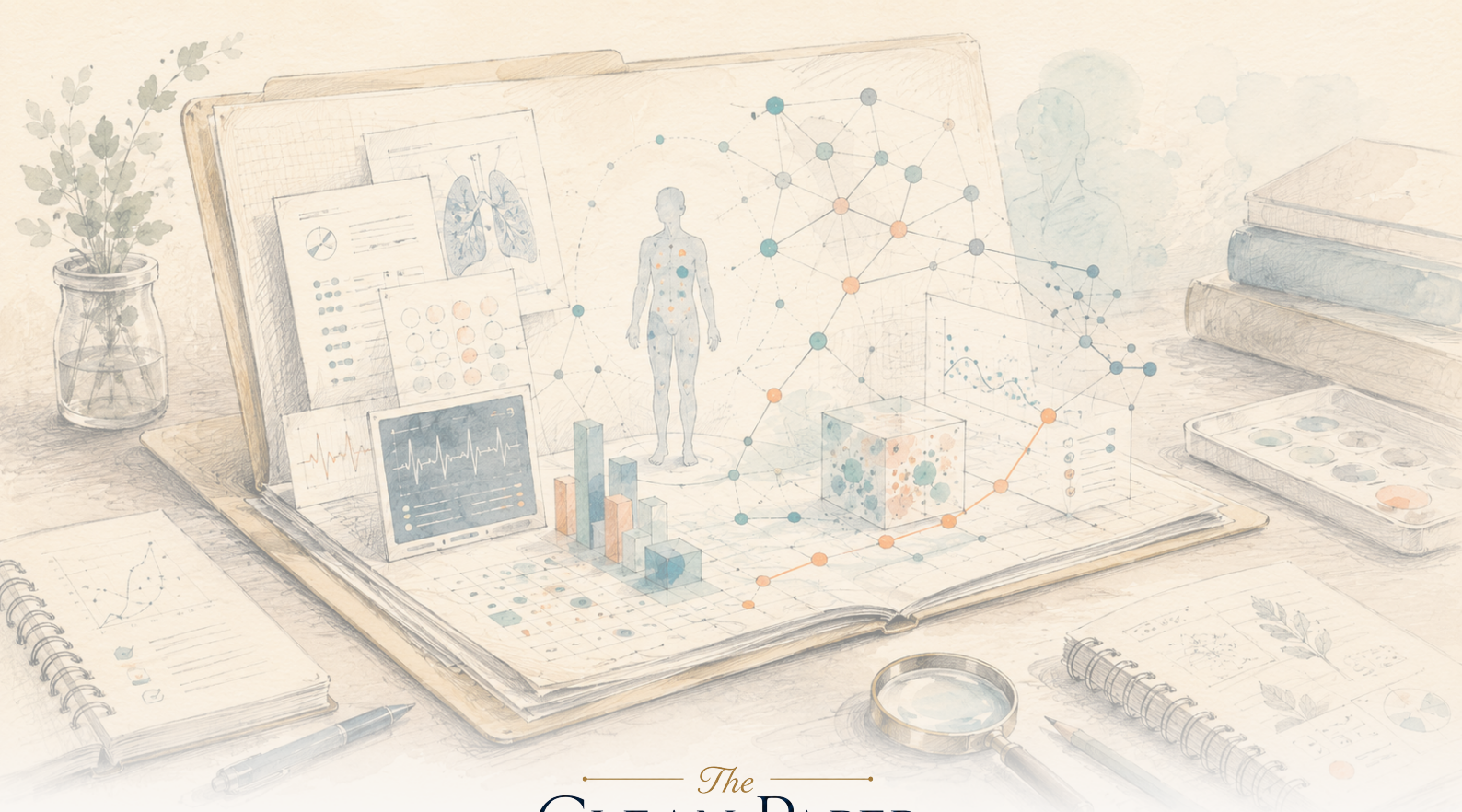


The Clean Paper

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

AI agent bütün patient case-i özbaşına işlədi — simulatorda, keçmiş records üzərində

