The Clean Paper

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

# Një agjent AI punoi vetë raste të plota pacientësh — në simulator, mbi kartela të së kaluarës

*MIRA është një lloj i ri AI-je mjekësore: në vend që t'i përgjigjet një pyetjeje të vetme, punon një rast të tërë brenda një kartelë spitalore të simuluar — merr histori, porosit e lexon teste, arrin diagnozë dhe shkruan urdhra. Në 574 raste retrospektive nga një bazë publike, në tetë diagnoza të zgjedhura paraprakisht, autorët raportojnë saktësi diagnostike më të lartë se mjekët dhe vendime kryesisht në përputhje me udhëzimet e të sigurta për barnat. Por çdo kualifikim ka peshë: sistemi punoi në sandbox mbi kartela të vjetra, vetëm në tekst; një pjesë e avantazhit erdhi te gjendjet me rezultate testesh të qarta; porositi afërsisht dy herë më shumë analiza gjaku se mjekët; dhe disa rezultate u vlerësuan kundrejt asaj që ishte dokumentuar në kartelën origjinale. Përparimi real është një agjent që vepron përgjatë gjithë workflow-it — jo provë se makina tani diagnostikon më mirë se mjeku.*

---

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

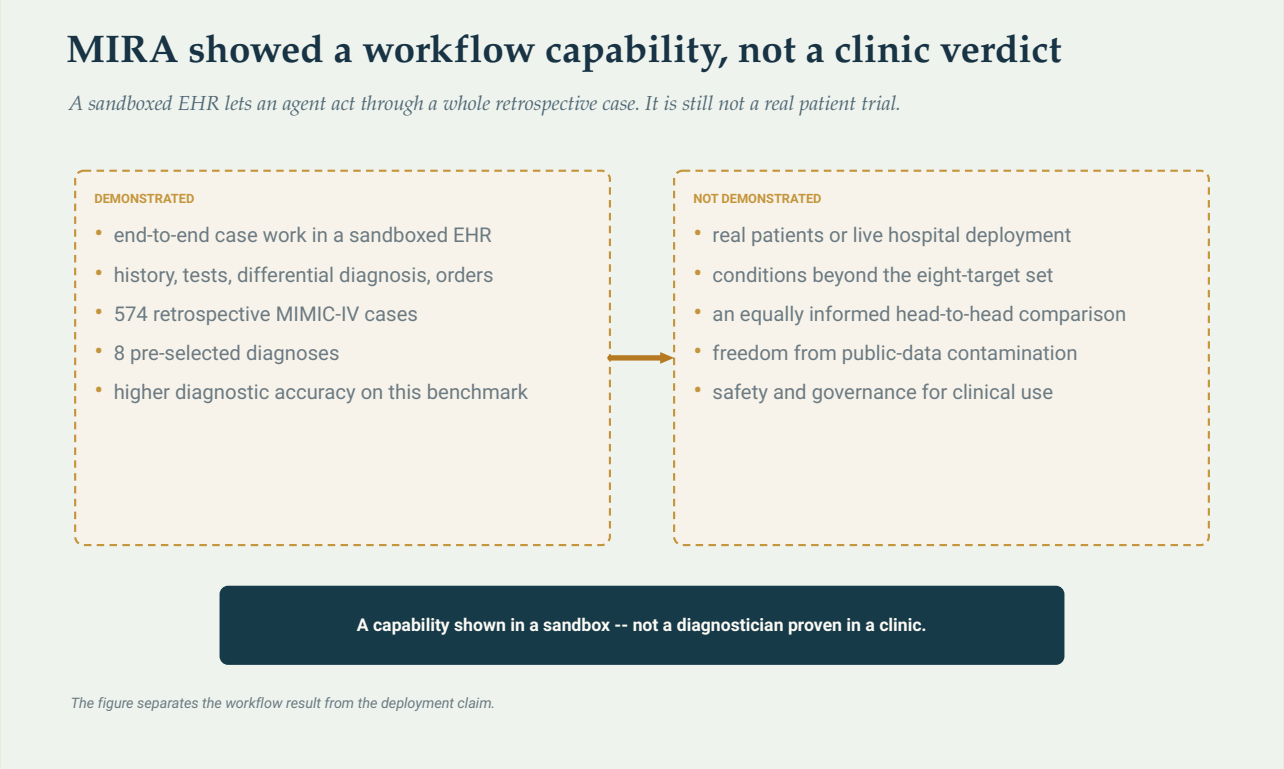
# T'i përgjigjesh një pyetjeje nuk është e njëjta gjë me ta menaxhosh gjithë rastin

