The Clean Paper

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

AI-agent töötas terved patsiendijuhud iseseisvalt läbi — simulaatoris, mineviku haiguslugudel

MIRA on uut tüüpi meditsiiniline AI: ühe küsimuse vastamise asemel töötab ta simuleeritud haigusloos läbi terve juhu — võtab anamneesi, tellib ja loeb teste, jõuab diagnoosini ning kirjutab korraldusi. Avaliku andmebaasi 574 retrospektiivsel haigusjuhul kaheksa eelvalitud diagnoosi piires teatasid autorid paremast diagnostilisest täpsusest kui arstidel ning valdavalt ravijuhistega kooskõlas ja ravimiohututest otsustest. Iga täpsustus selles lauses loeb: see töötas liivakastis mineviku andmetel, ainult tekstis; suur osa eelisest tuli selgete testitulemustega haigustest; ta tellis umbes kaks korda rohkem vereanalüüse kui arstid; ja mitut tulemust hinnati algses haigusloos kirjapandu järgi. Päris edasiminek on kogu töövoos tegutsev agent — mitte tõend, et masin diagnoosib nüüd arstist paremini. Autorid ütlevad sama: üldistamine, ohutus ja juhtimine vajavad veel prospektiivseid reaalelu uuringuid.

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

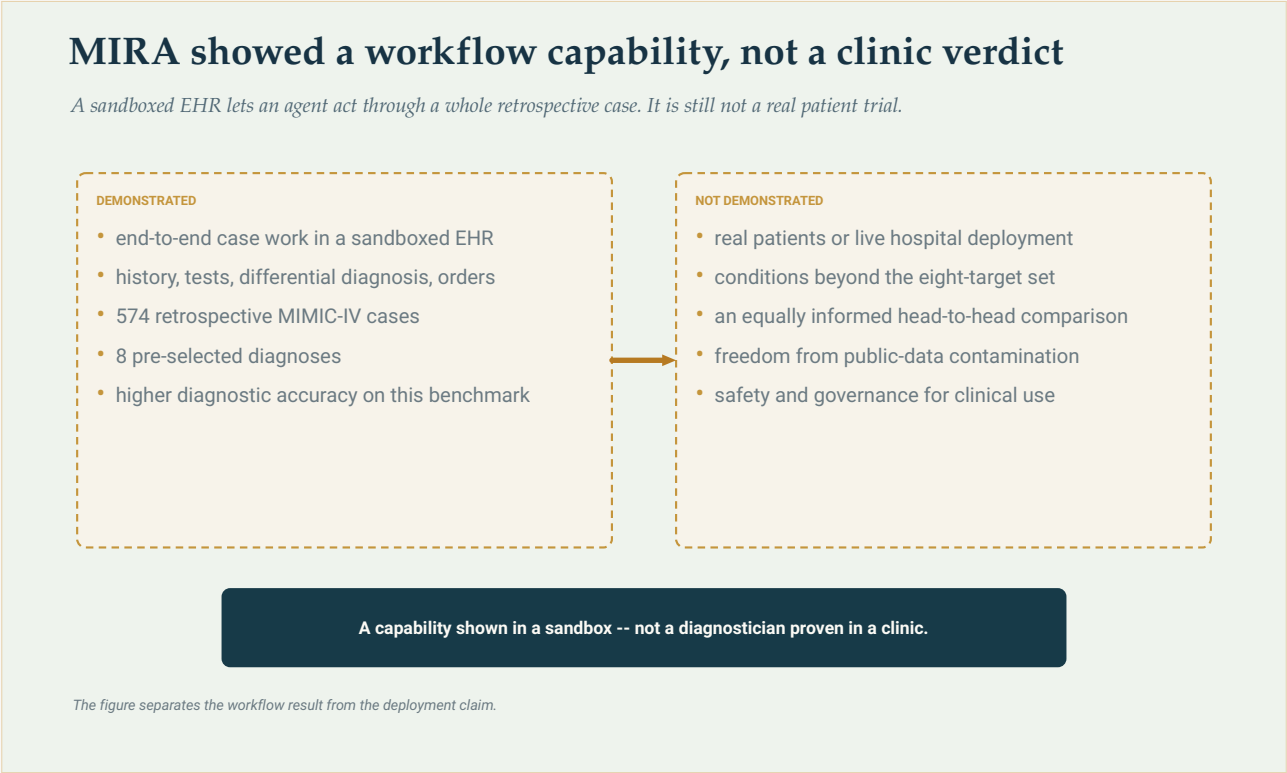
Küsimusele vastamine ei ole sama mis kogu haigusjuhu juhtimine

