The Clean Paper

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

Agent umetne inteligence je samostojno obdelal celotne primere bolnikov — v simulatorju, na preteklih zapisih

MIRA je nova vrsta medicinske umetne inteligence: namesto odgovora na eno vprašanje v simuliranem bolnišničnem zapisu obravnava celoten primer — zbere anamnezo, naroči in prebere preiskave, postavi diagnozo in napiše naročila. Na 574 retrospektivnih primerih iz javne baze, pri osmih vnaprej izbranih diagnozah, avtorji poročajo o višji diagnostični natančnosti kot pri zdravnikih ter večinoma zmerno skladnih in varnih odločitvah glede zdravil. Vsak kvalifikator v tej povedi šteje: sistem je deloval v peskovniku na preteklih zapisih, samo z besedilom; velik del prednosti je izhajal iz bolezni z jasnimi rezultati preiskav; naročil je približno dvakrat več krvnih testov kot zdravniki; več izidov pa je bilo ocenjenih glede na to, kar je bilo zapisano v prvotni dokumentaciji. Resnični napredek je agent, ki deluje čez celoten potek dela — ne dokaz, da stroj zdaj diagnosticira bolje od zdravnika. Avtorji to izrecno priznavajo: generalizacijo, varnost in upravljanje je treba še preizkusiti prospektivno in v resničnem 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

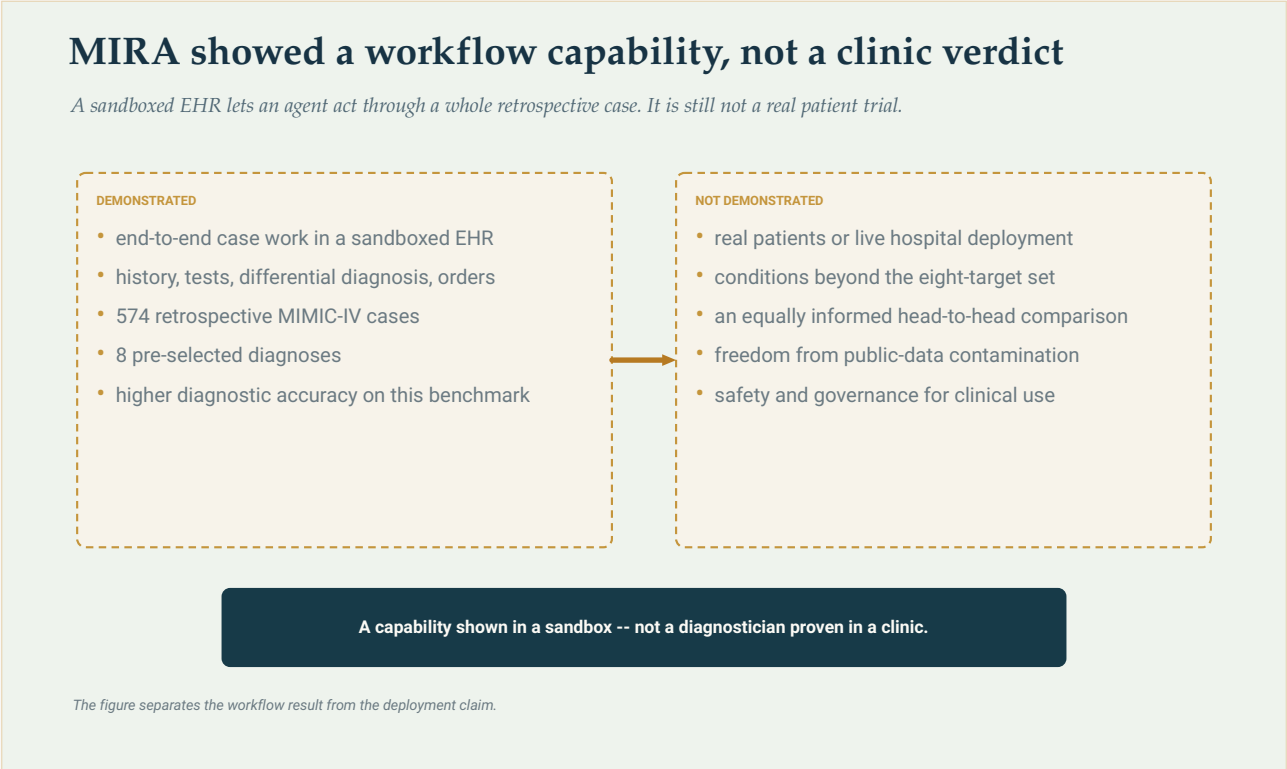
Odgovoriti na vprašanje ni isto kot voditi celoten primer

Večina zgodb o medicinski umetni inteligenci govori o modelu, ki *odgovarja*. Damo mu klinično vinjeto ali izpitno vprašanje, vrne diagnozo ali odstavek nasveta in doseže visoko oceno. Uporabno, vendar ozko: pameten svetovalec, ki nikoli ne poseže v kartoteko.