Shumica e historive për AI mjekësore flasin për një model që *përgjigjet*. I jep një vignette klinike ose një pyetje provimi, ai kthen një diagnozë ose një paragraf këshille dhe merr rezultat të mirë. E dobishme, por e ngushtë: një konsulent i zgjuar që nuk e prek kurrë kartelën.

MIRA është ndërtuar për të bërë diçka tjetër, dhe pikërisht ky ndryshim është lajmi. Në vend që t'i përgjigjet një pyetjeje, ajo punon një rast të tërë: lexon kartelën e pacientit, vendos çfarë historie i mungon ende, porosit analiza laboratorike, imazheri dhe mikrobiologji, lexon rezultatet, ngushton diagnozën diferenciale dhe pastaj shkruan urdhrat që pasojnë — receta, shtrim në spital, referim për kirurgji. E bën këtë duke ndërmarrë veprime *brenda* një kartelë elektronike shëndetësore, si një klinicist, jo duke prodhuar tekst të lirë që një njeri duhet ta kopjojë më pas.

Ky është një hap real: nga një sistem që këshillon te një sistem që vepron përgjatë gjithë workflow-it. Dhe është pikërisht lloji i hapit që në mbulim mediatik sheshohet në "AI ua kalon mjekëve". Prandaj ia vlen të jemi të saktë për atë që u testua dhe ku.

Versioni me një fjali: MIRA i punoi rastet e plota vetë dhe, në këtë benchmark, pati saktësi diagnostike më të lartë se mjekët — por e bëri këtë në një sandbox, mbi kartela retrospektive, në tetë diagnoza të zgjedhura paraprakisht, vetëm me komunikim tekstual, dhe një pjesë e madhe e avantazhit erdhi te gjendjet me rezultate testesh më të qarta. Përparimi është real. Tabela e rezultateve nuk është klinika.




MIRA demonstroi aftësi për të kryer një workflow të plotë brenda një EHR-je të izoluar. Studimi nuk demonstroi përdorim klinik real, mbulim të gjerë diagnostik ose një krahasim krejtësisht të barabartë me mjekët me të njëjtin

informacion.



# Çfarë bënë autorët

MIRA — Medical Intelligence for Reasoning and Action — është një agjent autonom i ndërtuar mbi modelet e OpenAI-t: pjesa që bisedon dhe vepron përdor **GPT-4o**, ndërsa një hap i veçantë planifikimi përdor modelin arsyetues **o1** të OpenAI-t. Këto vendosen brenda një kuadri të posaçëm që respekton standardet — të ndërtuar mbi HL7 FHIR, standardin e të dhënave që spitalet reale përdorin për shkëmbimin e kartelave — me më shumë se tetëdhjetë mijë veprime klinike të mundshme. Brenda sandbox-it MIRA mund të marrë historinë e pacientit, të porosisë e interpretojë laboratorë, imazheri dhe mikrobiologji, të gjenerojë diagnozë diferenciale dhe të formulojë plan trajtimi — receta, planifikim ndërhyrjesh kirurgjikale, shtrim. Një hollësi që duhet theksuar që në fillim: "gjyqtarët" automatikë që vlerësuan përgjigjet e MIRA-s ishin vetë GPT-4o — e njëjta familje modeli që po testohej — dhe autorët e plotësuan këtë me rishikim nga një mjek i certifikuar pasi një peer reviewer ngriti shqetësimin.