Enamik meditsiinilise AI lugusid räägib mudelist, mis *vastab*. Annad talle haigusjuhu kirjelduse või eksamiküsimuse, tema annab diagnoosi või nõuandva lõigu ning saab hea skoori. Kasulik, kuid kitsas: nutikas konsultant, kes haiguslugu kunagi ei puuduta.

MIRA on ehitatud tegema midagi muud ja just see erinevus on päris uudis. Ühele küsimusele vastamise asemel töötab ta läbi terve haigusjuhu: loeb patsiendi haiguslugu, otsustab, millist anamneesi veel vaja on, tellib labori-, pildistamis- ja mikrobioloogiauuringuid, loeb tulemusi, kitsendab diferentsiaaldiagnoosi ning kirjutab lõpuks järgnevad korraldused — retseptid, hospitaliseerimise, kirurgilise suunamise. Ta teeb seda, sooritades toiminguid *elektroonilise terviseloo sees* nagu klinitsist, mitte kirjutades vabateksti, mille inimene peaks ümber sisestama.

See on päris samm: nõu andvast süsteemist kogu töövoos tegutseva süsteemini. Ja just selline samm surutakse kajastuses kergesti kokku lauseks „AI edestab arste". Seepärast tasub olla täpne, mida testiti ja kus.

Ühe lause versioon: MIRA töötas haigusjuhud iseseisvalt lõpuni läbi ning edestas selles võrdlustestis arste diagnostilise täpsuse poolest — kuid tegi seda liivakastis, retrospektiivsetel andmetel, kaheksa eelvalitud diagnoosi piires, ainult teksti kaudu, ning suur osa eelisest tuli haigustest, millel olid kõige selgemad testitulemused. Edasiminek on päris. Edetabel ei ole kliinik.



MIRA demonstreeris liivakasti-EHR-is tervikliku töövoo võimekust. See ei demonstreerinud reaalset kliinilist kasutuselevõttu, laia diagnostilist katvust ega arstidega õiglast, võrdselt informeeritud otsest võrdlust.

Mida autorid tegid

MIRA — Medical Intelligence for Reasoning and Action — on OpenAI mudelitele ehitatud autonoomne agent: vestlev ja tegutsev osa töötab **GPT-4o** peal ning eraldi planeerimisetapp kasutab OpenAI **o1** arutlusmudelit. Need asuvad kohandatud standarditele vastavas raamistikus, mis põhineb HL7 FHIR-il — andmestandardil, mida päris haiglad haiguslugude liigutamiseks kasutavad — ning mille tööriistakastis on üle kaheksakümne tuhande võimaliku kliinilise toimingu. Liivakastis saab MIRA tuua patsiendi anamneesi, tellida ja tõlgendada labori-, pildistamis- ja mikrobioloogiauuringuid, koostada diferentsiaaldiagnoosi ning formuleerida raviplaane — määrata ravimeid, planeerida kirurgilisi protseduure ja hospitaliseerimist. Üks detail tasub kohe välja tuua: MIRA vastuseid hinnanud automatiseeritud „kohtunikud" olid ise GPT-4o — sama mudeliperekond, mida testiti — ning autorid lisasid sellele pärast eelretsensendi tõstatatud muret eriarsti sertifikaadiga arsti ülevaatuse.

