u **88.9%** slučajeva — ocjenjivano prema otpusnoj dijagnozi stvarno zapisanoj za svakog pacijenta u MIMIC-IV. U direktnoj uporedbi s ljudima ljekari nisu obrađivali svih 574 slučaja; obrađivali su zajednički **podskup od 311 slučajeva**, uz iste kartone i alate kojima je raspolagala MIRA. Na tih istih 311 slučajeva MIRA je postigla **87.8%**, a upoređena je s dvije zasebno okupljene grupe ljekara. Prva je bila **seniorna grupa**: četiri certificirana specijalista sa 7 do 11 godina iskustva, koji su postigli **78.1%**. Druga je bila **grupa mješovitog iskustva**: uglavnom mlađi specijalizanti uz dva specijalista, bliža stvarnoj kadrovskoj slici hitnog prijma, koja je postigla **71.1%**. Isti slučajevi, isti alati — redoslijed je bio AI prvi, iskusni ljekari drugi, mlađa grupa treća. Oprezna formulacija samog rada jest da je MIRA bila „dosljedno jednaka ljekarima, a često ih i nadmašivala" kroz svih osam bolesti.

I odluke nakon dijagnoze uglavnom su se održale: bile su većinom usklađene sa smjernicama, sigurne u pogledu lijekova i primjerene za odluku o prijamu. Autori navode da nije bilo teških grešaka u propisivanju kroz pet sigurnosnih kategorija (interakcije lijekova, doziranje kod bubrežne funkcije, alergije, QT-rizik i propisivanje opioida), a recepti su bili ispravni u 467 od 468 slučajeva — uz napomenu da sistem „nije postigao 100% pouzdanost". Ako se uzme doslovno, to je upečatljiv rezultat: agent vodi cijeli slučaj i završava ispred ljekara u dijagnozi.

Ali *oblik* te pobjede važan je koliko i sama pobjeda. Nezavisni stručnjaci koji su čitali rad istaknuli su odakle zapravo dolazi prednost. Na stručnom panelu Science Media Centrea dr. Wei Xing (University of Sheffield) primijetio je da je naslovna brojka — AI nadmašuje ljekare u dijagnostičkoj tačnosti — **uglavnom potaknuta stanjima s jasnim rezultatima pretraga**, poput apendicitisa i pankreatitisa, gdje odlučujuća snimka ili laboratorijska vrijednost rješava pitanje. Kod pneumonije i infekcija mokraćnog sistema — dva vrlo česta razloga dolaska u hitnu službu — *i* AI *i* ljekari imali su najlošije rezultate, a razlika među njima bila je najmanja.

Postojala je i asimetrija u načinu na koji su dvije strane „igrale", i ona je vidljiva u brojkama samog rada. MIRA se znatno više oslanjala na laboratorij: koristila je oko **51%** laboratorijskih analita dostupnih u rutinskoj zdravstvenoj zaštiti, naspram otprilike **28%** kod certificiranih ljekara — medijan od sedam dodatnih krvnih parametara po slučaju. Autori pažljivo navode da je to *ispod* rutinske razine

u samom skupu podataka, a ne strategija „naruči sve", te da nije bilo sistemskog povećanja skupljeg presječnog snimanja. Ipak, više informacija samo po sebi može dati veću dijagnostičku tačnost — pa ovo nije sasvim uporedba jednako informiranih učesnika, što je Xing direktno istaknuo. I kliničar koji smije naručiti svaku jeftinu krvnu pretragu bez vaganja cijene, neugode ili kašnjenja izgledao bi preciznije na papiru.

Zašto „pobijedila ljekare" ne znači „bolji ljekar"

Pobjedu na benchmarku lako je previše protumačiti. Tri stvari stoje između „MIRA je postigla viši rezultat" i „MIRA je bolja".

Prvo, **prema čemu se ocjenjivala.** Profesorica Julie Jacko (University of Edinburgh) primijetila je da je nekoliko ključnih MIRINIH ishoda definirano u odnosu na ono što je zapisano u izvornom skupu podataka — što znači da je sistem nagrađen za **ponavljanje zabilježenog kliničkog ponašanja**, a ne nužno za dokazivanje optimalne zdravstvene zaštite. Historijski karton postaje ključ odgovora. To je razuman način izgradnje benchmarka, ali mjeri slaganje s onime što je učinjeno, a ne apsolutnu ispravnost. Pošteno prema radu, to najviše pogađa metrike *usklađenosti liječenja* — podudaraju li se MIRINI nalozi s kartonom? — dok se naslovna dijagnostička tačnost mjeri prema otpusnoj dijagnozi, što je bliže stvarnom ishodu nego odjeku dokumentacije.

Drugo, već spomenuta **asimetrija informacija**: mnogo više krvnih pretraga (oko 51% dostupnih analita prema 28%) znači i više dokaza. Dio razlike u tačnosti vjerovatno je *kupljen* dodatnim informacijama, a ne izveden boljim rasuđivanjem.

