ИИ агент обработи цели пациентски случаи самостоятелно — в симулатор, върху стари записи

MIRA е нов тип медицински ИИ: вместо да отговаря на един въпрос, той обработва целия случай в симулирано болнично досие — снема анамнеза, назначава и разчита изследвания, стига до диагноза и оформя назначенията. При 574 ретроспективни случая от публична база данни, обхващащи осем предварително избрани диагнози, авторите съобщават, че той е постигнал по-висока диагностична точност от лекарите и е вземал решения, които до голяма степен съответстват на клиничните препоръки и са били безопасни по отношение на лекарствата. Всяко уточнение в това изречение има значение: системата е работила в sandbox върху стари записи, само с текст; голяма част от преднината ѝ е дошла при състояния с ясни резултати от изследвания; назначавала е около два пъти повече кръвни показатели от лекарите; а няколко изхода са били оценявани спрямо това, което е било записано в оригиналното досие. Истинският напредък е агент, който действа по целия работен процес — не доказателство, че машина вече диагностицира по-добре от лекар. Самите автори го казват: обобщимостта, безопасността и управлението все още изискват проспективни проучвания в реална среда.

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

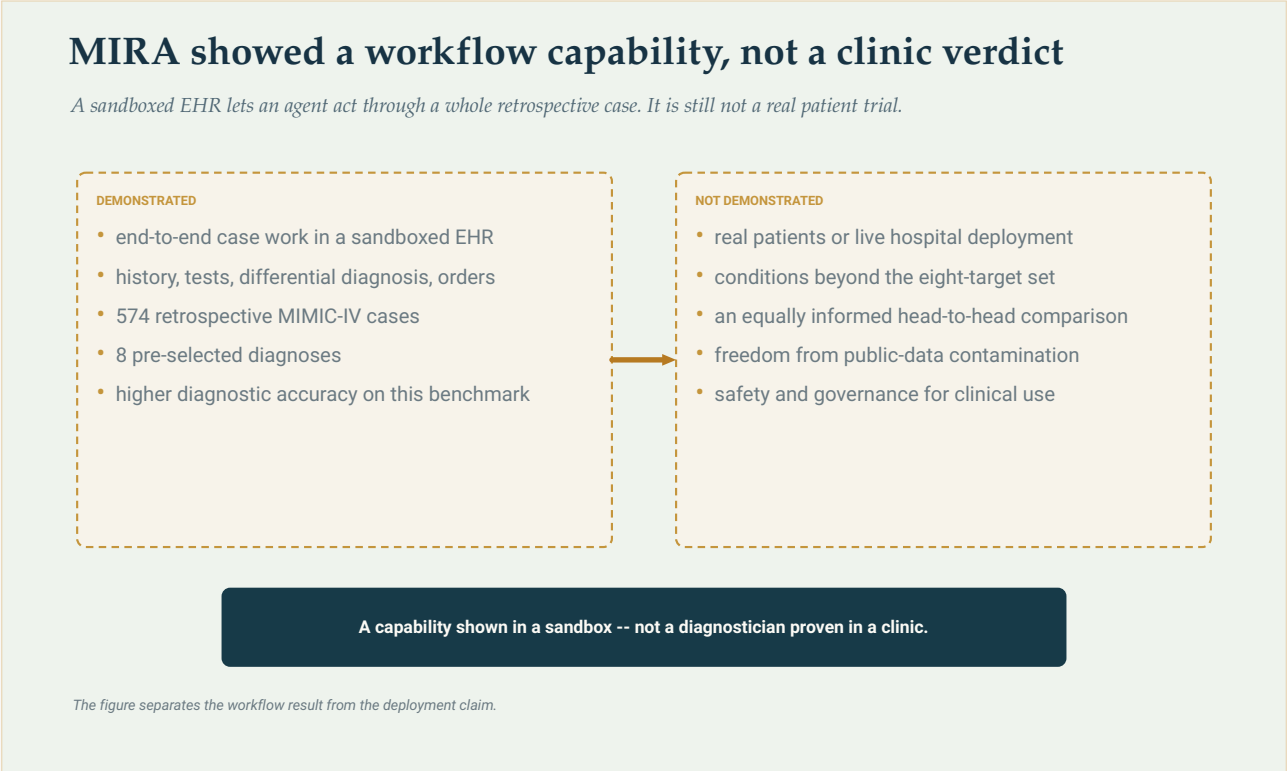
Да отговориш на въпрос не е същото като да водиш целия случай