Testikomplekt pärines **MIMIC-IV-st**, suurest avalikult kättesaadavast anonümiseeritud elektrooniliste terviseandmete andmebaasist. Umbes 300 000 patsiendist, keda raviti Beth Israel Deaconess Medical Centeris 2008.–2019. aastal, valisid autorid **574 haigusjuhtu kaheksa sihtdiagnoosi** kohta — kõhupiirkonna patoloogiad ja sisehaiguste erakorralised seisundid, muu hulgas apenditsiit, pankreatiit, pneumoonia ja kuseteede infektsioon. Iga juhu puhul alustas MIRA piiratud infoga ja pidi samm-sammult otsustama, mida edasi teha — sama kujuga protsess nagu päris diagnostiline töö, kuid ajaloolise haigusloo vastu, mille tegelik lõpptulemus oli juba kirjas.

Seejärel võrreldi tema tulemust arstidega samadel juhtudel.

Mida „autonoomne agent liivakasti-EHR-is" tegelikult tähendab

Kolm sõna teevad siin palju tööd, seega tasub need lahti võtta.
Agent tähendab, et süsteem pole lihtsalt viibale vastav vestlusrobot. See töötab tsüklina: vaata praegust seisundit, vali toiming (telli see test, küsi seda anamneesi), vaata tulemust, vali järgmine toiming — kuni jõuab diagnoosi ja plaanini. „Intelligents" on keelemudel; *agentsus* on selle ümber olev taristu, mis muudab mudeli teksti lubatud EHR-toiminguteks ja söödab tulemused tagasi.
Liivakasti-EHR tähendab simuleeritud ja eraldatud meditsiinilise infosüsteemi koopiat, mitte päris haigla süsteemi. MIRA võib „tellida" testi ja saada tulemuse, mille see patsient päriselus tegelikult sai, sest juhtum on ajalooline ja vastus on juba kirjas. Ükski MIRA toiming ei puuduta päris patsienti ega päris kliinikut.
Autonoomne tähendab, et ta viib haigusjuhu algusest lõpuni läbi ilma inimeseta tsüklis — *liivakasti sees*. See **ei** tähenda, et süsteem lasti järelevalveta päris patsiente juhtima. Need on väga erinevad väited ja testiti ainult esimest.
Veel üks detail liivakasti kohta, sest seda on lihtne valesti ette kujutada: MIRA võib *tellida* ükskõik millise testi, kuid süsteem saab tulemuse tagastada ainult siis, kui see test **päriselt**

> **tehti** selle patsiendi ravis. Küsi verepaneeli, mille patsient sai, ja saad päris väärtused; küsi skaneeringut, mida keegi ei tellinud, ja vastuseks tuleb „N/A — could not be performed", mitte väljamõeldud arv. Anamnees töötab samamoodi: kui haigusloos pole vastust MIRA küsimusele, ütleb simuleeritud patsient lihtsalt, et ei tea. Seega ei saa MIRA võluda välja üht otsustavat testi, mida kunagi ei tehtud — ta saab töötada ainult selle materjaliga, mida päris diagnostiline protsess juhtumisi sisaldas.
> Eristus on tähtis, sest pealkirjasõna „autonoomne" on just see, mida kõige kergemini loetakse teise tähendusega.

Mida nad leidsid

Võrdlustestis **edestas MIRA arste diagnostilise täpsuse poolest** — kuid pealkiri peidab kolme eri arvu ja neid on lihtne segi ajada, seega tasub need lahus hoida.