Treće, **gdje je sistem radio.** Ovo je bio sandbox na retrospektivnim slučajevima, kroz osam unaprijed odabranih dijagnoza i samo u tekstu. Stvarna klinička procjena, kako je rekao dr. Dominic Oliver (University of Oxford), ne zavisi samo o tome što pacijenti kažu nego i kako to kažu — uz fizikalni pregled, opaženo ponašanje i govor tijela, od čega tekstualni agent koji čita gotov karton ne vidi ništa. Pacijent čiji problem ne pripada jednoj od osam odabranih dijagnoza jednostavno je izvan onoga o čemu ova studija može govoriti.

Recenzenti su istaknuli i tišu brigu: **kontaminaciju podataka.** MIMIC-IV je javan i o njemu se mnogo piše, pa je jezični model treniran na otvorenom internetu možda već vidio radove, rasprave o slučajevima ili same podatke. Dr. Midhun Parakkal Unni (University of Sheffield) direktno je upozorio na to — ako su neki odgovori bili u podacima za treniranje, dio izvedbe je pamćenje, a ne kliničko rasuđivanje, i samo nezavisna replikacija može razlikovati jedno od drugoga. Autori taj problem ne umanjuju: pišu da se rezultati „mogu oprezno tumačiti kao moguća gornja granica" i da „mogu precijeniti generalizaciju na druge javne slučajeve" — rad sam postavlja strop vlastitom naslovu.

Ništa od toga MIRU ne čini manje zanimljivom. Samo zahtijeva odmjereno čitanje: na retrospektivnom benchmarku agent koji je mogao djelovati kroz cijeli karton nadmašio je ljekare kod dijagnoza s čistim testovima — dijelom naručujući više pretraga, dijelom slažući se sa zabilježenim kartonom. To je stvarna demonstracija sposobnosti, a ne presuda da mašine sada dijagnosticiraju bolje od ljekara.

Šta ovo *ne* dokazuje

- **Ne** pokazuje da MIRA radi na stvarnim pacijentima. Svaki slučaj bio je retrospektivan i simuliran; MIRA nikada nije vodila stvarnog pacijenta.
- **Ne** pokazuje da radi izvan osam dijagnoza. Stanja izvan unaprijed odabranog skupa — neuredna, nediferencirana većina medicine — nisu testirana.
- **Ne** pokazuje da je sigurna za upotrebu. Sami autori pišu da generalizacija, sigurnost i upravljanje još zahtijevaju prospektivne studije u stvarnom svijetu.
- **Ne** uspostavlja poštenu direktnu uporedbu s ljekarima. MIRA je koristila mnogo više krvnih pretraga (oko 51% dostupnih analita naspram 28%), a nekoliko ishoda liječenja djelomično je ocjenjivano prema historijskom kartonu umjesto prema nezavisnoj najboljoj zdravstvenoj zaštiti.
- **Ne** pokazuje da je rezultat oslobođen pamćenja. Javni podaci MIMIC-IV mogu se preklapati s podacima za treniranje modela, što bi nezavisna replikacija morala isključiti.
- **Ne** znači „AI zamjenjuje ljekare". To je podrška odlučivanju koja može *djelovati* kroz karton; stručni konsenzus je da će se stvarna upotreba odvijati u partnerstvu s kliničarima, koji zadržavaju ovlast i daju sve ono što tekstualni karton izostavlja.

Koliko su dokazi jaki?

Tvrdnju treba podijeliti na dva dijela jer su dokazi za njih vrlo različiti.

Kao demonstracija sposobnosti — da agent temeljen na jezičnom modelu može provesti cijeli klinički slučaj unutar EHR sandboxa, povezati anamnezu, pretrage, dijagnozu i naloge od početka do kraja — rad je zaista nov i prilično uvjerljiv. To je novi dio i on zaslužuje pozornost.

Kao tvrdnja o nadmoći — da je MIRA *bolja od ljekara* — dokazi su ograničeni i treba ih oprezno čitati. Vrijede na određenom retrospektivnom benchmarku, na osam dijagnoza, uz asimetriju informacija (više pretraga), ključ odgovora izveden iz historijskog kartona i stvarnu mogućnost kontaminacije podataka za treniranje. To je dovoljno da se kaže „agent je impresivno radio na ovom benchmarku". Nije dovoljno za „agent je bolji dijagnostičar od ljekara", a autori ni ne tvrde to drugo.

Najkorisniji stav nije ni „AI pobjeđuje ljekare" ni „samo igračka". On glasi: nova vrsta medicinske vještačke inteligencije — ona koja *djeluje* kroz karton umjesto da samo odgovara na pitanja — dobro se pokazala na zahtjevnom retrospektivnom benchmarku i sada se mora dokazati tamo gdje još nije isprobana: na stvarnim, nediferenciranim pacijentima, prospektivno i pod odgovarajućim upravljanjem.

Zašto je to važno

Posljednjih nekoliko godina zanimljivo pitanje u medicinskoj vještačkoj inteligenciji tiho se promijenilo. Nekad je bilo „može li model pogoditi dijagnozu?" — i modeli su sve češće odgovarali da mogu, na

sve čišćim testnim skupovima. MIRA označava pomak prema težem pitanju: „može li model *obaviti posao* — prikupiti, naručiti, protumačiti, odlučiti, djelovati — kroz cijeli tok rada?" To je korisnije i poštenije pitanje jer su upravo u djelovanju i teškoća i rizik.

