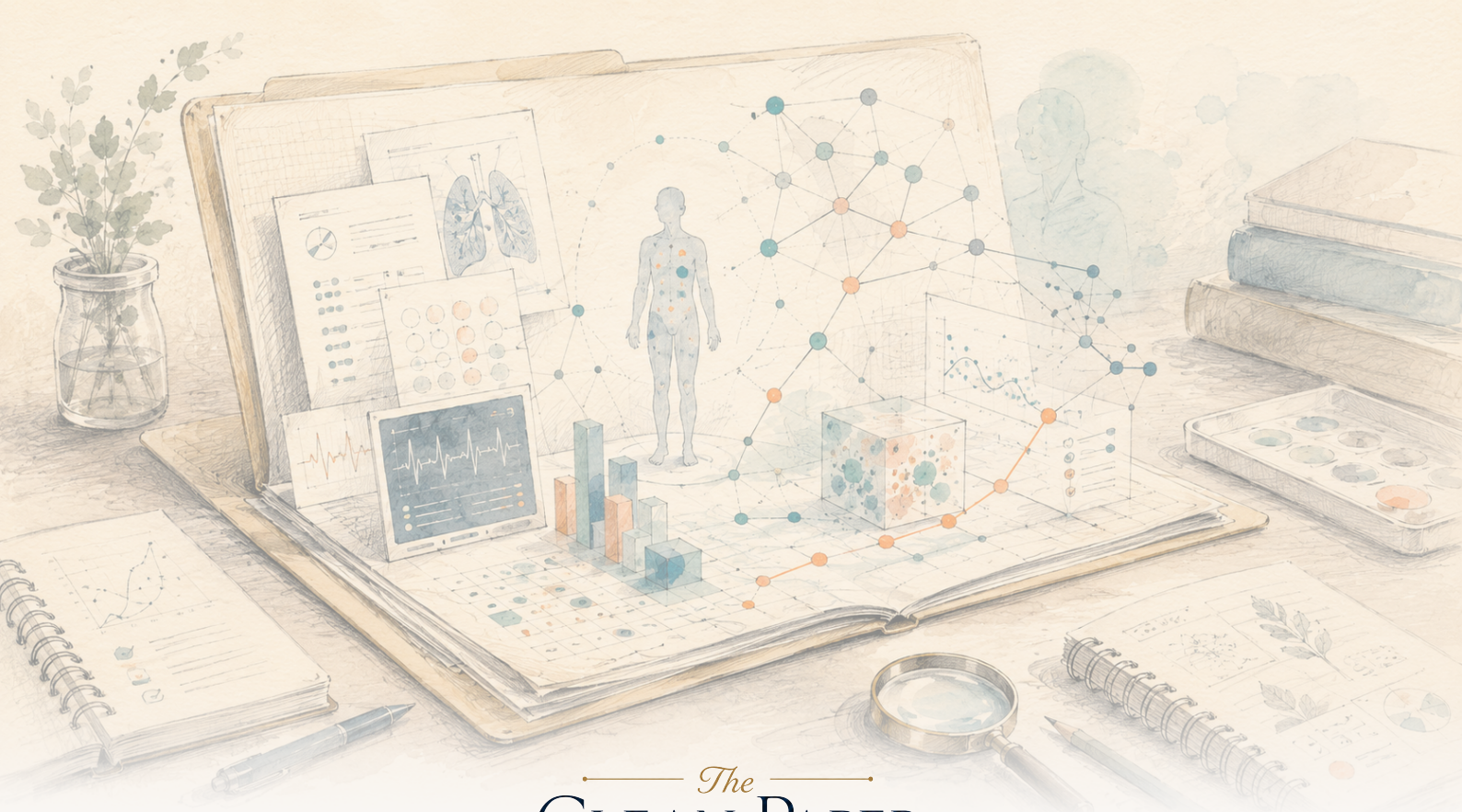


The Clean Paper

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

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

MIRA je nova vrsta medicinske veštačke inteligencije: umesto odgovora na jedno pitanje, obrađuje celi 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 unapred odabranih dijagnoza, autori navode da je nadmašila lečnike u dijagnostičkoj tačnosti i donosila odluke koje su uglavnom bile usklađene sa smernicama i sigurne u pogledu lekova. Svaka ograda u toj rečenici važna je: radila je u sandboxu na starim kartonima i samo u