MIRA yeni tip medical AI-dır: bir suala cavab vermək əvəzinə simulated hospital record daxilində bütün case-i işləyir — history götürür, tests order edib oxuyur, diagnosis-a çatır, orders yazır. Public database-dən səkkiz pre-selected diagnosis üzrə 574 retrospective case-də authors diagnostic accuracy üzrə physicians-ı keçdiyini və əsasən guideline-concordant, medication-safe decisions verdiyini bildirirlər. Bu cümlədə hər qualifier vacibdir: past records üzərində sandbox-da, yalnız text ilə işləyib; edge-in çoxu clear-cut test results olan conditions-dan gəlib; doctors-dan təxminən iki dəfə çox blood tests order edib; bir neçə outcome original chart-da record edilənə görə score edilib. Həqiqi yenilik whole workflow boyunca action götürən agent-dir — machine-in artıq doctor-dan daha yaxşı diagnosis qoyduğunun sübutu deyil. Authors özləri deyir: generalization, safety və governance üçün prospective, real-world studies hələ lazımdır.

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

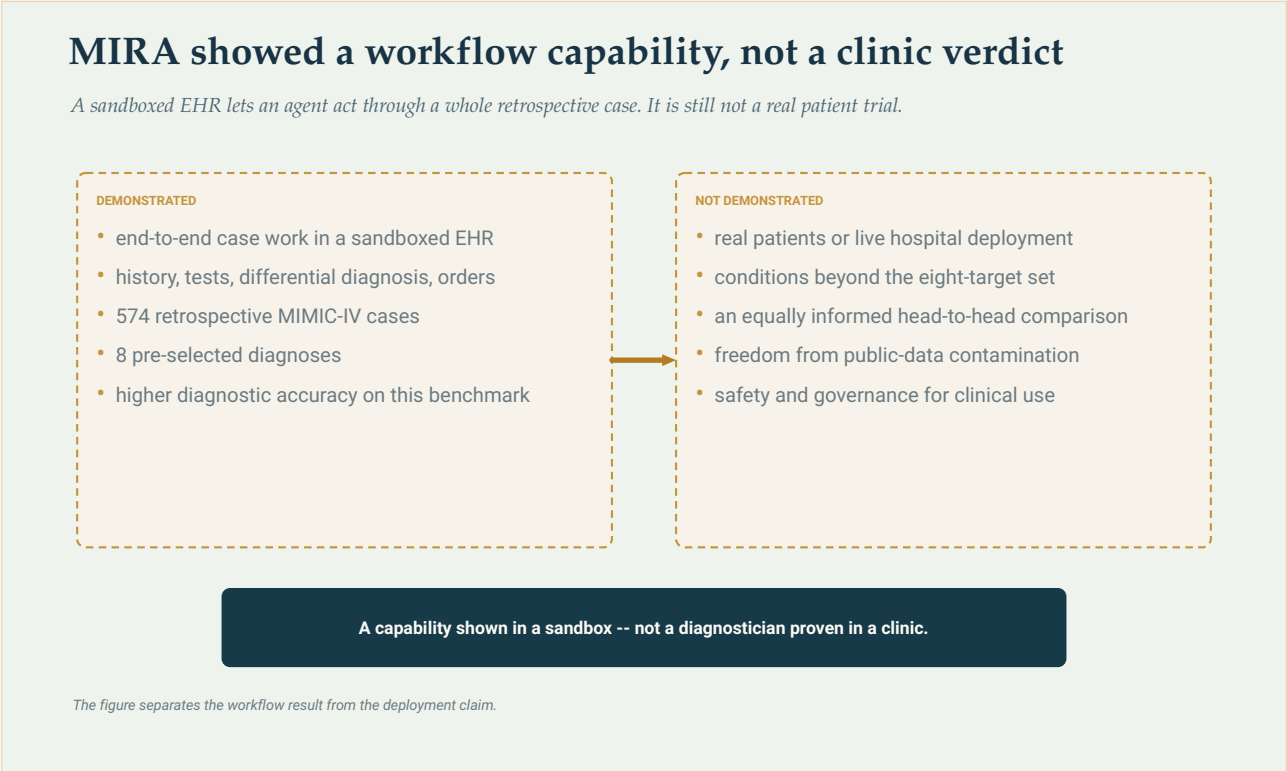
Suala cavab vermək bütün klinik işi aparmaqla eyni deyil

Medical-AI hekayələrinin çoxu *cavab verən* model haqqındadır. Ona vignette və ya exam question verirsiniz, o diaqnoz və ya məsləhət paraqrafı qaytarır və yüksək bal alır. Faydalıdır, amma dardır: chart-a heç vaxt toxunmayan ağıllı consultant.