MIRA je zasnovana za nekaj drugega in prav ta razlika je dejanska novica. Namesto enega odgovora obravnava celoten primer: prebere zdravstveni zapis, se odloči, katere podatke iz anamneze še potrebuje, naroči laboratorijske, slikovne in mikrobiološke preiskave, prebere rezultate, zoži diferencialno diagnozo in nato napiše naslednja naročila — zdravila, sprejem v bolnišnico, napotitev na operacijo. To počne z dejanji *znotraj* elektronskega zdravstvenega zapisa, podobno kot klinik, ne zgolj z ustvarjanjem prostega besedila, ki bi ga človek prepisal.

To je resničen korak: od sistema, ki svetuje, do sistema, ki deluje čez celoten potek dela. In prav tak korak se v poročanju hitro splošči v »AI prekaša zdravnike«. Zato je vredno natančno povedati, kaj je bilo preizkušeno in kje.

Enostavčna različica: MIRA je samostojno obravnavala celotne primere in na tem benchmarku dosegla višjo diagnostično natančnost kot zdravniki — vendar v peskovniku, na retrospektivnih zapisih, pri osmih vnaprej izbranih diagnozah, samo z besedilom, pri čemer je velik del prednosti dobila pri boleznih z najjasnejšimi rezultati preiskav. Napredek je resničen. Rezultat na lestvici ni klinika.



MIRA je v peskovniškem elektronskem zdravstvenem zapisu pokazala zmožnost vodenja celotnega poteka dela. Ni pokazala resnične klinične uporabe, širokega diagnostičnega obsega ali poštene neposredne primerjave z zdravniki, ki bi imeli povsem enake informacije.

Kaj so naredili avtorji

MIRA — Medical Intelligence for Reasoning and Action — je avtonomen agent, zgrajen na modelih OpenAI: del, ki se pogovarja in izvaja dejanja, uporablja **GPT-4o**, ločen načrtovalni korak pa model sklepanja **o1**. Modela delujeta v prilagojenem, standardom skladnem ogrodju, zgrajenem na HL7 FHIR, podatkovnem standardu, ki ga resnične bolnišnice uporabljajo za izmenjavo zapisov, in z več kot osemdeset tisoč možnimi kliničnimi dejanji. V tem peskovniku lahko MIRA pridobi anamnezo, naroči in interpretira laboratorijske, slikovne in mikrobiološke preiskave, ustvari diferencialno diagnozo in pripravi načrt zdravljenja — predpiše zdravila, načrtuje kirurški poseg ali sprejem. Ena podrobnost je vredna opozorila že na začetku: avtomatizirani »sodniki«, ki so ocenjevali odgovore MIRE, so bili sami GPT-4o — ista družina modelov, ki je bila preizkušana — kar so avtorji po pomisleku recenzenta dopolnili s pregledom zdravnika s specialističnim certifikatom.

Testni primeri so prišli iz **MIMIC-IV**, velike javno dostopne, deidentificirane baze elektronskih zdravstvenih zapisov. Iz približno 300.000 bolnikov, zdravljenih v Beth Israel Deaconess Medical Center med letoma 2008 in 2019, so avtorji izbrali **574 primerov** z **osmimi ciljnimi diagnozami** — abdominalnimi in internističnimi urgentnimi stanji, med drugim apendicitisom, pankreatitisom, pljučnico in okužbo sečil. Pri vsakem primeru je MIRA začela z omejeno sliko in se morala korak za korakom odločati, kaj storiti naprej — struktura je podobna resnični obravnavi, le da poteka nad zapisom, katerega dejanski konec je že znan.

Uspešnost so nato primerjali z zdravniki na istih primerih.

Kaj v resnici pomeni »avtonomen agent v peskovniškem EHR«

Tri besede tukaj opravljajo veliko dela, zato jih je vredno razpakirati.