Повечето истории за медицински ИИ са за модел, който *отговаря*. Давате му клиничен казус или изпитен въпрос, той връща диагноза или абзац със съвет и получава висок резултат. Полезно, но тясно: умен консултант, който никога не докосва досието.

MIRA е създаден да прави нещо различно и именно тази разлика е истинската новина. Вместо да отговаря на един въпрос, той обработва целия случай: прочита досието на пациента, решава каква анамнеза още му е нужна, назначава лабораторни, образни и микробиологични изследвания, прочита резултатите, стеснява диференциалната диагноза и след това оформя следващите назначения — рецепти, прием в болница, насочване за операция. Прави това чрез действия *вътре* в електронното здравно досие, както би действал клиницист, а не като генерира свободен текст, който човек после да преписва.

Това е реална стъпка: от система, която съветва, към система, която действа по целия работен процес. Но е и точно онзи вид стъпка, която лесно се сплесква в медийното отразяване до „ИИ превъзхожда лекарите“. Затова си струва да сме точни какво е било изпитано и къде.

Версията в едно изречение: MIRA е обработвал цели случаи самостоятелно и в този тест за оценяване е постигнал по-висока диагностична точност от лекарите — но го е направил в изолирана тестова среда, върху ретроспективни записи, при осем предварително избрани диагнози, общувайки само чрез текст, а голяма част от преднината му е дошла при състоянията с най-ясни резултати от изследвания. Напредъкът е реален. Класацията не е клиниката.



MIRA демонстрира способност да изпълнява цялостен работен процес от начало до край в изолирана EHR среда. Проучването не демонстрира реално клинично внедряване, широко диагностично покритие или справедливо сравнение с лекари при еднакво количество информация.

Original Aurora diagram — The Clean Paper CC BY 4.0

Какво са направили авторите

MIRA — Medical Intelligence for Reasoning and Action — е автономен агент, изграден върху модели на OpenAI: частта, която разговаря и действа, работи с **GPT-4o**, а отделна стъпка за планиране използва модела за разсъждение **o1** на OpenAI. Те са поставени в собствена, съвместима със стандартите рамка — базирана на HL7 FHIR, стандарта за данни, чрез който реалните болници обменят записи — с набор от повече от осемдесет хиляди възможни клинични действия. В тази тестова среда MIRA може да извлича анамнеза, да назначава и интерпретира лабораторни, образни и микробиологични изследвания, да съставя диференциална диагноза и да формулира план за лечение — да предписва лекарства, да планира хирургични процедури и болничен прием. Един детайл заслужава да бъде отбелязан още в началото: автоматизираните „съдии“, които са оценявали отговорите на MIRA, също са били GPT-4o — същото семейство модели, което се тества — а след като рецензент е повдигнал този проблем, авторите са добавили преглед от сертифициран лекар специалист.

Тестовият набор идва от **MIMIC-IV**, голяма публично достъпна база с деидентифицирани електронни здравни досиета. От приблизително 300 000 пациенти, лекувани в Beth Israel Deaconess Medical Center между 2008 и 2019 г., авторите са избрали **574**

пациентски случая, обхващащи **осем целеви диагнози** — коремни патологии и спешни вътрешни заболявания, сред които апендицит, панкреатит, пневмония и инфекция на пикочните пътища. За всеки случай MIRA е започвал с ограничена информация и е трябвало стъпка по стъпка да решава какво да направи след това — структура, подобна на реално диагностично уточняване, но върху досие, чийто истински край вече е известен.