MIRA başqa bir şey etmək üçün qurulub və əsl xəbər məhz bu fərqdir. Bir suala cavab vermək əvəzinə bütöv case-i işləyir: patient record-u oxuyur, hansı history-nin hələ lazım olduğuna qərar verir, laboratory, imaging və microbiology tests order edir, results-u oxuyur, differential diagnosis-i daraldır və sonra ardınca gələn orders-u yazır — prescriptions, admission, surgery referral. Bunu insanın köçürməsi üçün free text çıxarmaq əvəzinə, clinician kimi electronic health record-un *içində* actions götürərək edir.

Bu real addımdır: məsləhət verən sistemdən workflow boyunca hərəkət edən sistemə. Həm də coverage-də tam olaraq "AI həkimləri keçir"ə düzləşdirilən addımdır. Buna görə nə test edildiyini və harada test edildiyini dəqiq saxlamaq lazımdır.

Bir cümləlik versiya: MIRA bütün case-ləri özbaşına apardı və bu benchmarkda diagnostic accuracy üzrə physicians-ı keçdi — amma bunu sandbox-da, retrospective records üzərində, əvvəlcədən seçilmiş səkkiz diagnosis arasında, yalnız text vasitəsilə etdi və üstünlüyünün böyük hissəsini ən aydın test nəticələri olan conditions-da qazandı. Yenilik realdır. Scoreboard klinika deyil.



MIRA sandboxed EHR daxilində end-to-end workflow qabiliyyətini nümayiş etdirdi. Real-world clinical deployment, geniş diagnostic coverage və doctors-la bərabər məlumatlı, tam ədalətli head-to-head göstərmədi.

Original Aurora diagram — The Clean Paper CC BY 4.0

Müəlliflər nə etdilər

MIRA — Medical Intelligence for Reasoning and Action — OpenAI modelləri üzərində qurulan autonomous agent-dir: danışan və action götürən hissə **GPT-4o**, ayrıca planning addımı isə OpenAI **o1** reasoning model-dən istifadə edir. Bunlar HL7 FHIR — real hospitals-un records ötürmək üçün istifadə etdiyi data standard — üzərində qurulmuş custom, standards-compliant framework daxilində, səksən mindən çox mümkün clinical action toolbox-u ilə işləyir. Sandbox daxilində MIRA patient history-ni çəkə, labs, imaging və microbiology order edib interpret edə, differential diagnosis yarada və treatment plans formalaşdıra bilir — medications prescribe etmək, surgical procedures schedule etmək, admissions planlamaq. Əvvəldən vurğulamağa dəyən detal: MIRA-nın cavablarını avtomatik qiymətləndirən "judges" özləri də GPT-4o idi — test edilən eyni model ailəsi — və peer reviewer bu narahatlığı qaldırdıqdan sonra müəlliflər bunu board-certified physician review-u ilə dəstəklədilər.

Test set **MIMIC-IV**-dən — böyük, ictimai, de-identified EHR database-dən — gəldi. Beth Israel Deaconess Medical Center-də 2008–2019 arasında müalicə olunan təxminən 300,000 patient arasından müəlliflər **səkkiz target diagnosis** üzrə **574 patient case** seçdilər — abdominal pathologies və internal-medicine emergencies, o cümlədən appendicitis, pancreatitis, pneumonia və urinary-tract infection. Hər case üçün MIRA məhdud başlanğıc məlumatı ilə başlayıb addım-addım nə etməli olduğuna qərar verməli idi — real workup-a bənzər struktur, amma real sonluğu artıq record-da olan historical case üzərində.

Sonra performance eyni case-lərdə physicians ilə müqayisə edildi.