Kõigi **574 juhu** peale nimetas MIRA õige diagnoosi **88,9%** juhtudest — hinnatuna MIMIC-IV-s iga patsiendi kohta tegelikult kirja pandud väljakirjutamisdiagnoosi järgi. Inimestega otseses võrdluses ei töötanud arstid läbi kõiki 574 juhtu; nad lahendasid ühise **311 juhu alamhulga**, kasutades samu haiguslugusid ja tööriistu nagu MIRA. Neil samadel 311 juhul sai MIRA **87,8%** ning teda võrreldi kahe eraldi värvatud rühmaga. Esimene oli **vanem rühm**: neli eriarsti sertifikaadiga arsti 7–11 aasta kogemusega, tulemusega **78,1%**. Teine oli **segase staažiga rühm**: peamiselt nooremad residentuuris arstid koos kahe spetsialistiga, mis sarnaneb rohkem päris erakorralise meditsiini osakonna personalile, tulemusega **71,1%**. Samad juhud, samad tööriistad — järjestus oli AI esimene, kogenud arstid teised, nooremate osakaaluga meeskond kolmas. Töö enda ettevaatlik sõnastus on, et MIRA oli kõigi kaheksa haiguse lõikes arstidega „järjekindlalt samaväärne ja sageli parem".

Ka järgnevad otsused pidasid üsna hästi: need olid valdavalt juhistega kooskõlas, ravimiohutud ja hospitaliseerimise osas sobivad. Autorid ei teata ühestki kõrge raskusastmega ravimiveast viies ohutuskategoorias — koostoimed, neerufunktsiooni järgi annustamine, allergiad, QT-risk ja opioidide määramine — ning retseptid olid õiged 467 juhul 468-st, märkides samas, et süsteem „ei saavutanud 100% töökindlust". Otse vaadates on see tähelepanuväärne: kogu haigusjuhu läbi töötav agent tuli diagnoosis arstidest ette.

Kuid võidu *kuju* loeb sama palju kui võit ise. Tööd lugenud sõltumatud spetsialistid osutasid sellele, kust eelis tegelikult tuli. Science Media Centre'i ekspertpaneelis märkis dr Wei Xing (Sheffieldi ülikool), et pealkirjanumber — AI edestab arste diagnostilises täpsuses — oli **suures osas tingitud selgete testitulemustega haigustest**, nagu apenditsiit ja pankreatiit, kus otsustav pilt või laborivääritus lahendab küsimuse. Pneumoonia ja kuseteede infektsioonide puhul — kaks väga tavalist põhjust erakorralise meditsiini osakonda pöördumiseks — said *nii* AI kui arstid kõige kehvemad tulemused ning vahe nende vahel oli väikseim.

Ka mängureeglites oli asümmeetria ja see on töö enda arvudes nähtav. MIRA toetus laborile tugevamalt: ta kasutas umbes **51%** tavaravis saadaolevatest laborianalüütidest, võrreldes eriarsti sertifikaadiga arstide ligikaudu **28%-ga** — mediaanina seitse vereparameetrit rohkem ühe juhu

kohta. Autorid rõhutavad, et see jääb *alla* andmestiku enda tavaravi lähtejoone ega tähenda „telli kõik" strateegiat, ning ei leia süstemaatilist kasvu kallima ristlõikepildistamise kasutuses. Ometi võib rohkem infot iseenesest diagnostilist täpsust parandada — seega pole see päriselt võrdne võistlus võrdselt informeeritud osalejate vahel, millele Xing otseselt tähelepanu juhtis. Ka arst, kes võiks tellida kõik odavad vereanalüüsid ilma kulu, ebamugavust või viivitust kaalumata, paistaks paberil teravam.

Miks „edestas arste" ei tähenda „parem arst"

Võrdlustesti võitu on lihtne üle lugeda. „MIRA sai kõrgema skoori" ja „MIRA on parem" vahel seisab kolm asja.