Seti i testimit erdhi nga **MIMIC-IV**, një bazë e madhe publike me EHR të de-identifikuara. Nga afërsisht 300,000 pacientë të trajtuar në Beth Israel Deaconess Medical Center mes 2008 dhe 2019, autorët zgjodhën **574 raste pacientësh** që mbulonin **tetë diagnoza të synuara** — patologji abdominale dhe urgjenca të mjekësisë së brendshme, ndër to apendicit, pankreatit, pneumoni dhe infeksion të traktit urinar. Për çdo rast MIRA nisej nga një pamje e kufizuar dhe duhej të vendoste hap pas hapi çfarë të bënte më tej — e njëjta formë si një workup real, por e ekzekutuar mbi një kartelë, përfundimi real i së cilës tashmë dihej.

Më pas performanca e saj u krahasua me atë të mjekëve në të njëjtat raste.

### Çfarë do të thotë në të vërtetë "agjent autonom në një EHR të izoluar"

Tre fjalë bëjnë shumë punë këtu, ndaj ia vlen t'i zbërthejmë.
**Agjent** do të thotë se sistemi nuk është thjesht chatbot që i përgjigjet një prompt-i. Ai ekzekuton një cikël: shikon gjendjen aktuale, zgjedh një veprim (porosit këtë analizë, pyet për atë pjesë të historisë), sheh rezultatin, zgjedh veprimin tjetër — derisa arrin te një diagnozë dhe plan. "Inteligjenca" është modeli gjuhësor; *agjencia* është struktura rreth tij që e kthen tekstin e modelit në operacione të lejuara të EHR-së dhe i kthen rezultatet prapë modelit.
**EHR e izoluar në sandbox** do të thotë një kopje e simuluar dhe e mbyllur e një sistemi kartelash mjekësore, jo një sistem spitalor live. MIRA mund të "porosisë" një analizë dhe të marrë rezultatin që pacienti realisht pati, sepse rasti është historik dhe rezultati ndodhet tashmë në kartelë. Asgjë që bën nuk prek pacient real ose klinikë reale.
**Autonom** do të thotë se e përfundon rastin nga fillimi në fund pa njeri në loop — *brenda sandbox-it*. **Nuk** do të thotë se u la të menaxhonte pacientë pa mbikëqyrje. Janë dy pretendime shumë të ndryshme, dhe vetëm i pari u testua.

> Edhe një gjë për sandbox-in, sepse është e lehtë të imagjinohet gabim: MIRA mund të *porosisë* çfarëdo testi që dëshiron, por sistemi mund t'i japë rezultat vetëm nëse ai test ishte **realizuar vërtet** për atë pacient në jetën reale. Kërko një panel gjaku që pacienti e kishte bërë, dhe merr vlerat reale; kërko një skanim që askush nuk e kishte porositur, dhe merr "N/A — nuk mund të kryhet", kurrë një numër të shpikur. Historia funksionon njësoj: nëse kartela nuk përmban përgjigjen për atë që pyet MIRA, pacienti i simuluar thjesht thotë se nuk e di. Pra MIRA nuk mund të thërrasë testin vendimtar që nuk ishte bërë kurrë — mund të punojë vetëm me atë që workup-i real përfshinte.
> Dallimi ka rëndësi sepse fjala e titullit — "autonom" — është ajo që lexohet më lehtë në kuptimin e dytë.

# Çfarë gjetën

Në benchmark, **MIRA pati saktësi diagnostike më të lartë se mjekët** — por ky titull fsheh tre numra të ndryshëm, që është e lehtë të përzihen.

