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

AI-агент самостојно обработуваше цели случаи на пациенти — во симулатор, врз стари записи

MIRA е нов тип медицинска AI: наместо да одговори на едно прашање, таа обработува цел случај во симулиран болнички запис — зема анамнеза, нарачува и чита тестови, стигнува до дијагноза и ги пишува наредбите. На 574 ретроспективни случаи од јавна база, низ осум однапред избрани дијагнози, авторите пријавуваат дека ја надминала дијагностичката точност на лекарите и донесувала главно одлуки усогласени со насоките и безбедни во однос на лекови. Но секој квалификатор е важен: работела во песочна средина врз стари записи и само преку текст; голем дел од предноста дошол кај состојби со јасни тест-резултати; нарачувала околу двојно повеќе крвни тестови од лекарите; а неколку исходи биле оценети според тоа што било запишано во оригиналното досие. Вистинскиот напредок е агент што дејствува низ целиот работен тек — не доказ дека машина сега дијагностицира подобро од лекар. Самите автори велат дека генерализацијата, безбедноста и управувањето сè уште бараат проспективни студии во реалниот свет.

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

Да одговориш на прашање не е исто што и да водиш цел клинички случај

Повеќето приказни за медицинска AI се за модел што *одговара*. Му давате клиничко сценарио или испитно прашање, тој враќа дијагноза или пасус совет и добива висок резултат. Корисно, но тесно: паметен консултант што никогаш не го допира досието.

MIRA е направена за нешто поинакво, а токму таа разлика е вистинската новост. Наместо да одговори на едно прашање, таа обработува цел случај: го чита досието на пациентот, одлучува кои дополнителни податоци од анамнезата ѝ се потребни, нарачува лабораториски, радиолошки и микробиолошки испитувања, ги чита резултатите, ја стеснува диференцијалната дијагноза и потоа ги пишува наредбите што следуваат — рецепти, прием во болница, упатување на операција. Тоа го прави со преземање дејства *внатре* во електронскиот здравствен запис, како клиничар, наместо само да создава слободен текст што човек треба да го препише.

Тоа е вистински чекор: од систем што советува кон систем што дејствува низ целиот работен тек. Но токму таков чекор лесно се претвора во медиумскиот наслов „AI ги надминува лекарите“. Затоа вреди прецизно да се каже што било тестирано и каде.

Верзијата во една реченица: MIRA самостојно обработувала цели случаи и на овој бенчмарк ги надминала лекарите по дијагностичка точност — но работела во песочна средина, врз ретроспективни записи, само со осум однапред избрани дијагнози, комуницирала само преку текст, а голем дел од предноста го добила кај состојби со најјасни тест-резултати. Напредокот е реален. Табелата со резултати не е клиника.

MIRA showed a workflow capability, not a clinic verdict

A sandboxed EHR lets an agent act through a whole retrospective case. It is still not a real patient trial.

DEMONSTRATED

- end-to-end case work in a sandboxed EHR
- history, tests, differential diagnosis, orders
- 574 retrospective MIMIC-IV cases
- 8 pre-selected diagnoses
- higher diagnostic accuracy on this benchmark

→

NOT DEMONSTRATED

- real patients or live hospital deployment
- conditions beyond the eight-target set
- an equally informed head-to-head comparison
- freedom from public-data contamination
- safety and governance for clinical use

A capability shown in a sandbox -- not a diagnostician proven in a clinic.

The figure separates the workflow result from the deployment claim.

MIRA покажала способност да спроведе цел работен тек од почеток до крај во изолирана EHR-средина. Не покажала реална клиничка примена, широка дијагностичка покриеност ниту целосно праведна директна споредба со лекари што располагаат со еднакви информации.

Original Aurora diagram — The Clean Paper CC BY 4.0

Што направиле авторите

MIRA — Medical Intelligence for Reasoning and Action — е автономен агент изграден врз модели на OpenAI: делот што разговара и дејствува работи со **GPT-4o**, а посебен чекор за планирање го користи резонирачкиот модел **o1** на OpenAI. Тие се поставени во сопствена рамка усогласена со стандардите — изградена врз HL7 FHIR, стандардот за податоци што вистинските болници го користат за размена на записи — со повеќе од осумдесет илјади можни клинички дејства. Во таа песочна средина MIRA може да ја преземе анамнезата на пациентот, да нарача и интерпретира лабораториски, радиолошки и микробиолошки испитувања, да состави диференцијална дијагноза и да формулира терапевтски план — да препише лекови, да закаже хируршка постапка, да планира прием. Важен детал е дека автоматските „судии“ што ги оценувале одговорите на MIRA биле и самите GPT-4o — истото семејство модели што се тестира — а по забелешка од рецензент авторите го дополниле тоа со преглед од лекар со специјалистичка сертификација.