"Sandboxed EHR-də autonomous agent" əslində nə deməkdir

Burada üç söz çox yük daşıyır, ona görə açmaq faydalıdır.
Agent o deməkdir ki, sistem prompt-a cavab verən chatbot deyil. Loop işləyir: cari state-ə bax, action seç (bu testi order et, həmin history-ni soruş), result-u gör, növbəti action-u seç — diagnosis və plan-a çatana qədər. "Intelligence" language model-dir; *agency* isə model text-ini icazəli EHR operations-a çevirən və results-u geri verən scaffolding-dir.
Sandboxed EHR live hospital sistemi yox, medical-record sisteminin simulated, divarla ayrılmış surətidir. MIRA test "order" edib həmin patient-ın real həyatda aldığı nəticəni ala bilir, çünki case historical-dır və cavab artıq record-da var. Etdiyi heç nə real patient-a və ya real clinic-ə toxunmur.
Autonomous o deməkdir ki, case-i human-in-the-loop olmadan end-to-end tamamlayır — *sandbox daxilində*. Bu, patients-ı supervision olmadan idarə etmək üçün sərbəst buraxıldığı demək **deyil**. Bunlar çox fərqli iddialardır və yalnız birincisi test edilib.
Sandbox barədə bir detal da vacibdir, çünki onu səhv təsəvvür etmək asandır: MIRA istədiyi testi *order* edə bilər, amma sistem result-u yalnız həmin test patient üçün **real həyatda həqiqətən aparılıbsa** qaytara bilər. Patient aldığı blood panel-i soruşur

> — real values gəlir; heç kimin order etmədiyi scan soruşur — “N/A — could not be performed” gəlir, uydurma rəqəm heç vaxt yox. History də eyni işləyir: record-da MIRA-nın soruşduğu məlumat yoxdursa, simulated patient bilmədiyini deyir. Deməli MIRA heç vaxt götürülməmiş “həlledici” testi sehrlə yarada bilməz — yalnız real workup-da olanlarla işləyir.
> Bu fərq vacibdir, çünki headline sözü — “autonomous” — ən çox ikinci məna kimi yanlış oxunandır.

Nə tapdılar

Benchmarkda **MIRA diagnostic accuracy üzrə physicians-ı keçdi** — amma headline üç ayrı rəqəmi gizlədir və onları qarışdırmaq asandır.

Özbaşına, bütün **574 case** üzrə MIRA düzgün diagnosis-u **88.9%** hallarda dedi — hər patient üçün MIMIC-IV-də record edilmiş discharge diagnosis-a görə score edildi. Humans-la head-to-head üçün doctors bütün 574 case-i işləmədi; MIRA ilə eyni records və tools istifadə edərək ortaq **311-case subset** işlədilər. Eyni 311 case-də MIRA **87.8%** score etdi və ayrıca recruited iki group-la müqayisə edildi. Birincisi **senior group** idi: 7–11 illik təcrübəli dörd board-certified physician, score **78.1%**. İkincisi **mixed-seniority group** idi: əsasən junior residents və iki specialist, real emergency department staffing-ə daha yaxın, score **71.1%**. Eyni cases, eyni tools — sıra AI birinci, seasoned doctors ikinci, junior-heavy team üçüncü oldu. Məqalənin diqqətli ifadəsi MIRA-nın bütün səkkiz disease üzrə physicians-a “consistently equivalent to, and often exceeded” olmasıdır.

Downstream decisions də əsasən dayandı: guideline-concordant, medication-safe və admission baxımından uyğun. Müəlliflər beş safety category üzrə (drug interactions, renal dosing, allergies, QT-risk, opioid prescribing) high-severity medication error olmadığını və 468 case-dən 467-də prescriptions-un correct olduğunu bildirirlər — eyni zamanda sistemin “100% reliability” əldə etmədiyini qeyd edirlər. Səthi oxunuşda bu striking result-dır: agent bütün case-i işləyir və diagnosis-da doctors-dan öndə çıxır.