След това представянето му е сравнено с това на лекари по същите случаи.

Какво всъщност означава „автономен агент в изолирана EHR среда“

Три думи тук носят много смисъл, затова си струва да ги разглобим.
Агент означава, че системата не е чатбот, който просто отговаря на инструкция. Тя изпълнява цикъл: разглежда текущото състояние, избира действие (назначи това изследване, попитай за онази част от анамнезата), вижда резултата, избира следващото действие — докато стигне до диагноза и план. „Интелигентността“ е езиков модел; *агентността* идва от инфраструктурата около него, която превръща текста на модела в разрешени операции в EHR и връща резултатите обратно към него.
Изолирана EHR среда означава симулирано, изолирано копие на система за медицински досиета, а не жива болнична система. MIRA може да „назначи“ изследване и да получи резултата, който пациентът действително е имал, защото случаят е исторически и отговорът вече е записан. Нищо, което системата прави, не засяга реален пациент или реална клиника.
Автономен означава, че той завършва случая от начало до край без човек в цикъла — *вътре в тестовата среда*. Това **не** означава, че е бил пуснат да управлява пациенти без надзор. Това са много различни твърдения и е изпитано само първото.
Още нещо за тестовата среда, защото лесно може да си го представим погрешно: MIRA може да *назначи* каквото изследване пожелае, но системата може да върне резултат само ако това изследване е било **действително извършено** на този пациент в реалния живот. Ако поиска кръвен панел, който пациентът е имал, получава реалните стойности; ако поиска скенер, който никой не е назначавал, получава „N/A — не може да бъде извършено“, никога измислено число. С анамнезата е същото: ако досието не съдържа информацията, за която MIRA пита, симулираният пациент просто казва, че не знае. Така MIRA не може да извика от нищото решаващо изследване, което никога не е било направено — може да работи само с това, което реалното диагностично уточняване е включвало.
Разликата е важна, защото заглавната дума — „автономен“ — е именно тази, която най-лесно се прочита във втория смисъл.

Какво са установили

В теста за оценяване **MIRA е постигнал по-висока диагностична точност от лекарите** — но заглавието скрива три различни числа, които лесно могат да се смесят, затова е добре да ги разделим.

Самостоятелно, върху всички **574 случая**, MIRA е посочил правилната диагноза в **88,9%** от случаите — като оценката е спрямо диагнозата при изписване, действително записана за всеки пациент в MIMIC-IV. При директното сравнение с хора лекарите не са работили по всичките 574 случая; те са работили върху общ **поднабор от 311 случая**, използвайки същите записи и инструменти, с които е разполагал MIRA. На тези същите 311 случая MIRA е постигнал **87,8%** и е бил сравнен с две отделно набрани групи. Първата е **група от опитни лекари**: четирима сертифицирани специалисти със 7 до 11 години опит, които са постигнали **78,1%**. Втората е **група със смесена старшинност**: предимно млади специализанти плюс двама специалисти — по-близо до начина, по който реално е окомплектовано едно спешно отделение — и тя е постигнала **71,1%**. Същите случаи, същите инструменти — а подреждането е било ИИ първи, опитните лекари втори, по-младият екип трети. Внимателната формулировка на самата статия е, че MIRA е бил „последователно еквивалентен на лекарите и често ги е превъзхождал“ при всичките осем заболявания.

Решенията му след диагнозата също са издържали: до голяма степен съответстващи на клиничните препоръки, безопасни по отношение на лекарствата и подходящи при решение за прием. Авторите съобщават за нито една лекарствена грешка с висока тежест в пет категории за безопасност (лекарствени взаимодействия, дозиране при бъбречна функция, алергии, QT-риск и предписване на опиоиди) и за правилни предписания в 467 от 468 случая — като едновременно отбелязват, че системата „не е постигнала 100% надеждност“. Взето буквално, това е впечатляващ резултат: агент води целия случай и излиза пред лекарите по диагнозата.