tekstu; velik deo prednosti došao je kod stanja s jasnim rezultatima pretraga; naručivala je otprilike dvostruko više krvnih pretraga od lekara; a nekoliko ishoda ocenjivano je prema onome što je originalni karton zabeležio. Stvarni napredak je agent koji deluje kroz celi tok rada — ne dokaz da mašina sada dijagnosticira bolje od lekara. Autori to i sami kažu: generalizacija, bezbednost i upravljanje još zahtevaju prospektivne studije u stvarnom svetu.

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

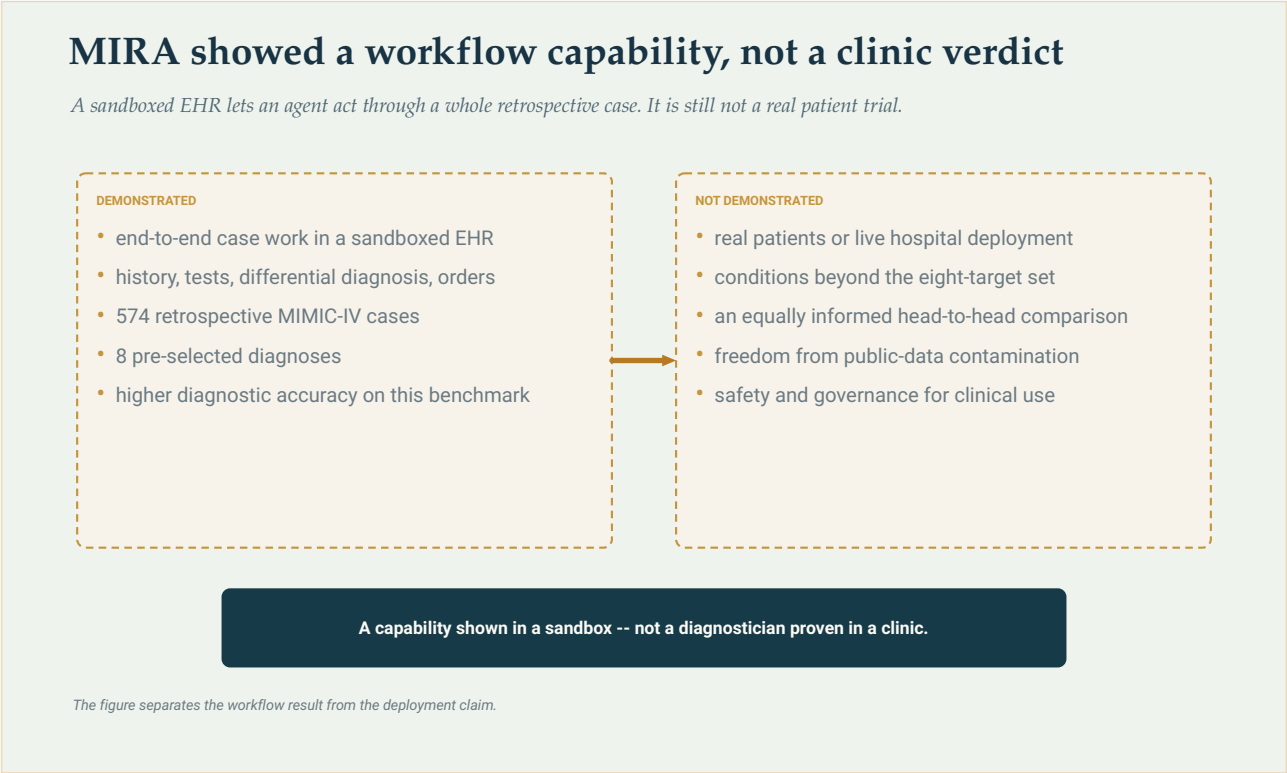
Odgovoriti na pitanje nije isto što i voditi celi slučaj