E vetme, në të gjitha **574 rastet**, MIRA dha diagnozën e saktë në **88.9%** të rasteve — duke u vlerësuar kundrejt diagnozës së daljes nga spitali të regjistruar realisht për çdo pacient në MIMIC-IV. Në krahasimin kokë më kokë me njerëzit, mjekët nuk punuan të gjitha 574 rastet; ata punuan një **nënset të përbashkët prej 311 rastesh**, me të njëjtat kartela dhe mjete që kishte MIRA. Në ato 311 raste, MIRA arriti **87.8%**, dhe u krahasua me dy grupe të rekrutuar veçmas. I pari ishte një **grup senior**: katër mjekë të certifikuar me 7–11 vjet përvojë, me **78.1%**. I dyti ishte një **grup me senioritet të përzier**: kryesisht rezidentë juniorë plus dy specialistë, më afër mënyrës si stafizohet realisht një urgjencë, me **71.1%**. Të njëjtat raste, të njëjtat mjete — dhe renditja doli: AI e para, mjekët me përvojë të dytët, grupi me më shumë juniorë i treti. Formulimi i kujdesshëm i punimit është se MIRA ishte "në mënyrë të qëndrueshme e barasvlershme me, dhe shpesh i tejkalonte" mjekët, në të tetë sëmundjet.

Vendimet që pasuan diagnozën qëndruan gjithashtu mirë: kryesisht në përputhje me udhëzimet, të sigurta për barnat dhe të përshtatshme për shtrim. Autorët raportojnë asnjë gabim medikamentoz me ashpërsi të lartë në pesë kategori sigurie (ndërveprime barnash, dozimi renal, alergji, rrezik QT dhe përshkrim opioidësh) dhe 467 receta të sakta nga 468 — duke theksuar megjithatë se sistemi "nuk arriti besueshmëri 100%". Marrë siç është, ky është rezultat mbresëlënës: një agjent që punon gjithë rastin dhe del përpara mjekëve në diagnozë.

Por *forma* e fitores ka po aq rëndësi sa fitorja. Specialistë të pavarur që lexuan punimin treguan se nga vinte avantazhi. Në panelin e ekspertëve të Science Media Centre, Dr Wei Xing (University of Sheffield) vuri në dukje se rezultati kryesor — AI që kalon mjekët në saktësi diagnostike — ishte **kryesisht i shtyrë nga gjendjet me rezultate testesh të qarta**, si apendiciti dhe pankreatiti, ku një skanim ose vlerë laboratorike vendimtare e mbyll çështjen. Për pneumoninë dhe infeksionet e traktit urinar — dy nga arsyet më të zakonshme pse njerëzit paraqiten realisht në urgjencë — *si* AI *ashtu edhe* mjekët patën performancën më të dobët dhe diferenca mes tyre ishte më e vogla.

Kishte edhe një asimetri në mënyrën si luajtën dy palët, dhe ajo shihet në numrat e vetë punimit.

MIRA u mbështet më shumë te laboratori: përdori rreth **51%** të analitëve laboratorikë të disponueshëm në kujdesin rutinë, kundrejt afërsisht **28%** për mjekët e certifikuar — një medianë prej shtatë parametrash shtesë gjaku për rast. Autorët e paraqesin me kujdes këtë si *më pak* se baseline-i i kujdesit rutinë të vetë dataset-it, jo si strategji "porosit gjithçka", dhe raportojnë *asnjë* rritje sistematike të imazherisë cross-sectional më të kushtueshme. Megjithatë, më shumë informacion mund, vetvetiu, të rrisë saktësinë diagnostike — pra nuk është plotësisht një duel mes palëve me informacion të barabartë, një boshllëk që Xing e theksoi drejtpërdrejt. Edhe një klinicist i lirë të porosisë çdo analizë të lirë gjaku pa peshuar koston, parehatinë ose vonesën do të dukej më i mprehtë në letër.

## Pse "ua kaloi mjekëve" nuk do të thotë "mjek më i mirë"

Një fitore në benchmark është e lehtë të mbilexohet. Tri gjëra ndodhen mes "MIRA mori rezultat më të lartë" dhe "MIRA është më e mirë".