Но *формата* на тази победа е също толкова важна, колкото самата победа. Независими специалисти, които са прочели статията, посочват откъде всъщност идва преднината. В експертния панел на Science Media Centre д-р Wei Xing (University of Sheffield) отбелязва, че заглавното число — ИИ да превъзхожда лекарите по диагностична точност — е **до голяма степен движено от състоянията с ясни резултати от изследвания**, като апендицит и панкреатит, при които решаващ образен или лабораторен резултат установява диагнозата. При пневмония и инфекции на пикочните пътища — две от най-честите причини хората действително да стигат до спешно отделение — *и* ИИ, *и* лекарите са се представили най-зле, а разликата между тях е била най-малка.

Има и асиметрия в начина, по който двете страни са „играли“, и тя личи от собствените числа на статията. MIRA е разчитал повече на лабораторията: използвал е около **51%** от лабораторните аналити, достъпни при рутинни грижи, срещу приблизително **28%** за сертифицираните лекари — медианно със седем повече кръвни показателя на случай. Авторите внимателно подчертават, че това е *под* рутинното изходно ниво в самия набор данни, а не стратегия „назначи всичко“, и не намират *систематично*

увеличение на по-скъпите напречни образни изследвания. Все пак повече информация сама по себе си може да повиши диагностичната точност — така че това не е напълно равнопоставено състезание между еднакво информирани участници, което Xing изрично отбелязва. Лекар, който може свободно да назначава всеки евтин кръвен тест, без да отчита цена, дискомфорт или забавяне, също би изглеждал по-точен на хартия.

Защо „победи лекарите“ не означава „по-добър лекар“

Лесно е една победа в тест за оценяване да бъде прочетена прекалено широко. Между „MIRA получи по-висок резултат“ и „MIRA е по-добър“ стоят три неща.

Първо, **спрямо какво е оценяван.** Проф. Julie Jacko (University of Edinburgh) отбелязва, че няколко от ключовите резултати на MIRA са дефинирани спрямо това, което е било документирано в изходния набор данни — тоест системата е възнаграждавана за **възпроизвеждане на записаното клинично поведение**, а не непременно за демонстриране на оптимална грижа. Историческото досие се превръща в ключ с верните отговори. Това е разумен начин да се построи тест за оценяване, но измерва съгласие с направеното, а не правилност в някакъв абсолютен смисъл. За да сме справедливи към статията, този проблем засяга най-силно метриките за *съответствие на лечението* — съвпадат ли назначенията на MIRA с досието? — докато заглавната диагностична точност се оценява спрямо диагнозата при изписване, което е по-близо до реален изход, отколкото до ехо на документацията.

Второ, **асиметрията на информацията**, спомената по-горе: много повече кръвни изследвания (около 51% от наличните аналити срещу 28%) означават повече доказателства. Част от разликата в точността вероятно е *купена* с информация, а не получена чрез по-добро разсъждение.

Трето, **къде е работила системата.** Това е изолирана тестова среда, върху ретроспективни случаи, при осем предварително избрани диагнози, само с текст. Реалната клинична оценка, както посочва д-р Dominic Oliver (University of Oxford), зависи не само от това какво казват пациентите, а и от начина, по който го казват — плюс физикалния преглед, наблюдаваното поведение и езика на тялото, които текстов агент, четящ вече завършено досие, изобщо не вижда. А пациент, чийто проблем не попада в една от осемте избрани диагнози, просто е извън това, за което това проучване може да говори.

Има и по-тихо притеснение, повдигнато от рецензентите: **замърсяване на данните**. MIMIC-IV е публичен и е широко обсъждан, така че езиков модел, обучаван върху отворения интернет, може вече да е виждал статии, обсъждания на случаи или самите данни. Д-р Midhun Parakkal Unni (University of Sheffield) посочва това директно — ако част от отговорите са присъствали в обучителните данни, част от представянето може да е памет, а не клинично разсъждение, и само независимо възпроизвеждане може да раздели двете. Забележително е, че авторите не подминават проблема: те

пишат, че резултатите им „могат предпазливо да се интерпретират като възможна горна граница“ и „може да надценяват обобщимостта към други публични случаи“ — статия, която сама поставя таван на собственото си заглавие.