Većina priča o medicinskoj veštačkoj inteligenciji govori o modelu koji *odgovara*. Date mu kliničku vinjetu ili ispitno pitanje, a on vrati dijagnozu ili odlomak saveta 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. Umesto odgovora na jedno pitanje, obrađuje celi 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 slede — recepte, prijam, upućivanje na operaciju. To čini izvodeći radnje *unutar* elektronskog zdravstvenog kartona, poput kliničara, umesto da samo proizvodi slobodan tekst koji bi čovek morao prepisati.

To je stvaran korak: od sistema koji savetuje prema sistemu koji deluje kroz celi tok rada. I upravo je to vrsta koraka koja se u medijskom izveštavanju lako spljošti u „AI nadmašuje lečnike". Zato vredi precizno reći što je testirano i gde.

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



MIRA je pokazala sposobnost vođenja celog toka rada unutar sandboxiranog EHR-a. Nije pokazala kliničku upotrebu u stvarnom svetu, široku dijagnostičku pokrivenost ni poštenu uporedbu s lečnicima 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: deo koji razgovara i deluje koristi **GPT-4o**, a zaseban korak planiranja koristi OpenAI-jev model za rasuđivanje **o1**. smešteni su u prilagođeni okvir usklađen sa standardima — izgrađen na HL7 FHIR-u, podatkovnom standardu kojim stvarne bolnice razmenjuju 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 lečenja — propisati lekove, zakazati kirurški zahvat, planirati prijam. Vredi odmah istaknuti jednu pojedinost: automatizirani „suci" koji su ocenjivali 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 lečenih 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 lečnicima na istim slučajevima.

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

Tri reči ovde nose mnogo značenja, pa ih vredi 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 sledeću radnju — sve dok ne dođe do dijagnoze i plana. „Inteligencija" je jezički model; *agentnost* daje okolni softverski okvir koji tekst modela pretvara u dozvoljene operacije nad EHR-om i rezultate vraća modelu.
Sandboxirani EHR znači simuliranu, odvojenu kopiju sistema medicinskih kartona, a ne aktivni bolnički sistem. MIRA može „naručiti" pretragu i dobiti rezultat koji je taj pacijent stvarno imao, jer je slučaj istorijski i odgovor je već zabeležen. Ništa što učini ne dotiče stvarnog pacijenta ni stvarnu kliniku.
Autonomno znači da dovršava slučaj od početka do kraja bez čoveka 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 sme *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 vrednosti; zatraži li snimku koju niko nije napravio, vraća se „N/A — could not be performed", nikada izmišljena brojka. Isto vredi 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 sadržavala.
> Ta je razlika važna jer je reč iz naslova — „autonomno" — upravo ona koju će se najlakše pogrešno pročitati kao drugo značenje.

Šta su pronašli

Na benčmarku je **MIRA nadmašila lečnike u dijagnostičkoj tačnosti** — ali naslov skriva tri različite brojke koje je lako pomešati pa ih vredi razdvojiti.

Sama, na svih **574 slučaja**, MIRA je navela ispravnu dijagnozu u **88.9%** slučajeva — ocenjivano prema otpusnoj dijagnozi stvarno zapisanoj za svakog pacijenta u MIMIC-IV. U direktnoj poređenju s ljudima lekari 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 dve zasebno okupljene grupe lekara. Prva je bila **seniorna grupa**: četiri certificirana specijalista sa 7 do 11 godina iskustva, koji su postigli **78.1%**. Druga je bila **grupa mešovitog iskustva**: uglavnom mlađi specijalizanti uz dva specijalista, bliža stvarnoj kadrovskoj slici hitnog prijma, koja je postigla **71.1%**. Isti slučajevi, isti alati — redosled je bio AI prvi, iskusni lekari drugi, mlađa grupa treća. Oprezna formulacija samog rada je da je MIRA bila „dosledno jednaka lečnicima, a često ih i nadmašivala" kroz svih osam bolesti.