Së pari, **kundrejt kujt u vlerësua.** Profesorja Julie Jacko (University of Edinburgh) vuri në dukje se disa nga rezultatet kyçe të MIRA-s përcaktohen në raport me atë që ishte dokumentuar në dataset-in bazë — pra sistemi shpërblehet për **riprodhimin e sjelljes klinike të regjistruar**, jo domosdoshmërisht për kujdes optimal. Kartela historike bëhet çelësi i përgjigjeve. Kjo është mënyrë e arsyeshme për të ndërtuar benchmark, por mat përputhjen me atë që u bë, jo korrektësinë në një kuptim absolut. Për drejtësi ndaj punimit, ky problem godet më shumë metrikat e *përputhjes së trajtimit* — a përputheshin urdhrat e MIRA-s me kartelën? — ndërsa saktësia diagnostike kryesore vlerësohet kundrejt diagnozës së daljes nga spitali, që është më afër një rezultati real sesa jehona e dokumentacionit.

Së dyti, **asimetri informacioni** e përmendur më lart: shumë më tepër analiza gjaku (rreth 51% e analitëve të disponueshëm kundrejt 28%) do të thotë më shumë provë. Një pjesë e hendekut në saktësi ka gjasa të jetë *blerë* me më shumë informacion, jo prodhuar vetëm nga arsyetimi.

Së treti, **ku u ekzekutua.** Ishte sandbox, raste retrospektive, tetë diagnoza të zgjedhura paraprakisht dhe vetëm tekst. Vlerësimi klinik real, siç tha Dr Dominic Oliver (University of Oxford), varet jo vetëm nga çfarë thonë pacientët, por nga *si* e thonë — së bashku me ekzaminimin fizik, sjelljen e vëzhguar dhe gjuhën e trupit, asnjë prej të cilave një agjent tekstual që lexon një kartelë të përfunduar nuk e sheh. Dhe një pacient, problemi i të cilit nuk hyn në një nga tetë diagnozat e zgjedhura, është thjesht jashtë asaj për të cilën ky studim mund të flasë.

Ka edhe një shqetësim më të heshtur që recensentët ngritën: **kontaminimi i të dhënave**. MIMIC-IV është publik dhe është diskutuar gjerësisht, kështu që një model gjuhësor i trajnuar në internetin e hapur mund të ketë parë më parë punime, diskutime rastesh ose vetë të dhënat. Dr Midhun Parakkal Unni (University of Sheffield) e theksoi drejtpërdrejt — nëse disa përgjigje ishin në të dhënat e trajnimit, një pjesë e performancës është kujtesë, jo arsyetim klinik, dhe vetëm replikimi i pavarur mund t'i ndajë. Veçanërisht, autorët nuk e anashkalojnë pikën: shkruajnë se rezultatet e tyre "mund të interpretohen me kujdes si një kufi i sipërm i mundshëm" dhe "mund të mbivlerësojnë përgjithësimin te raste të tjera publike" — një punim që i vendos vetë tavan titullit të vet.

Asgjë nga kjo nuk e bën MIRA-n më pak interesante. E bën leximin e duhur më të kalibruar: në

një benchmark retrospektiv, një agjent që mund të vepronte në gjithë kartelën ia kaloi mjekëve te diagnozat me teste të qarta — pjesërisht duke porositur më shumë teste, pjesërisht duke u përputhur me kartelën e regjistruar. Kjo është demonstrim real aftësie, jo vendim se makinat tani diagnostikojnë më mirë se mjekët.

## Çfarë *nuk* provon kjo

- **Nuk** tregon se MIRA funksionon te pacientë realë. Çdo rast ishte retrospektiv dhe i simuluar; asnjë pacient nuk u menaxhua prej saj.
- **Nuk** tregon se funksionon përtej tetë diagnozave. Gjendjet jashtë setit të zgjedhur — shumica e çrregullt dhe e padiferencuar e mjekësisë — nuk u testuan.
- **Nuk** tregon se është e sigurt për t'u vendosur. Vetë autorët shkruajnë se përgjithësimi, siguria dhe qeverisja kërkojnë ende studime prospektive në botën reale.
- **Nuk** vendos një duel të drejtë me mjekët. MIRA përdori shumë më tepër analiza gjaku (rreth 51% të analitëve të disponueshëm kundrejt 28%), dhe disa rezultate trajtimi u vlerësuan pjesërisht kundrejt kartelës historike e jo kundrejt kujdesit optimal të njohur.
- **Nuk** tregon se rezultati është i lirë nga memorizimi. Të dhënat publike MIMIC-IV mund të mbivendosen me trajnimin e modelit; këtë do ta përjashtonte vetëm replikimi i pavarur.
- **Nuk** do të thotë "AI zëvendëson mjekët". Është decision support që mund të *veprojë* në kartelë; konsensusi i recensentëve është se përdorimi real do të jetë në partneritet me klinicistët, të cilët ruajnë autoritetin dhe sjellin gjithçka që mungon në një kartelë tekstuale.