Нищо от това не прави MIRA по-малко интересен. То просто налага калибриран прочит: на ретроспективен тест за оценяване агент, способен да действа по цялото досие, е превъзхождал лекарите при диагнози с ясни изследвания — отчасти като е назначавал повече тестове, отчасти като е съвпадал със записаното в досието. Това е реална демонстрация на възможност, а не присъда, че машините вече диагностицират по-добре от лекарите.

Какво *не* доказва това

- То **не** показва, че MIRA работи при реални пациенти. Всеки случай е бил ретроспективен и симулиран; нито един пациент не е бил управляван от системата.
- То **не** показва, че системата работи извън осемте диагнози. Състоянията извън предварително избрания набор — тоест по-голямата, неясна и неразграничена част от медицината — не са тествани.
- То **не** показва, че системата е безопасна за внедряване. Самите автори пишат, че обобщимостта, безопасността и управлението все още изискват проспективни проучвания в реална среда.
- То **не** установява справедливо директно сравнение с лекарите. MIRA е използвал много повече кръвни изследвания (около 51% от наличните аналити срещу 28%), а няколко изхода от лечението са оценявани отчасти спрямо записаното в досието, а не спрямо обективно най-добрата грижа.
- То **не** показва, че резултатът е свободен от запаметяване. Публичните данни MIMIC-IV може да се припокриват с обучението на модела и това трябва да бъде изключено чрез независимо възпроизвеждане.
- То **не** означава „ИИ заменя лекарите“. Това е подпомагане на решенията, което може да *действа* по досието; консенсусът на експертите е, че реалната употреба ще бъде в партньорство с клиницисти, които запазват правомощията и добавят всичко, което текстовото досие не съдържа.

Колко силни са доказателствата?

Твърдението трябва да се раздели на две, защото доказателствата за двете половини са много различни.

Като демонстрация на възможност — че агент с езиков модел може да води цял клиничен случай в изолирана EHR среда, свързвайки анамнеза, изследвания, диагноза и назначения от начало до край — работата е действително нова и сравнително

убедителна. Това е новата част и именно тя заслужава внимание.

Като твърдение за превъзходство — че MIRA е *по-добър от лекарите* — доказателствата са ограничени и трябва да се четат внимателно. Те важат за конкретен ретроспективен тест за оценяване, при осем диагнози, с информационна асиметрия (повече изследвания), с ключ за отговорите, извлечен от историческото досие, и с реална възможност за замърсяване на обучителните данни. Това е достатъчно, за да кажем „агентът се представи впечатляващо в този тест за оценяване“. Не е достатъчно, за да кажем „агентът е по-добър диагностик от лекар“, а авторите и не твърдят второто.

Най-полезната позиция не е нито „ИИ победи лекарите“, нито „това е просто играчка“. Тя е: нов тип медицински ИИ — такъв, който *действа* по досието вместо само да отговаря на въпроси — се е представил добре на труден ретроспективен тест за оценяване и сега трябва да се докаже там, където още не е изпробван: върху реални, неразграничени пациенти, проспективно и под управление.

Защо това има значение

През последните няколко години интересният въпрос в медицинския ИИ тихо се промени. Преди беше „може ли моделът да постави правилната диагноза?“ — и моделите все по-често отговаряха „да“ на все по-чисти тестови набори. MIRA бележи прехода към по-труден въпрос: „може ли моделът *да свърши работата* — да събере информация, да назначи, да интерпретира, да реши, да действа — по целия работен процес?“ Това е по-полезен и по-честен въпрос, защото именно в действието живеят и трудността, и рискът.