Amma qələbənin *forması* qələbənin özü qədər vacibdir. Məqaləni oxuyan müstəqil specialists üstünlüyün haradan gəldiyini göstərdilər. Science Media Centre expert panel-ində Dr Wei Xing (University of Sheffield) qeyd etdi ki, AI-nın doctors-u diagnostic accuracy üzrə keçməsi barədə headline rəqəmi **əsasən aydın test results olan conditions** — appendicitis və pancreatitis kimi, decisive scan və ya lab value sualı həll edən hallardan — gəlirdi. Pneumonia və urinary-tract infections üçün — insanların emergency department-a müraciət etdiyi ən yayğın səbəblərdən ikisi — *həm* AI, *həm* doctors ən pis nəticəni verdi və aralarındakı gap ən kiçik idi.

İki tərəfin oyunu arasında asymmetry də vardı və bu məqalənin öz rəqəmlərində görünür. MIRA routine care-da mövcud laboratory analytes-in təxminən **51%-indən**, board-certified physicians isə təxminən **28%-indən** istifadə etdi — case başına median yeddi əlavə blood parameter. Müəlliflər bunu “order everything” strategiyası yox, dataset-in öz routine-care baseline-dan *aşağı* səviyyə kimi təqdim etməkdə diqqətlidirlər və higher-cost cross-

sectional imaging-də systematic artım olmadığını bildirirlər. Yenə də daha çox information öz-özlüyündə higher diagnostic accuracy yarada bilər — buna görə bu, equally-informed players arasında tam like-for-like contest deyil; Xing bunu birbaşa qeyd edib. Cost, discomfort və delay düşünmədən bütün ucuz blood tests-i order etmək azadlığı olan clinician da kağız üzərində daha dəqiq görünərdi.

Niyə “doctors-u keçdi” “daha yaxşı doctor” demək deyil

Benchmark win-i həddən artıq oxumaq asandır. “MIRA higher score aldı” ilə “MIRA daha yaxşıdır” arasında üç şey var.

Birincisi, **nəyə görə score edildi**. Professor Julie Jacko (University of Edinburgh) qeyd etdi ki, MIRA-nın bir neçə əsas outcome-u underlying dataset-də sənədləşdirilənlə müqayisədə təyin olunur — yəni sistem optimal care nümayiş etdirdiyinə görə yox, **recorded clinical behaviour-i təkrarladığına görə** reward alır. Historical chart answer key-ə çevrilir. Benchmark qurmaq üçün ağlabatandır, amma absolute correctness yox, edilənlə agreement ölçür. Ədalət üçün bu problem ən çox *treatment-alignment* metrics-də — MIRA-nın orders-u chart-la match edirdimi? — təsirlidir; headline diagnostic accuracy isə discharge diagnosis-a görə score edilir və documentation echo-dan real outcome-a daha yaxındır.

İkincisi, artıq qeyd olunan **information asymmetry**: daha çox blood tests (available analytes-in təxminən 51%-i vs 28%) daha çox evidence deməkdir. Accuracy gap-in bir hissəsi yəqin reasoning ilə qazanılmır, *satın alınır*.

Üçüncüsü, **harada işlədiyi**. Bu sandbox idi, retrospective cases, səkkiz pre-selected diagnoses, yalnız text. Dr Dominic Oliver (University of Oxford)-ın dediyi kimi, real clinical assessment yalnız patients-ın nə dediyindən yox, necə dediyindən də asılıdır — physical examination, observed behaviour və body language ilə birlikdə; bitmiş record-u oxuyan text-only agent bunların heç birini görmür. Problemi seçilmiş səkkiz diagnosis-dan birinə düşməyən patient isə bu tədqiqatın danışa bildiyi sahədən kənardadır.

Reviewers daha sakit bir narahatlıq da qaldırdılar: **data contamination**. MIMIC-IV public-dır və onun haqqında çox yazılıb; open internet-də trained language model artıq papers, case discussions və ya data-nın özünü görmüş ola bilər. Dr Midhun Parakkal Unni (University of Sheffield) bunu birbaşa qeyd etdi — cavabların bəzisi training data-da idisə, performance-in bir hissəsi clinical reasoning yox, memory-dir və yalnız independent replication ikisini ayıra bilər. Maraqlıdır ki, müəlliflər bunu hand-wave etmirlər: results-un “could be cautiously interpreted as a possible upper bound” və “may overestimate generalization to other public cases” olduğunu yazırlar — öz headline-nə tavan qoyan məqalə.