## Sa e fortë është prova?

Pretendimi duhet ndarë në dy pjesë, sepse prova është shumë e ndryshme për secilën.

**Si demonstrim aftësie** — që një agjent me model gjuhësor mund të punojë një rast të plotë klinik brenda një sandbox-i EHR, duke lidhur historinë, testet, diagnozën dhe urdhrat nga fillimi në fund — puna është vërtet e re dhe mjaft bindëse. Kjo është pjesa e re dhe ajo që meriton vëmendje.

**Si pretendim epërsie** — se MIRA është *më e mirë se mjekët* — prova është e kufizuar dhe duhet lexuar me kujdes. Vlen në një benchmark të caktuar retrospektiv, me tetë diagnoza, me asimetri informacioni (më shumë teste) dhe një çelës përgjigjesh të nxjerrë nga kartela historike, plus mundësinë reale të kontaminimit të të dhënave të trajnimit. Kjo mjafton për të thënë "agjenti pati performancë mbresëlënëse në këtë benchmark". Nuk mjafton për të thënë "agjenti është diagnostikues më i mirë se një mjek", dhe autorët nuk pretendojnë gjënë e dytë.

Qëndrimi më i dobishëm nuk është as "AI i mund mjekët", as "është vetëm lodër". Është: një lloj i ri AI-je mjekësore — që *vepron* në kartelë në vend që thjesht t'u përgjigjet pyetjeve — doli mirë në një benchmark të vështirë retrospektiv dhe tani duhet të provojë veten aty ku ende nuk është provuar: te pacientë realë, të padiferencuar, prospektivisht dhe nën qeverisje.

# Pse ka rëndësi

Prej disa vitesh pyetja interesante në AI mjekësore ka ndryshuar në heshtje. Dikur ishte “a mund ta gjejë modeli diagnozën e saktë?” — dhe modelet vazhdonin të përgjigjeshin po, në test sets gjithnjë e më të pastra. MIRA shënon kalimin te një pyetje më e vështirë: “a mund modeli *ta bëjë punën* — të mbledhë, porosisë, interpretojë, vendosë, veprojë — përgjatë gjithë workflow-it?” Kjo është pyetje më e dobishme dhe më e ndershme, sepse veprimi është aty ku jetojnë edhe vështirësia, edhe rreziku.

Prandaj korniza ka rëndësi. Refleksi i titullit — *AI ua kalon mjekëve* — tregon nga pjesa më pak e re dhe më pak e mbështetur e rezultatit. Gjëja vërtet e re është më e vogël dhe më me pasoja: një agjent që mund të lëvizë nëpër të gjithë kartelën. Nëse kjo aftësi qëndron, ajo ndryshon **workflow-in** shumë përpara se të ndryshojë kush e drejton atë. Mjeku nuk zëvendësohet; porositja, interpretimi dhe ndjekja e rezultateve — workflow-i — është vendi ku një mjet i tillë do të prekë së pari praktikën.

Dhe e rivendos barrën e provës në drejtimin e duhur. Një fitore në leaderboard me raste retrospektive është arsye për të bërë provën prospektive, jo zëvendësim për të. Autorët thonë pikërisht këtë. Leximi i ndershëm i MIRA-s është ftesë për studimin tjetër — të matur me standardin që fusha tashmë e njeh: te pacientët, jo në benchmark-e.