Esiteks **mille vastu seda hinnati**. Professor Julie Jacko (Edinburghi ülikool) märkis, et mitu MIRA põhiväljundit on määratletud algandmestikus dokumenteeritu suhtes — süsteemi premeeritakse **kirja pandud kliinilise käitumise kordamise** eest, mitte tingimata optimaalse ravi demonstreerimise eest. Ajaloolisest haigusloost saab vastusevõti. See on mõistlik viis võrdlustesti ehitada, kuid mõõdab kooskõla tehtuga, mitte õigsust mingis absoluutses mõttes. Töö suhtes aus olles mõjutab see kõige tugevamalt *ravikooskõla* mõõdikuid — kas MIRA korraldused vastasid haigusloole? — samal ajal kui pealkirjas olev diagnostiline täpsus hinnatakse väljakirjutamisdiagnoosi järgi, mis on päris tulemusele lähemal kui dokumentatsiooni kaja.

Teiseks juba mainitud **infoasümmeetria**: palju rohkem vereanalüüse — umbes 51% saadaolevatest analüütidest võrreldes 28%-ga — tähendab rohkem tõendeid. Osa täpsuse erinevusest on tõenäoliselt *ostetud*, mitte välja arutletud.

Kolmandaks **kus see töötas**. See oli liivakast, retrospektiivsetel juhtudel, kaheksa eelvalitud diagnoosi piires ja ainult tekstis. Nagu dr Dominic Oliver (Oxfordi ülikool) märkis, sõltub päris kliiniline hindamine mitte ainult sellest, mida patsiendid ütlevad, vaid ka *kuidas* nad seda ütlevad — lisaks füüsilisest läbivaatusest, nähtavast käitumisest ja kehakeelest, millest valmis tekstilist haiguslugu lugev agent ei näe midagi. Ja patsient, kelle probleem ei kuulu ühegi kaheksa valitud diagnoosi alla, on lihtsalt väljaspool seda, mille kohta uuring midagi öelda saab.

Arvustajad tõstatasid ka vaiksema mure: **andmesaaste**. MIMIC-IV on avalik ja sellest on palju kirjutatud, seega võib avatud internetil treenitud keelemudel olla juba näinud artikleid, haigusjuhtude arutelusid või andmeid endid. Dr Midhun Parakkal Unni (Sheffieldi ülikool) märkis seda otse — kui osa vastuseid oli treeningandmetes, pärineb osa sooritusest mälust, mitte kliinilisest arutlusest, ning ainult sõltumatu replikatsioon saab need eristada. Oluline on, et autorid ei lükka seda käega kõrvale: nad kirjutavad, et tulemusi „võib ettevaatlikult tõlgendada võimaliku ülemise piirina" ja et need „võivad üle hinnata üldistumist teistele avalikele juhtudele" — töö paneb omaenda pealkirjale lae.

Miski sellest ei tee MIRAt vähem huvitavaks. See teeb õige lugemise kalibreerituks: retrospektiivses võrdlustestis edestas kogu haigusloo sees tegutseda suutnud agent arste selgete testidega diagnoosides — osaliselt rohkem teste tellides, osaliselt ajaloolise haiguslooga nõustudes. See on päris võimekuse demonstratsioon, mitte otsus, et masinad diagnoosivad nüüd arstidest paremini.

Mida see *ei* tõesta

- See **ei** näita, et MIRA töötab päris patsientidega. Kõik juhud olid retrospektiivsed ja simuleeritud; MIRA ei juhtinud ühtki päris patsienti.
- See **ei** näita toimimist väljaspool kaheksat diagnoosi. Eelvalitud komplektist väljapoole jäävaid seisundeid — meditsiini segast ja diferentseerimata enamust — ei testitud.
- See **ei** näita, et süsteemi oleks ohutu kasutusele võtta. Autorid ise kirjutavad, et üldistamine, ohutus ja juhtimine vajavad veel prospektiivseid reaalelu uuringuid.
- See **ei** loo arstidega täiesti õiglast otsest võrdlust. MIRA kasutas palju rohkem vereanalüüse (umbes 51% saadaolevatest analüütidest vs 28%) ning mitut ravitulemust hinnati osaliselt selle järgi, mida ajaloolises haigusloos tehti, mitte parima ravi objektiivse tõe järgi.
- See **ei** näita, et tulemus oleks mälust mõjutamisest vaba. Avalik MIMIC-IV andmestik võib kattuda mudeli treeningandmetega ning selle välistamiseks on vaja sõltumatut replikatsiooni.
- See **ei** tähenda „AI asendab arste". See on otsustustugi, mis suudab haigusloo sees *tegutseda*; arvustajate ühine hinnang on, et päris kasutus toimub koostöös klinitsistidega, kellele jääb otsustusõigus ja kes annavad kõik selle, mille tekstiline haiguslugu välja jätab.