Agent pomeni, da sistem ni klepetalnik, ki odgovori na prompt. Izvaja zanko: pogleda trenutno stanje, izbere dejanje (naroči test, vpraša za podatek iz anamneze), vidi rezultat, izbere naslednje dejanje — dokler ne pride do diagnoze in načrta. »Inteligenca« je jezikovni model; *agencija* je ogrodje okoli njega, ki besedilo modela spremeni v dovoljena dejanja v EHR in rezultate vrne nazaj modelu.

Peskovniški EHR pomeni simulirano, od resnične bolnišnice ločeno kopijo sistema zdravstvenih zapisov, ne živega bolnišničnega sistema. MIRA lahko »naroči« preiskavo in dobi rezultat, ki ga je ta bolnik dejansko imel, ker je primer zgodovinski in je odgovor že v dokumentaciji. Nič, kar naredi, ne vpliva na pravega bolnika ali pravo ambulanto.

Avtonomen pomeni, da primer dokonča od začetka do konca brez človeka v zanki — *znotraj peskovnika*. Ne pomeni, da so jo spustili v kliniko, da bi brez nadzora obravnavala bolnike. To sta zelo različni trditvi in preizkušena je bila le prva.

Še nekaj o peskovniku, ker si ga je zlahka predstavljati napačno: MIRA lahko *naroči* katerikoli test, sistem pa ji rezultat vrne le, če je bil ta test pri bolniku **v resnici opravljen**. Če zahteva krvno preiskavo, ki jo je bolnik imel, dobi prave vrednosti; če zahteva slikanje, ki ga nihče

> ni naročil, dobi »N/A — ni bilo mogoče izvesti«, nikoli izmišljene številke. Enako velja za anamnezo: če zapisa ni, simulirani bolnik preprosto pove, da tega ne ve. MIRA torej ne more priklicati odločilne preiskave, ki v resnici ni bila narejena — dela lahko samo s tem, kar je resnična obravnava dejansko vsebovala.
> Razlika je pomembna, ker je naslovna beseda »avtonomen« prav tista, ki jo je najlažje napačno razumeti v drugem pomenu.

Kaj so ugotovili

Na benchmarku je **MIRA dosegla višjo diagnostično natančnost kot zdravniki** — vendar naslov skriva tri različne številke, ki jih je zlahka zamegliti, zato jih ločimo.

Samostojno je na vseh **574 primerih** pravilno diagnozo navedla v **88,9 %** primerov — ocenjeno glede na odpustno diagnozo, zapisano za bolnika v MIMIC-IV. V neposredni primerjavi z ljudmi zdravniki niso obravnavali vseh 574 primerov; obdelali so skupni podnabor **311 primerov** z istimi zapisi in orodji, kot jih je imela MIRA. Na teh istih 311 primerih je MIRA dosegla **87,8 %**, primerjali pa so jo z dvema ločeno rekrutiranima skupinama. Prva je bila **starejša skupina** štirih specialistično certificiranih zdravnikov s 7–11 leti izkušenj, ki je dosegla **78,1 %**. Druga je bila **mešana skupina po senioriteti**, večinoma mlajši specializanti in dva specialista, bliže tipični kadrovski sestavi urgentnega oddelka; dosegla je **71,1 %**. Isti primeri, ista orodja — vrstni red pa AI, izkušeni zdravniki, nato mlajša skupina. Previdna formulacija članka je, da je bila MIRA pri vseh osmih boleznih »dosledno enakovredna in pogosto boljša« od zdravnikov.

Dobro so se odrezale tudi odločitve po diagnozi: večinoma so bile skladne s smernicami, varne glede zdravil in primerne glede sprejema. Avtorji poročajo, da v petih varnostnih kategorijah — interakcije zdravil, odmerjanje pri okvarjenih ledvicah, alergije, tveganje QT in predpisovanje opioidov — ni bilo napak visoke resnosti, zdravila pa so bila pravilna v 467 od 468 primerov. Hkrati opozarjajo, da sistem »ni dosegel 100-odstotne zanesljivosti«. Če številke vzamemo dobesedno, je to osupljiv rezultat: agent vodi cel primer in pri diagnozi prehiti zdravnike.