I odluke nakon dijagnoze uglavnom su se održale: bile su većinom usklađene sa smernicama, sigurne u pogledu lekova i primerene za odluku o prijamu. Autori navode da nije bilo teških grešaka u propisivanju kroz pet bezbednosnih kategorija (interakcije lekova, 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 celi slučaj i završava ispred lekara u dijagnozi.

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

Postojala je i asimetrija u načinu na koji su dve 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 nezi, naspram otprilike **28%** kod certificiranih lekara — medijan od sedam dodatnih krvnih parametara po slučaju. Autori pažljivo navode da je to *ispod* rutinskog nivoa u samom skupu podataka,

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

Zašto „pobedila lečnike" ne znači „bolji lekar"

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

Prvo, **prema čemu se ocenjivala.** Profesorica Julie Jacko (University of Edinburgh) primetila je da je nekoliko ključnih MIRINIH ishoda definisano u odnosu na ono što je zapisano u originalnom skupu podataka — što znači da je sistem nagrađen za **ponavljanje zabeleženog kliničkog ponašanja**, a ne nužno za dokazivanje optimalne nezi. Istorijski karton postaje ključ odgovora. To je razuman način izgradnje benčmarka, ali meri slaganje s onime što je učinjeno, a ne apsolutnu ispravnost. Pošteno prema radu, to najviše pogađa metrike *usklađenosti lečenja* — podudaraju li se MIRINI nalozi s kartonom? — dok se naslovna dijagnostička tačnost meri 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. Deo razlike u tačnosti verovatno je *kupljen* dodatnim informacijama, a ne izveden boljim rasuđivanjem.

Treće, **gde je sistem radio.** Ovo je bio sandbox na retrospektivnim slučajevima, kroz osam unapred odabranih dijagnoza i samo u tekstu. Stvarna klinička procena, kako je rekao dr. Dominic Oliver (University of Oxford), ne zavisi samo o tome što pacijenti kažu nego i kako to kažu — uz fizički pregled, uočeno ponašanje i govor tela, 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 jezički 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, deo 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 preceniti generalizaciju na druge javne slučajeve" — rad sam postavlja strop sopstvenom naslovu.

Ništa od toga MIRU ne čini manje zanimljivom. Samo zahteva odmereno čitanje: na retrospektivnom benčmarku agent koji je mogao delovati kroz celi karton nadmašio je lečnike kod dijagnoza s čistim testovima — delom naručujući više pretraga, delom slažući se sa zabeleženim kartonom. To je stvarna demonstracija sposobnosti, a ne presuda da mašine sada dijagnostikuju bolje od lekara.

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 unapred odabranog skupa — neuredna, nediferencirana većina medicine — nisu testirana.
- **Ne** pokazuje da je sigurna za upotrebu. Sami autori pišu da generalizacija, bezbednost i upravljanje još zahtevaju prospektivne studije u stvarnom svetu.
- **Ne** uspostavlja poštenu direktnu uporedbu s lečnicima. MIRA je koristila mnogo više krvnih pretraga (oko 51% dostupnih analita naspram 28%), a nekoliko ishoda lečenja delimično je ocenjivano prema istoresnom kartonu umesto prema nezavisnoj najboljoj nezi.
- **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 zamenjuje lečnike". To je podrška odlučivanju koja može *delovati* 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 podeliti na dva dela jer su dokazi za njih vrlo različiti.

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

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

Najkorisniji stav nije ni „AI pobeđuje lečnike" ni „samo igračka". On glasi: nova vrsta medicinske veštačke inteligencije — ona koja *deluje* kroz karton umesto da samo odgovara na pitanja — dobro se pokazala na zahtevnom retrospektivnom benčmarku i sada se mora dokazati tamo gde još nije isprobana: na stvarnim, nediferenciranim pacijentima, prospektivno i pod odgovarajućim upravljanjem.