Тест-збирката потекнувала од **MIMIC-IV**, голема јавно достапна деидентификувана база на електронски здравствени записи. Од приближно 300.000 пациенти лекувани во Beth Israel Deaconess Medical Center меѓу 2008 и 2019 година, авторите избрале **574 случаи** што опфаќале **осум целни дијагнози** — абдоминални патологии и итни состојби од интерната медицина, меѓу кои апендицит, панкреатит, пневмонија и инфекција на уринарниот тракт. За секој случај MIRA започнувала со ограничена слика и морала, чекор по чекор, да одлучува што следно да направи — по форма слично на вистинска дијагностичка обработка, но врз запис чии вистински исходи веќе се познати.

Потоа нејзините резултати биле споредени со резултатите на лекари на истите случаи.

> **Што навистина значи „автономен агент во изолирана EHR-средина“**
>
> Трите зборови носат многу значење, па вреди да ги раздвоиме.
> **Агент** значи дека системот не е чатбот што одговара на еден промпт. Тој работи во циклус: ја разгледува моменталната состојба, избира дејство (нарачај го овој тест, прашај ја оваа анамнеза), го добива резултатот и избира следно дејство — сè додека не стигне до дијагноза и план. „Интелигенцијата“ е јазичен модел; *агентноста* е рамката околу него што го претвора текстот на моделот во дозволени EHR-операции и му ги враќа резултатите.
> **Изолирана EHR-средина** значи симулирана, одвоена копија на систем за

> медицински записи, а не активен болнички систем. MIRA може да „нарача“ тест и да го добие резултатот што пациентот навистина го имал, бидејќи случајот е историски и резултатот веќе постои во записот. Ништо што прави не допира вистински пациент или вистинска клиника.
> **Автономен** значи дека го завршува случајот од почеток до крај без човек во циклусот — *внатре во песочната средина*. Тоа **не** значи дека била пуштена сама да лекува пациенти. Тоа се многу различни тврдења, а тестирано е само првото.
> Уште еден важен детал за песочната средина: MIRA може да *нарача* каков било тест, но системот може да врати резултат само ако тој тест **навистина бил направен** кај пациентот во реалниот случај. Ако побара лабораториски панел што пациентот го имал, ги добива реалните вредности; ако побара снимка што никогаш не била направена, добива „N/A — не може да се изврши“, а не измислена бројка. Исто е и со анамнезата: ако во досието ја нема информацијата за која MIRA прашува, симулираниот пациент едноставно вели дека не знае. Значи MIRA не може магично да повика одлучувачки тест што никогаш не бил направен — мора да работи со она што реалната обработка го содржела.
> Разликата е важна затоа што зборот во насловот — „автономен“ — најлесно се чита како второто, посилно тврдење.

Што откриле

На бенчмаркот **MIRA ги надминала лекарите по дијагностичка точност** — но зад таа реченица стојат три различни бројки што лесно се мешаат.

Самостојно, низ сите **574 случаи**, MIRA ја погодила правилната дијагноза во **88,9%** од случаите — оценето според отпусната дијагноза што навистина била заведена за секој пациент во MIMIC-IV. За директната споредба со луѓе, лекарите не ги обработувале сите 574 случаи; работеле на заедничка подгрупа од **311 случаи**, со истите записи и алатки што ги имала MIRA. На тие исти 311 случаи MIRA постигнала **87,8%**, а била споредувана со две одделно регрутирани групи. Првата била **сениорска група**: четворица специјалистички сертифицирани лекари со 7 до 11 години искуство, со резултат **78,1%**. Втората била **група со мешано искуство**: претежно помлади специјализанти плус двајца специјалисти, поблиску до тоа како навистина е екипиран ургентен центар, со резултат **71,1%**. Исти случаи, исти алатки — редоследот бил AI прва, искусните лекари втори, помладата група трета. Внимателната формулација на самиот труд е дека MIRA била „доследно еквивалентна на, а често и подобра од“ лекарите низ сите осум болести.