Toda *oblika* te prednosti je enako pomembna kot njen obstoj. Neodvisni strokovnjaki, ki so brali članek, so pokazali, od kod je velik del prednosti prišel. Na strokovnem panelu Science Media Centre je dr. Wei Xing z Univerze v Sheffieldu opozoril, da je naslovna prednost AI pred zdravniki pri diagnostični natančnosti **večinoma izhajala iz bolezni z jasnimi rezultati preiskav**, kot sta apendicitis in pankreatitis, kjer vprašanje pogosto razreši odločilen slikovni ali laboratorijski rezultat. Pri pljučnici in okužbi sečil — dveh najpogostejših razlogih za obisk urgence — so se najslabše odrezali *tako* AI kot zdravniki, razlika med njimi pa je bila najmanjša.

Obstajala je tudi asimetrija pri količini informacij. MIRA se je bolj opirala na laboratorij: uporabila je približno **51 %** laboratorijskih analitov, ki so bili na voljo v rutinski oskrbi, specialistično certificirani zdravniki pa približno **28 %** — mediana je pomenila sedem dodatnih krvnih parametrov na primer. Avtorji previdno poudarijo, da je to še vedno *manj* od rutinske uporabe v samem podatkovnem naboru in da sistem ni sistematično naročal več dražjih presečnih slikovnih preiskav. Kljub temu več infor-

macij samo po sebi lahko poveča diagnostično natančnost — zato to ni povsem enaka tekma med stranema z enako količino podatkov. Kliniku, ki bi lahko brez omejitev naročal vsak poceni krvni test, ne da bi tehtal strošek, nelagodje ali zamudo, bi se prav tako izboljšala uspešnost na papirju.

Zakaj »premagala zdravnike« ne pomeni »boljši zdravnik«

Zmago na benchmarku je lahko hitro preceniti. Med »MIRA je dosegla višji rezultat« in »MIRA je boljša« stojijo vsaj tri stvari.

Prvič, **proti čemu je bila ocenjena**. Profesorica Julie Jacko z Univerze v Edinburgu je opozorila, da je več ključnih izidov MIRE opredeljenih glede na dokumentacijo v osnovnem podatkovnem naboru — sistem je torej nagrajen za **ponavljanje zabeleženega kliničnega vedenja**, ne nujno za dokaz optimalne oskrbe. Zgodovinska kartoteka postane ključ z odgovori. To je razumen način za benchmark, vendar meri skladnost s tem, kar je bilo narejeno, ne absolutne pravilnosti. Pošteno do članka: ta pomislek najbolj zadeva metrike *skladnosti zdravljenja* — ali so naročila MIRE ustrezala kartoteki — medtem ko je naslovna diagnostična natančnost merjena glede na odpustno diagnozo, kar je bliže resničnemu izidu kot zgolj odmev dokumentacije.

Drugič, **asimetrija informacij**: precej več krvnih preiskav, približno 51 % razpoložljivih analitov proti 28 %, pomeni več dokazov. Del razlike v natančnosti je verjetno *kupljen* z dodatnimi informacijami, ne zgolj izpeljan z boljšim sklepanjem.

Tretjič, **okolje**. To je bil peskovnik na retrospektivnih primerih, pri osmih vnaprej izbranih diagnozah in samo z besedilom. Resnična klinična ocena, kot je opozoril dr. Dominic Oliver z Univerze v Oxfordu, ni odvisna le od tega, kaj bolniki povedo, temveč tudi *kako* to povedo — poleg telesnega pregleda, opazovanega vedenja in govorice telesa, česar besedilni agent, ki bere končan zapis, sploh ne vidi. Bolnik, katerega težava ne sodi med osem izbranih diagnoz, pa je preprosto zunaj tega, kar lahko študija pove.

Recenzenti so opozorili tudi na tišjo skrb: **kontaminacijo učnih podatkov**. MIMIC-IV je javen in pogosto opisan, zato je lahko jezikovni model, izurjen na odprtem internetu, že videl članke, razprave o primerih ali same podatke. Dr. Midhun Parakkal Unni z Univerze v Sheffieldu je to izrecno izpostavil — če so bili nekateri odgovori v učnih podatkih, je del uspeha spomin, ne klinično sklepanje, in le neodvisna ponovitev lahko to loči. Avtorji tega ne odpravijo z zamahom roke: zapišejo, da je njihove rezultate mogoče »previdno razumeti kot možno zgornjo mejo« in da »lahko precenjujejo generalizacijo na druge javne primere« — članek torej sam postavlja strop nad lastni naslovni rezultat.