Kui tugevad on tõendid?

Jagagem väide kaheks, sest nende kahe poole tõendid on väga erinevad.

Võimekuse demonstratsioonina — et keelemudeli agent suudab EHR-liivakastis kogu kliinilise haigusjuhu läbi viia, sidudes anamneesi, testid, diagnoosi ja korraldused algusest lõpuni — on töö päriselt uudne ja üsna veenev. See on uus osa ja just sellele tasub tähelepanu pöörata.

Üleolekuväitena — et MIRA on *arstidest parem* — on tõendusmaterjal piiratud ja seda tuleb lugeda ettevaatlikult. See kehtib ühe konkreetse retrospektiivse võrdlustesti, kaheksa diagnoosi, infoasümmeetria ehk rohkemate testide, ajaloolisest haigusloost pärit vastusevõtme ja reaalse treeningandmete saastumise võimaluse tingimustes. Sellest piisab ütlemaks „agent esines selles võrdlustestis muljetavaldavalt". Sellest ei piisa ütlemaks „agent on arstist parem diagnostik" ning autorid teist väidet ei esita.

Kõige kasulikum hoiak pole „AI võidab arste" ega „lihtsalt mänguasi". Pigem: uut tüüpi meditsiiniline AI — mis küsimustele vastamise asemel *tegutseb* haigusloo sees — esines raskes retrospektiivses võrdlustestis hästi ning peab nüüd ennast tõestama seal, kus teda veel testitud pole: päris, diferentseerimata patsientidel, prospektiivselt ja juhtimisraamistiku all.

Miks see oluline on

Viimastel aastatel on huvitav küsimus meditsiinilises AI-s vaikselt muutunud. Varem oli see „kas mudel saab diagnoosi õigesti?" — ja mudelid vastasid üha puhtamatel testikomplektidel aina „jah".

MIRA tähistab nihet raskema küsimuseni: „kas mudel suudab *teha töö ära* — koguda, tellida, tõlgendada, otsustada, tegutseda — läbi terve töövoo?" See on kasulikum ja ausam küsimus, sest tegutsemises elavad nii raskus kui risk.

Seepärast loeb raamistik. Pealkirjarefleks — *AI edestab arste* — osutab tulemuse kõige vähem uudsele ja kõige nõrgemini toetatud osale. Päriselt uus asi on väiksem ja tagajärjekam: agent, mis suudab liikuda läbi kogu haigusloo. Kui see võimekus vastu peab, muudab see **töövoogu** ammu enne seda, kui muudab, kes seda juhib. Arst ei kao; tellimine, tõlgendamine, tulemuste tagaajamine — töövoog — on see, kuhu selline tööriist kõigepealt maandub.

Ja see seab tõendamiskohustuse õigesse suunda. Edetabelivõit retrospektiivsetel juhtudel on põhjus teha prospektiivne uuring, mitte selle asendus. Autorid ütlevad sama. MIRA aus lugemine on kutse sellele järgmisele uuringule — standardiga, mida valdkond juba tunneb: mõõdetuna patsientidel, mitte võrdlustestidel.