Bunların heç biri MIRA-nı daha az maraqlı etmir. Sadəcə düzgün oxunuşu calibrate edir: retrospective benchmarkda whole record boyunca action götürən agent clean tests olan diagnoses-da physicians-ı keçdi — qismən daha çox test order etməklə, qismən recorded chart-la agreement sayəsində. Bu real capability demonstration-dır, machines-in artıq doctors-dan daha yaxşı diagnosis qoyduğuna hökm deyil.

Bu nəyi *sübut etmir*

- MIRA-nın real patients-da işlədiyini **göstərmir**. Bütün cases retrospective və simulated idi; heç bir real patient MIRA tərəfindən idarə edilməyib.
- Səkkiz diagnosis-dan kənarda işlədiyini **göstərmir**. Pre-selected set-dən kənardakı conditions — medicine-in qarışıq, undifferentiated çoxluğu — test edilməyib.
- Deploy üçün təhlükəsiz olduğunu **göstərmir**. Müəlliflər özləri generalization, safety və governance üçün prospective, real-world studies lazım olduğunu yazırlar.
- Doctors-la ədalətli head-to-head **müəyyən etmir**. Xeyli daha çox blood tests-dən istifadə etdi (available analytes-in ~51%-i vs 28%) və bir neçə treatment outcome ground-truth best care əvəzinə qismən recorded chart-a görə score edildi.
- Nəticənin memorization-dan azad olduğunu **göstərmir**. Public MIMIC-IV data model training-i ilə overlap edə bilər və bunu independent replication istisna etməlidir.
- "AI doctors-u əvəz edir" **demək deyil**. Record boyunca *action* götürə bilən decision support-dur; reviewers consensus real istifadənin authority-ni saxlayan və text record-un buraxdığı hər şeyi gətirən clinicians-lə partnership olacağıdır.

Sübut nə qədər güclüdür?

İddianı iki yerə ayırmaq lazımdır, çünki hər yarı üçün evidence çox fərqlidir.

Capability demonstration kimi — language-model agent-in EHR sandbox daxilində history, tests, diagnosis və orders-u end-to-end birləşdirərək bütün clinical case-i işləyə bilməsi — iş həqiqətən yenidir və kifayət qədər inandırıcıdır. Yeni olan hissə budur və diqqətə layiq olan da budur.

Superiority claim kimi — MIRA-nın *physicians-dan daha yaxşı olması* — evidence məhduddur və ehtiyatla oxunmalıdır. Konkret retrospective benchmarkda, səkkiz diagnosis-da, information asymmetry (daha çox tests), historical chart-dan gələn answer key və canlı training-data contamination ehtimalı ilə doğrudur. Bu, "agent bu benchmarkda impressively performed" deməyə kifayətdir. "Agent doctor-dan daha yaxşı diagnostician-dır" deməyə kifayət etmir və authors ikinci şeyi iddia etmir.

Ən faydalı mövqe nə "AI doctors-u keçir", nə də "sadəcə oyuncaqdır"dır. Budur: suallara cavab vermək əvəzinə record boyunca *action* götürən yeni medical AI növü çətin retrospective benchmarkda yaxşı işləyib və indi heç sınanmadığı yerdə özünü sübut etməlidir: real, undifferentiated patients üzərində, prospective olaraq, governance altında.

Niyə vacibdir

Son bir neçə ildə medical AI-da maraqlı sual sakitcə dəyişib. Əvvəl "model diagnosis-u düz tapa bilərmi?" idi — modellər daha təmiz test sets-də "bəli" deməyə davam etdi. MIRA daha çətin sualı işarələyir: "model *işi görə bilərmi* — bütün workflow boyunca toplamaq, order etmək, interpret etmək, qərar vermək, action götürmək?" Bu daha faydalı və dürüst sualdır, çünki həm çətinlik, həm risk məhz action-da yaşayır.

Buna görə framing vacibdir. Headline reflex — *AI doctors-u keçir* — nəticənin ən az yeni və ən zəif dəstəklənən hissəsinə işarə edir. Həqiqətən yeni olan daha kiçik və daha nəticəli şeydir: bütün record boyunca hərəkət edən agent. Qabiliyyət dayansa, **workflow**-u ona kimin rəhbərlik etdiyini dəyişdirməzdən xeyli əvvəl dəyişir. Doctor əvəz olunmur; ordering, interpreting, results-u izləmək — workflow — belə tool-un ilk düşdüyü yerdir.