Nič od tega MIRE ne naredi manj zanimive. Pravilno branje je samo bolj umerjeno: na retrospektivnem benchmarku je agent, ki je lahko deloval čez celoten zapis, presegel zdravnike predvsem pri diagnozah z jasnimi preiskavami — deloma z naročanjem več testov, deloma s skladnostjo z zgodovinsko kartoteko. To je resničen prikaz zmogljivosti, ne razsodba, da stroji zdaj diagnosticirajo bolje od zdravnikov.

Česa to ne dokazuje

- **Ne** kaže, da MIRA deluje pri resničnih bolnikih. Vsi primeri so bili retrospektivni in simulirani; nobenega bolnika ni obravnavala v živo.
- **Ne** kaže delovanja zunaj osmih diagnoz. Bolezni zunaj vnaprej izbranega nabora — neurejena in nediferencirana večina medicine — niso bile preizkušene.
- **Ne** kaže, da je varna za uporabo. Avtorji sami pišejo, da generalizacija, varnost in upravljanje potrebujejo prospektivne študije v resničnem svetu.
- **Ne** vzpostavlja povsem poštene neposredne primerjave z zdravniki. Uporabila je precej več krvnih testov (okoli 51 % razpoložljivih analitov proti 28 %), več izidov zdravljenja pa je bilo deloma ocenjenih glede na zgodovinsko kartoteko in ne glede na absolutno najboljšo oskrbo.
- **Ne** kaže, da rezultat ni posledica pomnjenja. Javni MIMIC-IV se lahko prekriva z učnimi podatki modela, kar bi morala izključiti neodvisna ponovitev.
- **Ne** pomeni »AI nadomesti zdravnike«. Gre za podporo odločanju, ki lahko *deluje* v zapisu; soglasje recenzentov je, da bo resnična uporaba partnerstvo s kliniki, ki ohranijo pristojnost in dodajo vse, česar besedilni zapis ne vsebuje.

Kako močni so dokazi?

Trditev je treba razdeliti na dva dela, ker je moč dokazov zelo različna.

Kot prikaz zmogljivosti — da lahko agent jezikovnega modela v peskovniškem EHR vodi celoten klinični primer ter poveže anamnezo, preiskave, diagnozo in naročila od začetka do konca — je delo res novo in razmeroma prepričljivo. To je dejansko nov del in prav njega je vredno spremljati.

Kot trditev o superiornosti — da je MIRA *boljša od zdravnikov* — so dokazi omejeni in zahtevajo previdnost. Veljajo na določenem retrospektivnem benchmarku, pri osmih diagnozah, z informacijsko asimetrijo (več testov), ključem odgovorov, ki deloma izhaja iz zgodovinske kartoteke, in z resnično možnostjo kontaminacije učnih podatkov. To zadošča za »agent se je na tem benchmarku izkazal impresivno«. Ne zadošča za »agent je boljši diagnostik od zdravnika« in avtorji drugega niti ne trdijo.

Najuporabnejše stališče zato ni niti »AI premaga zdravnike« niti »to je igrača«. Je: nova vrsta medicinske AI — takšna, ki *deluje* čez zapis namesto da odgovarja na posamezna vprašanja — se je dobro odrezala na zahtevnem retrospektivnem benchmarku, zdaj pa se mora dokazati tam, kjer še ni bila preizkušena: prospektivno, pri resničnih nediferenciranih bolnikih in pod jasnim upravljanjem.

Zakaj je pomembno

V zadnjih nekaj letih se je zanimivo vprašanje v medicinski AI tiho premaknilo. Nekoč je bilo »ali model postavi pravo diagnozo?« — modeli pa so na vedno bolj urejenih testnih nizih odgovarjali da. MIRA pomeni prehod k težjemu vprašanju: »ali model lahko *opravi delo* — zbere podatke, naroči,

interpretira, odloči in ukrepa — čez celoten potek?« To je uporabnejše in poštenejše vprašanje, ker prav v dejanjih živita tako težavnost kot tveganje.

Zato je okvir pomemben. Refleksni naslov *AI prekaša zdravnike* kaže na najmanj nov in najmanj podprt del rezultata. Resnično novo je manjše, a pomembnejše: agent, ki se lahko premika skozi celoten zapis. Če se ta zmožnost potrdi, bo najprej spremenila **potek dela**, veliko prej kot vprašanje, kdo ga vodi. Zdravnik ni zamenjan; naročanje, interpretiranje in sledenje rezultatom — sam workflow — je mesto, kjer bo takšno orodje najprej pristalo.