И подоцнежните одлуки биле добри: главно усогласени со насоките, безбедни во однос на лекови и соодветни за прием. Авторите не пријавиле ниту една медикаментозна грешка со висока сериозност во пет категории на безбедност (интеракции меѓу лекови, дозирање при бубрежна функција, алергии, QT-ризик и пропишување опиоиди), а рецептите биле точни во 467 од 468 случаи — при што нагласуваат дека системот „не

постигнал 100% веродостојност“. Земено буквално, тоа е впечатлив резултат: агент го води целиот случај и излегува пред лекарите по дијагнозата.

Но *обликот* на победата е важен колку и самата победа. Независни експерти што го прочитале трудот посочиле од каде главно доаѓа предноста. На експертскиот панел на Science Media Centre, д-р Wei Xing (University of Sheffield) забележал дека главната бројка — AI подобра од лекарите по дијагностичка точност — била **пред сè поттикната од состојби со јасни тест-резултати**, како апендицит и панкреатит, каде одлучувачка снимка или лабораториска вредност го решава прашањето. Кај пневмонија и инфекции на уринарниот тракт — две од најчестите причини луѓето навистина да се појават во ургентен центар — *и AI и лекарите* имале најслаби резултати, а разликата меѓу нив била најмала.

Постоела и асиметрија во количината информации со кои работеле двете страни, видлива во бројките на трудот. MIRA се потпирала повеќе на лабораторијата: користела околу **51%** од лабораториските аналити достапни во рутинската грижа, наспроти приближно **28%** кај сертифицираните лекари — медијана од седум дополнителни крвни параметри по случај. Авторите внимателно истакнуваат дека тоа е *под* вообичаената практика во самата база и не е стратегија „нарачај сè“, а не нашле ниту систематско зголемување на поскапите пресечни снимања. Сепак, повеќе информации сами по себе можат да ја зголемат дијагностичката точност — па ова не е сосема споредба на две еднакво информирани страни, што Xing директно го нагласил. И лекар кому би му било дозволено да нарача секој евтин крвен тест без да ги зема предвид трошоците, непријатноста или доцнењето би изгледал подобро на хартија.

Зошто „ги надмина лекарите“ не значи „е подобар лекар“

Победа на бенчмарк лесно се чита предалеку. Меѓу „MIRA постигна повисок резултат“ и „MIRA е подобра“ стојат три важни работи.

Прво, **според што била оценета.** Професорката Julie Jacko (University of Edinburgh) забележала дека неколку клучни исходи за MIRA се дефинирани во однос на она што било заведено во основната база — што значи дека системот се наградува за **репродуцирање на документираното клиничко однесување**, а не нужно за покажување оптимална грижа. Историското досие станува клуч со одговори. Тоа е разумен начин да се изгради бенчмарк, но мери согласност со она што било направено, не апсолутна исправност. За волја на правичноста, ова најсилно ги засега метриките за *усогласеност на третманот* — дали наредбите на MIRA се совпаѓале со досието? — додека главната дијагностичка точност се мери според отпусната дијагноза, што е поблиску до реален исход отколку едноставно повторување на документацијата.

Второ, **веќе споменатата нерамнотежа во информациите**: многу повеќе крвни тестови (околу 51% од достапните аналити наспроти 28%) значат повеќе докази. Дел од јазот во точноста веројатно е *купен* со повеќе податоци, а не добиен само со подобро

резонирање.

Трето, **каде работела.** Ова била песочна средина, врз ретроспективни случаи, со осум однапред избрани дијагнози и само преку текст. Реалната клиничка проценка, како што истакнал д-р Dominic Oliver (University of Oxford), зависи не само од тоа што пациентите го кажуваат, туку и како го кажуваат — заедно со физичкиот преглед, набљудуваното однесување и говорот на телото, од кои ништо не гледа текстуален агент што чита завршено досие. А пациент чиј проблем не спаѓа во една од осумте избрани дијагнози е едноставно надвор од она за што оваа студија може да зборува.