Lühidalt

MIRA on autonoomne AI-agent, mis töötab liivakasti-elektroonilises terviseloos: saab võtta anamneesi, tellida ja tõlgendada labori-, pildistamis- ja mikrobioloogiauuringuid, jõuda diagnoosini ning kirjutada ravikorraldusi. Avaliku MIMIC-IV andmebaasi 574 retrospektiivsel haigusjuhul kaheksa eelvalitud diagnoosi piires edestas ta arste diagnostilise täpsuse poolest (88,9% kokku; otseses võrdluses 87,8% vs 78,1%) ning tegi valdavalt ravijuhistega kooskõlas ja ravimiohutuid otsuseid. Kuid hindamine oli mineviku haiguslugude simulatsioon, ainult tekstis; suur osa eelisest tuli selgete testitulemustega haigustest; MIRA kasutas arstidest palju rohkem vereanalüüse (umbes 51% saadaolevatest analüütidest vs 28%); mitut ravitulemust hinnati selle järgi, mida algne haiguslugu kirjeldas; ning avalik andmestik tekitab päris treeningandmete saastumise riski — mida autorid ise nimetavad oma arvude võimalikuks ülemiseks piiriks. Tõeline edasiminek on agent, mis tegutseb üksikküsimustele vastamise asemel kogu töövoos. Autorid ütlevad selgelt, et üldistamine, ohutus ja juhtimine vajavad endiselt prospektiivseid reaalelu uuringuid — ja see on õige lugemine: muljetavaldav võimekuse demonstratsioon, mitte tõend, et AI diagnoosib arstidest paremini, ega süsteem, mis on valmis päris kliinikusse.

Ilustamata kontroll

Mida töö näitab: Keelemudeli agent MIRA suudab liivakasti-EHR-is läbi viia terve kliinilise haigusjuhu — anamnees, testid, diagnoos, korraldused — ning 574 juhu retrospektiivsel võrdlustestil kaheksa diagnoosi piires edestas arste diagnostilise täpsuse poolest (88,9% kokku; 87,8% vs 78,1% eriarsti sertifikaadiga arstidega otseses võrdluses), ilma kõrge raskusastmega ravimivigadeta.

Mis on usutav, kuid tõestamata: Et see võimekus toob kasu päris diferentseerimata patsientidele; et täpsuse eelis peegeldab paremat kliinilist arutlust, mitte rohkem tellitud teste, kooskõla ajaloolise haigusloogа või avalike andmete meeldejätmist.

Mida see ei näita: Et MIRA töötab väljaspool kaheksat diagnoosi; et seda on ohutu kasutusele võtta; et see võidab arste õiglases, võrdselt informeeritud võrdluses; et see asendab klinitsiste; et tulemused on treeningandmete saastumisest vabad.

Peamised piirangud: Retrospektiivne, simuleeritud ja ainult tekstiline hindamine; kaheksa eelvalitud diagnoosi; infoasümmeetria (palju rohkem vereanalüüse — umbes 51% saadaolevatest analüütidest vs 28%, odavad vereanalüüsid, mitte pildistamine); ravitulemused osaliselt hinnatud ajaloolise haigusloo järgi; ning avalik võrdlustest MIMIC-IV, mida keelemudel võis treeningul näha — autorid märgivad seda võimaliku soorituse ülemise piirina.

Kui palju võiks tavalugeja seda usaldada? Kõrge kindlus, et MIRA on päris ja uudne võimekuse demonstratsioon — agent, mis *tegutseb* haigusloo sees, mitte lihtsalt vestlusrobot. Madal kindlus, et see on *arstist parem diagnostik*: väide on piiratud ühe retrospektiivse võrdlustestiga ning seda segavad testide tellimise erinevus, haigusloo järgi hindamine ja võimalik andmesaaste. Autorite enda lõppjäreldus on õige — enne kui sellel on patsiendiravi jaoks tähendus, vajab see prospektiivset reaalelu uuringut.

Allikad

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