In dokazno breme prestavi v pravo smer. Zmaga na lestvici retrospektivnih primerov je razlog za prospektivno preskušanje, ne njegovo nadomestilo. Avtorji pravijo isto. Pošteno branje MIRE je povabilo k naslednji študiji — z merilom, ki ga medicina že pozna: merjeno na bolnikih, ne na benchmarkih.

Kratek povzetek

MIRA je avtonomen agent AI, ki deluje v peskovniškem elektronskem zdravstvenem zapisu: lahko zbere anamnezo, naroči in interpretira laboratorijske, slikovne in mikrobiološke preiskave, postavi diagnozo in pripravi načrt zdravljenja. Na 574 retrospektivnih primerih iz javne baze MIMIC-IV, pri osmih vnaprej izbranih diagnozah, je dosegla višjo diagnostično natančnost kot zdravniki (88,9 % na vseh primerih; v neposredni primerjavi 87,8 % proti 78,1 %) in sprejemala večinoma s smernicami skladne ter varne odločitve glede zdravil. Toda ocenjevanje je bila simulacija na preteklih zapisih in samo z besedilom; velik del prednosti je prišel pri boleznih z jasnimi testi; MIRA je uporabila precej več krvnih preiskav kot zdravniki (okoli 51 % razpoložljivih analitov proti 28 %); več izidov zdravljenja je bilo ocenjenih glede na prvotno kartoteko; javni podatkovni niz pa ustvarja resnično tveganje kontaminacije učnih podatkov — kar avtorji sami opisujejo kot možno zgornjo mejo svojih številk. Resnični napredek je agent, ki deluje čez celoten workflow namesto da odgovarja na izolirana vprašanja. Avtorji izrecno poudarjajo, da generalizacija, varnost in upravljanje zahtevajo prospektivne študije v resničnem svetu — to je pravilno branje: impresiven prikaz zmogljivosti, ne dokaz, da AI diagnosticira bolje od zdravnikov, in ne sistem, pripravljen za pravo kliniko.

Preverjanje brez olepševanja

Kaj članek pokaže: Agent jezikovnega modela MIRA lahko v peskovniškem EHR vodi celoten klinični primer od začetka do konca — anamneza, preiskave, diagnoza, naročila — in je na retrospektivnem benchmarku 574 primerov z osmimi diagnozami dosegel višjo diagnostično natančnost kot zdravniki (88,9 % na vseh; 87,8 % proti 78,1 % specialistično certificiranih zdravnikov) brez napak visoke resnosti pri zdravilih.

Kaj je verjetno, vendar ne dokazano: Da se ta zmožnost prenese v korist pri resničnih, nediferenciranih bolnikih; da prednost v natančnosti pomeni boljše sklepanje in ne več naročenih testov, skladnost z zgodovinsko kartoteko ali pomnjenje javnih podatkov.

Česa ne pokaže: Da MIRA deluje zunaj osmih diagnoz; da jo je varno uvesti; da premaga zdravnike v pošteni primerjavi z enako količino informacij; da nadomešča klinike; da rezultat ni kontaminiran z učnimi podatki.

Glavne omejitve: Retrospektivno, simulirano, samo besedilno ocenjevanje; osem vnaprej izbranih diagnoz; informacijska asimetrija (veliko več krvnih testov — okoli 51 % razpoložljivih analitov proti 28 %, predvsem poceni krvne preiskave, ne slikanje); izidi zdravljenja deloma ocenjeni glede na zgodovinsko kartoteko; in javni benchmark MIMIC-IV, ki ga je jezikovni model lahko videl med učenjem — avtorji sami to označujejo kot možno zgornjo mejo zmogljivosti.

Koliko zaupanja naj ima splošni bralec? Visoko, da je MIRA resničen in nov prikaz zmogljivosti — agent, ki *deluje* čez zapis, ne zgolj klepetalnik. Nizko, da je *boljši diagnostik kot zdravnik*: ta trditev je omejena na en retrospektivni benchmark in zamegljena z naročanjem testov, ocenjevanjem glede na kartoteko ter možno kontaminacijo. Zaključek avtorjev je pravi — preden rezultat karkoli pomeni za oskrbo bolnikov, potrebuje prospektivno študijo v resničnem svetu.

Viri

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