Рецензентите покренале и потивка грижа: **контаминација со податоци.** MIMIC-IV е јавна и многу анализирана база, па јазичен модел обучуван на отворен интернет можеби веќе видел трудови, дискусии на случаи или самите податоци. Д-р Midhun Parakkal Unni (University of Sheffield) директно го посочил ова — ако дел од одговорите биле во тренинг-податоците, дел од резултатот е меморија, не клиничко резонирање, а само независна репликација може да ги раздели. Значајно е што авторите не ја игнорираат можноста: пишуваат дека нивните резултати „може внимателно да се толкуваат како можна горна граница“ и дека „може да ја преценуваат генерализацијата кон други јавни случаи“ — труд што самиот поставува плафон врз сопствениот наслов.

Ништо од ова не ја прави MIRA помалку интересна. Само бара правилно калибрирано читање: на ретроспективен бенчмарк, агент што може да дејствува низ целото досие ги надминал лекарите кај дијагнози со чисти тестови — делумно затоа што нарачувал повеќе тестови, делумно затоа што се согласувал со запишаното во досието. Тоа е вистинска демонстрација на способност, не пресуда дека машините сега дијагностицираат подобро од лекарите.

Што ова не докажува

- **Не** покажува дека MIRA работи на вистински пациенти. Секој случај бил ретроспективен и симулиран; ниту еден пациент не бил лекуван од неа.
- **Не** покажува дека работи надвор од осумте дијагнози. Состојбите надвор од однапред избраниот сет — неуредното, недиференцирано мнозинство од медицината — не биле тестирани.
- **Не** покажува дека е безбедна за примена. Самите автори пишуваат дека генерализацијата, безбедноста и управувањето мора да се проверуваат во проспективни студии во реалниот свет.
- **Не** утврдува праведна директна споредба со лекари. MIRA користела многу повеќе крвни тестови (околу 51% од достапните аналити наспроти 28%), а неколку терапевтски исходи биле оценети делумно според запишаното во историското досие, не според независно утврдена најдобра грижа.
- **Не** покажува дека резултатот е ослободен од меморирање. Јавните податоци MIMIC-IV можеби се преклопуваат со тренинг-податоците на моделот, што може да го

исклучи само независна репликација.
- **Не** значи „AI ги заменува лекарите“. Ова е поддршка за одлучување што може да *дејствува* низ досие; консензусот на рецензентите е дека реалната употреба би била во партнерство со клиничари, кои ја задржуваат одговорноста и го додаваат сето она што текстуалниот запис го нема.

Колку се убедливи доказите?

Тврдењето треба да се подели на два дела, затоа што доказите за нив се многу различни.

Како демонстрација на способност — дека агент базиран на јазичен модел може да води цел клинички случај во EHR-песочна средина, поврзувајќи анамнеза, тестови, дијагноза и наредби од почеток до крај — трудот е навистина нов и прилично убедлив. Тоа е новиот дел и токму на него вреди да се обрне внимание.

Како тврдење за надмоќ — дека MIRA е *подобра од лекарите* — доказите се ограничени и мора внимателно да се читаат. Резултатот важи на конкретен ретроспективен бенчмарк, со осум дијагнози, нерамнотежа во информациите (повеќе тестови), клуч со одговори изведен од историското досие и реална можност за контаминација на тренинг-податоците. Тоа е доволно за да се каже „агентот се покажа впечатливо на овој бенчмарк“. Не е доволно за да се каже „агентот е подобар дијагностичар од лекар“, а авторите не го тврдат второто.

Најкорисната позиција не е ниту „AI ги победи лекарите“ ниту „ова е само играчка“. Таа е: нов тип медицинска AI — што *дејствува* низ досието наместо само да одговара на прашања — се покажал добро на тежок ретроспективен бенчмарк и сега мора да се докаже таму каде што сè уште не бил испробан: на вистински, недиференцирани пациенти, проспективно и под соодветно управување.

Зошто е важно

Во последниве години интересното прашање во медицинската AI тивко се смени. Порано беше „може ли моделот да ја погоди дијагнозата?“ — а моделите сè почесто одговараа да, на сè почисти тест-збирки. MIRA го означува преминот кон потешко прашање: „може ли моделот *да ја врши работата* — да собира податоци, да нарачува, да толкува, да одлучува, да дејствува — низ цел работен тек?“ Тоа е покорисно и почесно прашање, затоа што токму во дејствувањето живеат и тешкотијата и ризикот.