Затова рамката е важна. Заглавният рефлекс — *ИИ превъзхожда лекарите* — сочи към най-малко новата и най-слабо подкрепената част от резултата. Истински новото е по-малко, но по-значимо: агент, който може да премине през цялото досие. Ако тази способност издържи, тя променя **работния процес** много преди да промени кой го ръководи. Лекарят не е заменен; назначаването, интерпретацията, проследяването на резултатите — именно там първо се намесва инструмент като този.

И това пренарежда тежестта на доказване в правилната посока. Победа в класация върху ретроспективни случаи е причина да се проведе проспективно изпитване, а не заместител на такова. Самите автори казват същото. Честният прочит на MIRA е покана за следващото проучване — според стандарт, който областта вече познава: измерване при пациенти, а не върху еталонни тестове.

Накратко

MIRA е автономен ИИ агент, който работи в изолирана среда на електронно здравно досие: може да снема анамнеза, да назначава и интерпретира лабораторни, образни и микробиологични изследвания, да стига до диагноза и да изготвя планове за лечение. Тестван върху 574 ретроспективни случая от публичната база MIMIC-IV, при осем предварително избрани диагнози, той е постигнал по-висока диагностична точност от лекарите (88,9% общо; 87,8% срещу 78,1% в директното сравнение) и е вземал решения, до голяма степен съответстващи на клиничните препоръки и безопасни по отношение на лекарствата. Но оценката е била симулация върху минали записи, само с текст; голяма част от преднината е дошла при състояния с ясни резултати от изследвания; MIRA е използвал много повече кръвни показатели от лекарите (около 51% от наличните аналити срещу 28%); няколко изхода от лечението са оценявани спрямо това, което е било записано в оригиналното досие; а публичният набор данни поражда реален риск от замърсяване на обучителните данни — което самите автори наричат възможна горна граница на числата си. Истинският напредък е агент, който действа по целия работен процес, вместо да отговаря на изолирани въпроси. Авторите изрично подчертават, че обобщимостта, безопасността и управлението все още изискват проспективни проучвания в реална среда — и това е правилният прочит: впечатляваща демонстрация на възможност, не доказателство, че ИИ диагностицира по-добре от лекар, и не система, готова за реална клиника.

Проверка без украса

Какво показва статията: Агент с езиков модел (MIRA) може да води цял клиничен случай от начало до край в изолирана EHR среда — анамнеза, изследвания, диагноза, назначения — и върху ретроспективен тест за оценяване от 574 случая при осем диагнози е постигнал по-висока диагностична точност от лекарите (88,9% общо; 87,8% срещу 78,1% за сертифицираните лекари), без лекарствени грешки с висока тежест.

Какво е правдоподобно, но недоказано: Че тази способност носи полза за реални, неразграничени пациенти; че преднината в точността отразява по-добро разсъждение, а не повече назначени изследвания, съвпадение със записаното досие или запаметени публични данни.

Какво не показва: Че MIRA работи извън осемте диагнози; че е безопасен за внедряване; че побеждава лекарите при справедливо сравнение с еднакво количество информация; че заменя клиницистите; че резултатите са свободни от замърсяване на обучителните данни.

Основни ограничения: Ретроспективна, симулирана, само текстова оценка; осем предварително избрани диагнози; информационна асиметрия (много повече кръвни изследвания — около 51% от наличните аналити срещу 28%, при евтини кръвни

показатели, а не при образна диагностика); изходи от лечението, оценявани отчасти спрямо историческото досие; и публичен набор за оценяване (MIMIC-IV), който езиковият модел може да е виждал по време на обучение — нещо, което авторите посочват като възможна горна граница на представянето.

Колко доверие трябва да има общият читател? Високо, че MIRA е реална и нова демонстрация на възможност — агент, който *действа* по досието, а не просто чатбот. Ниско, че е *по-добър диагностик от лекар*: това твърдение е ограничено до един ретроспективен тест за оценяване и е обърканo от броя на изследванията, оценяването спрямо досието и възможното замърсяване на данните. Заключението на самите автори е правилното — нужни са проспективни проучвания в реална среда, преди това да означава нещо за грижата за пациенти.

Източници

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