# Përmbledhje e shkurtër

MIRA është një agjent autonom AI që operon në një kartelë elektronike shëndetësore të izoluar: mund të marrë histori, të porosisë dhe interpretojë analiza laboratorike, imazheri dhe mikrobiologji, të arrijë një diagnozë dhe të shkruajë plane trajtimi. E testuar në 574 raste retrospektive nga baza publike MIMIC-IV, në tetë diagnoza të zgjedhura paraprakisht, ajo pati saktësi diagnostike më të lartë se mjekët (88.9% në përgjithësi; 87.8% kundrejt 78.1% në krahasimin kokë më kokë) dhe mori vendime kryesisht në përputhje me udhëzimet dhe të sigurta për barnat. Por vlerësimi ishte simulim mbi kartela të vjetra, vetëm në tekst; një pjesë e madhe e avantazhit erdhi te gjendjet me teste të qarta; MIRA përdori shumë më tepër analiza gjaku se mjekët (rreth 51% të analitëve të disponueshëm kundrejt 28%); disa rezultate trajtimi u vlerësuan kundrejt asaj që ishte regjistruar në kartelën origjinale; dhe dataset-i publik ngre rrezik real kontaminimi të të dhënave të trajnimit — diçka që vetë autorët e quajnë kufi të sipërm të mundshëm të numrave të tyre. Përparimi i vërtetë është një agjent që vepron përgjatë gjithë workflow-it në vend që t’u përgjigjet pyetjeve të izoluara. Autorët thonë qartë se përgjithësimi, siguria dhe qeverisja kërkojnë ende studime prospektive në botën reale — ky është leximi i duhur: demonstrim mbresëlënës aftësie, jo provë se AI diagnostikon më mirë se mjekët dhe jo sistem gati për një klinikë reale.

# Kontroll pa teprime

**Çfarë tregon punimi:** Një agjent me model gjuhësor (MIRA) mund të punojë një rast të plotë klinik nga fillimi në fund brenda një EHR-je të izoluar — histori, teste, diagnozë, urdhra — dhe, në një bench-

mark retrospektiv prej 574 rastesh në tetë diagnoza, pati saktësi diagnostike më të lartë se mjekët (88.9% në përgjithësi; 87.8% kundrejt 78.1% ndaj mjekëve të certifikuar), pa gabime medikamentoze me ashpërsi të lartë.

**Çfarë është e besueshme, por jo e provuar:** Se kjo aftësi përkthehet në përfitim për pacientë realë dhe të padiferencuar; ose se avantazhi në saktësi pasqyron arsyetim më të mirë e jo më shumë teste, përputhje me kartelën e regjistruar ose të dhëna publike të memorizuara.

**Çfarë nuk tregon:** Se MIRA funksionon jashtë tetë diagnozave; se është e sigurt për përdorim real; se i mund mjekët në një krahasim të drejtë me informacion të barabartë; se zëvendëson klinicistët; ose se rezultatet janë pa kontaminim nga të dhënat e trajnimit.

**Kufizimet kryesore:** Vlerësim retrospektiv, i simuluar dhe vetëm tekstual; tetë diagnoza të zgjedhura paraprakisht; asimetri informacioni (shumë më tepër analiza gjaku — rreth 51% të analitëve të disponueshëm kundrejt 28%, kryesisht analiza të lira, jo imazheri); rezultate trajtimi të vlerësuara pjesërisht kundrejt kartelës historike; dhe një benchmark publik (MIMIC-IV) që modeli mund ta ketë parë gjatë trajnimit — gjë që autorët e shënojnë si kufi të sipërm të mundshëm të performancës.

**Sa besim duhet të ketë një lexues i përgjithshëm?** Të lartë se MIRA është demonstrim real dhe i ri aftësie — një agjent që *vepron* në kartelë, jo thjesht chatbot. Të ulët se është *diagnostikues më i mirë se një mjek*: ky pretendim kufizohet në një benchmark retrospektiv dhe ndikohet nga porositja e testeve, vlerësimi kundrejt kartelës dhe kontaminimi i mundshëm. Përfundimi i vetë autorëve është ai i duhuri — nevojiten studime prospektive në botën reale para se kjo të ketë kuptim për kujdesin e pacientëve.

## Burimet

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