Bu, proof burden-i də doğru istiqamətə qaytarır. Retrospective cases-də leaderboard win prospective trial aparmaq üçün səbəbdir, onun əvəzi deyil. Müəlliflər də bunu deyirlər. MIRA-nın dürüst oxunuşu növbəti tədqiqata dəvətdir — sahənin onsuz da bildiyi standard-la: benchmarkda yox, patients-da ölçülmüş.

Təmiz xülasə

MIRA sandboxed electronic health record daxilində işləyən autonomous AI agent-dir: history götürə, labs, imaging və microbiology order edib interpret edə, diagnosis-a çata və treatment plans yaza bilir. Public MIMIC-IV database-dən səkkiz pre-selected diagnosis üzrə 574 retrospective case-də test ediləndə diagnostic accuracy üzrə physicians-ı keçdi (88.9% overall; head-to-head 87.8% vs 78.1%) və əsasən guideline-concordant, medication-safe decisions verdi. Amma evaluation past records üzərində simulation idi, yalnız text; üstünlüyün çoxu clear-cut test results olan conditions-da gəldi; MIRA doctors-dan xeyli daha çox blood tests-dən istifadə etdi (available analytes-in ~51%-i vs 28%); bir neçə treatment outcome original chart-da record edilənə görə score edildi; public dataset isə real training-data contamination riski yaradır — authors özləri bunu rəqəmlər üçün mümkün upper bound adlandırır. Həqiqi yenilik isolated questions-a cavab verən yox, bütün workflow boyunca action götürən agent-dir. Müəlliflər generalization, safety və governance üçün hələ prospective real-world studies lazım olduğunu açıq deyirlər — düzgün oxunuş budur: impressive capability demonstration, AI-nın doctors-dan daha yaxşı diagnosis qoyduğunun sübutu deyil və real clinic üçün hazır system deyil.

No-BS yoxlaması

Məqalə nəyi göstərir: Language-model agent (MIRA) sandboxed EHR daxilində bütün clinical case-i end-to-end — history, tests, diagnosis, orders — işləyə bilir və səkkiz diagnosis üzrə 574-case retrospective benchmarkda diagnostic accuracy ilə physicians-ı keçib (88.9% overall;

board-certified physicians-a qarşı 87.8% vs 78.1%), high-severity medication error olmadan.

Mümkündür, amma sübut olunmayıb: Bu qabiliyyətin real, undifferentiated patients üçün faydaya çevrilməsi; accuracy edge-in more tests ordered, recorded chart-la agreement və ya memorized public data əvəzinə daha yaxşı reasoning əks etdirməsi.

Nəyi göstərmir: MIRA-nın səkkiz diagnosis-dan kənarda işlədiyini; deploy üçün təhlükəsiz olduğunu; equally-informed ədalətli müqayisədə doctors-u keçdiyini; clinicians-i əvəz etdiyini; nəticələrin training-data contamination-dan azad olduğunu.

Əsas məhdudiyyətlər: Retrospective, simulated, text-only evaluation; səkkiz pre-selected diagnosis; information asymmetry (xeyli daha çox blood tests — available analytes-in ~51%-i vs 28%, low-cost bloods-da, imaging-də yox); treatment outcomes qismən historical chart-a görə score edilib; və language model-in training-də görə biləcəyi public benchmark (MIMIC-IV) — müəlliflər bunun performance üçün mümkün upper bound olduğunu bildirirlər.

Ümumi oxucu nə qədər əmin ola bilər? MIRA-nın real və yeni capability demonstration — yalnız chatbot yox, record boyunca *action* götürən agent — olduğuna yüksək. *Physician-dan daha yaxşı diagnostician* olduğuna aşağı: iddia bir retrospective benchmarkla məhduddur və test-ordering, chart-based scoring və possible contamination ilə confounded-dir. Authors-un öz bottom line-ı doğrudur — patient care üçün məna daşımadan əvvəl prospective, real-world study lazımdır.

Mənbələr

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