Zašto je to važno

poslednjih nekoliko godina zanimljivo pitanje u medicinskoj veštačkoj inteligenciji tiho se promenilo. 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, delovati — kroz celi tok rada?" To je korisnije i poštenije pitanje jer su upravo u delovanju i teškoća i rizik.

Zato je okvir priče važan. Refleksni naslov — *AI nadmašuje lečnike* — upire u najmanje nov i najslabije potkrepljen deo rezultata. Ono zaista novo manje je i važnije: agent koji se može kretati kroz celi karton. Ako ta sposobnost izdrži proveru, promeniće **tok rada** puno pre nego što promeni ko njime upravlja. Lekar nije zamenjen; naručivanje, tumačenje i praćenje rezultata — tok rada — mesto je gde alat poput ovoga prvo ulazi.

To i teret dokazivanja pomiče u pravom smeru. Pobeda na skali retrospektivnih slučajeva razlog je da se provede prospektivno ispitivanje, a ne zamena za njega. Autori kažu isto. Pošteno čitanje MIRE poziv je na sledeću studiju — prema standardu koji polje već poznaje: meriti na pacijentima, a ne na benčmarkovima.

Kratak rezime

MIRA je autonomni AI agent koji deluje unutar sandboxiranog elektronskog zdravstvenog kartona: može uzeti anamnezu, naručiti i protumačiti laboratorij, snimanje i mikrobiologiju, doći do dijagnoze i napisati plan lečenja. Testirana na 574 retrospektivna slučaja iz javne baze MIMIC-IV, kroz osam unapred odabranih dijagnoza, nadmašila je lečnike u dijagnostičkoj tačnosti (88.9% ukupno; 87.8% naspram 78.1% u direktnom poređenju) i donosila odluke koje su uglavnom bile usklađene sa smernicama i sigurne u pogledu lekova. Ali evaluacija je bila simulacija na starim kartonima i samo u tekstu; velik deo prednosti došao je kod stanja s jasnim rezultatima pretraga; MIRA je koristila mnogo više krvnih pretraga od lekara (oko 51% dostupnih analita naspram 28%); nekoliko ishoda lečenja ocenjivano je prema onome što je originalni karton zabeležio; 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 deluje kroz celi tok rada umesto da odgovara na izolovana pitanja. Autori izričito navode da generalizacija, bezbednost i upravljanje još zahtevaju prospektivne studije u stvarnom svetu — što je i ispravno čitanje: impresivna demonstracija sposobnosti, a ne dokaz da AI dijagnosticira bolje od lekara niti sistem spreman za stvarnu ambulantu.

Provera bez ulepšavanja

Šta rad pokazuje: Agent zasnovan na jezičnom modelu (MIRA) može provesti celi klinički slučaj od početka do kraja unutar sandboxiranog EHR-a — anamneza, pretrage, dijagnoza, nalozi — i na retrospektivnom benčmarku od 574 slučaja kroz osam dijagnoza nadmašiti lečnike u dijagnostičkoj tačnosti (88.9% ukupno; 87.8% naspram 78.1% u poređenju s certificiranim lečnicima), bez teških grešaka u lekovima.

Šta je verovatno, 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 zabeleženim kartonom ili zapamćene javne podatke.

Šta ne pokazuje: Da MIRA radi izvan osam dijagnoza; da je sigurna za upotrebu; da pobeđuje lečnike u poštenoj poređenju s jednakim informacijama; da zamenjuje kliničare; da su rezultati oslobođeni kontaminacije podataka za treniranje.

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

Koliko poverenja treba imati opšti čitalac? Visoko da je MIRA stvarna i nova demonstracija sposobnosti — agent koji *deluje* kroz karton, a ne samo chatbot. Nisko da je *bolji dijagnostičar od lekara*: ta je tvrdnja ograničena na jedan retrospektivni benčmark i zamagljena većim brojem pretraga, ocenjivanjem prema kartonu i mogućom kontaminacijom. Zaključak samih autora je pravi — pre nego što ovo znači išta za nega o pacijentima, potrebna je prospektivna studija u stvarnom svetu.

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