Затоа е важно како се формулира резултатот. Рефлексниот наслов — *AI ги надминува лекарите* — покажува кон најмалку новиот и најслабо поткрепениот дел од трудот. Вистински новото е потесно, но позначајно: агент што може да се движи низ целото досие. Ако се потврди, таа способност го менува **работниот тек** многу пред да смени

кој е надлежен за него. Лекарот не исчезнува; нарачувањето, интерпретацијата и следењето на резултатите — работниот тек — е местото каде ваква алатка најпрво би влегла.

И тоа го поставува товарот на докажување во правилна насока. Победа на табела со ретроспективни случаи е причина да се спроведе проспективно испитување, не замена за него. Авторите го кажуваат истото. Најчесното читање на MIRA е покана за следната студија — според стандардот што медицината веќе го познава: мерење на пациенти, не само на бенчмаркови.

Кратко резиме

MIRA е автономен AI-агент што работи во симулиран електронски здравствен запис: може да земе анамнеза, да нарача и толкува лабораторија, снимања и микробиологија, да стигне до дијагноза и да напише терапевтски план. Тестирана на 574 ретроспективни случаи од јавната база MIMIC-IV, низ осум однапред избрани дијагнози, таа ги надминала лекарите по дијагностичка точност (88,9% вкупно; 87,8% наспроти 78,1% во директната споредба) и донесувала претежно одлуки усогласени со насоките и безбедни во однос на лекови. Но проценката била симулација на стари записи и само преку текст; голем дел од предноста бил кај состојби со јасни тест-резултати; MIRA користела многу повеќе крвни тестови од лекарите (околу 51% од достапните аналити наспроти 28%); неколку терапевтски исходи биле оценети според тоа што било запишано во оригиналното досие; а јавната база отвора реален ризик од контаминација на тренинг-податоците — што самите автори го опишуваат како можна горна граница на нивните бројки. Вистинскиот напредок е агент што дејствува низ целиот работен тек наместо да одговара на изолирани прашања. Авторите експлицитно велат дека генерализацијата, безбедноста и управувањето сè уште бараат проспективни студии во реалниот свет — што е правилното читање: впечатлива демонстрација на способност, не доказ дека AI дијагностицира подобро од лекар и не систем подготвен за реална клиника.

Проверка без претерување

Што покажува трудот: Агент базиран на јазичен модел (MIRA) може да води цел клинички случај од почеток до крај во изолирана EHR-средина — анамнеза, тестови, дијагноза, наредби — и на ретроспективен бенчмарк со 574 случаи и осум дијагнози ги надминал лекарите по дијагностичка точност (88,9% вкупно; 87,8% наспроти 78,1% кај специјалистички сертифицирани лекари), без медикаментозни грешки со висока сериозност.

Што е веројатно, но не е докажано: Дека оваа способност ќе донесе корист за вистински, недиференцирани пациенти; и дека предноста во точноста се должи

на подобро резонирање, а не на повеќе нарачани тестови, согласност со историското досие или меморирани јавни податоци.

Што не покажува: Дека MIRA работи надвор од осумте дијагнози; дека е безбедна за примена; дека ги надминува лекарите во праведна споредба со еднакви информации; дека ги заменува клиничарите; или дека резултатите се без контаминација од тренинг-податоци.

Главни ограничувања: Ретроспективна, симулирана и текстуална проценка; осум однапред избрани дијагнози; нерамнотежа во информациите (многу повеќе крвни тестови — околу 51% од достапните аналити наспроти 28%, главно евтини лабораториски тестови, не снимања); терапевтски исходи делумно оценети според историското досие; и јавен бенчмарк (MIMIC-IV) што јазичниот модел можеби го видел за време на обуката — што авторите го означуваат како можна горна граница на резултатот.

Колкава доверба треба да има општ читател? Висока дека MIRA е реална и нова демонстрација на способност — агент што *дејствува* низ досието, а не само чатбот. Ниска дека е *подобар дијагностичар од лекар*: тоа тврдење е ограничено на еден ретроспективен бенчмарк и е заматено од различното нарачување тестови, оценувањето според досието и можна контаминација. Заклучокот на самите автори е правилниот — потребни се проспективни студии во реалниот свет пред ова да значи нешто за грижата за пациенти.

Извори

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