Zato je okvir priče važan. Refleksni naslov — *AI nadmašuje ljekare* — upire u najmanje nov i najslabije potkrijepljen dio rezultata. Ono zaista novo manje je i važnije: agent koji se može kretati kroz cijeli karton. Ako ta sposobnost izdrži provjeru, promijenit će **tok rada** puno prije nego što promijeni ko njime upravlja. Ljekar nije zamijenjen; naručivanje, tumačenje i praćenje rezultata — tok rada — mjesto je gdje alat poput ovoga prvo ulazi.

To i teret dokazivanja pomiče u pravom smjeru. Pobjeda na skali retrospektivnih slučajeva razlog je da se provede prospektivno ispitivanje, a ne zamjena za njega. Autori kažu isto. Pošteno čitanje MIRE poziv je na sljedeću studiju — prema standardu koji polje već poznaje: mjeriti na pacijentima, a ne na benchmarkovima.

Kratak sažetak

MIRA je autonomni AI agent koji djeluje unutar sandboxiranog elektroničkog zdravstvenog kartona: može uzeti anamnezu, naručiti i protumačiti laboratorij, snimanje i mikrobiologiju, doći do dijagnoze i napisati plan liječenja. Testirana na 574 retrospektivna slučaja iz javne baze MIMIC-IV, kroz osam unaprijed odabranih dijagnoza, nadmašila je ljekare u dijagnostičkoj tačnosti (88.9% ukupno; 87.8% naspram 78.1% u direktnoj uporedbi) i donosila odluke koje su uglavnom bile usklađene sa smjernicama i sigurne u pogledu lijekova. Ali evaluacija je bila simulacija na starim kartonima i samo u tekstu; velik dio prednosti došao je kod stanja s jasnim rezultatima pretraga; MIRA je koristila mnogo više krvnih pretraga od ljekara (oko 51% dostupnih analita naspram 28%); nekoliko ishoda liječenja ocjenjivano je prema onome što je izvorni karton zabilježio; a javni skup podataka stvara stvaran rizik kontaminacije podataka za treniranje — što sami autori nazivaju mogućom gornjom granicom svojih brojki. Stvaran napredak je agent koji djeluje kroz cijeli tok rada umjesto da odgovara na izolirana pitanja. Autori izričito navode da generalizacija, sigurnost i upravljanje još zahtijevaju prospektivne studije u stvarnom svijetu — što je i ispravno čitanje: impresivna demonstracija sposobnosti, a ne dokaz da AI dijagnosticira bolje od ljekara niti sistem spreman za stvarnu ambulantu.

Provjera bez uljepšavanja

Šta rad pokazuje: Agent temeljen na jezičnom modelu (MIRA) može provesti cijeli klinički slučaj od početka do kraja unutar sandboxiranog EHR-a — anamneza, pretrage, dijagnoza, nalozi — i na retrospektivnom benchmarku od 574 slučaja kroz osam dijagnoza nadmašiti ljekare u dijagnostičkoj tačnosti (88.9% ukupno; 87.8% naspram 78.1% u uporedbi s certificiranim ljekarima), bez teških grešaka u lijekovima.

Šta je vjerovatno, ali nije dokazano: Da se ta sposobnost pretvara u korist za stvarne, nediferencirane pacijente; da prednost u tačnosti odražava bolje rasuđivanje, a ne više naručenih pretraga, slaganje

sa zabilježenim kartonom ili zapamćene javne podatke.

Šta ne pokazuje: Da MIRA radi izvan osam dijagnoza; da je sigurna za upotrebu; da pobjeđuje ljekare u poštenoj uporedbi s jednakim informacijama; da zamjenjuje kliničare; da su rezultati oslobođeni kontaminacije podataka za treniranje.

Glavna ograničenja: Retrospektivna, simulirana evaluacija samo u tekstu; osam unaprijed odabranih dijagnoza; asimetrija informacija (mnogo više krvnih pretraga — oko 51% dostupnih analita naspram 28%, i to jeftinih krvnih pretraga, ne snimanja); ishodi liječenja dijelom ocjenjivani prema historijskom kartonu; javni benchmark MIMIC-IV koji je jezični model možda vidio tokom treniranja — što autori označavaju kao moguću gornju granicu izvedbe.

Koliko povjerenja treba imati opći čitalac? Visoko da je MIRA stvarna i nova demonstracija sposobnosti — agent koji *djeluje* kroz karton, a ne samo chatbot. Nisko da je *bolji dijagnostičar od ljekara*: ta je tvrdnja ograničena na jedan retrospektivni benchmark i zamagljena većim brojem pretraga, ocjenjivanjem prema kartonu i mogućom kontaminacijom. Zaključak samih autora je pravi — prije nego što ovo znači išta za briga o pacijentima, potrebna je prospektivna studija u stvarnom svijetu.

Izvori

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