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

ШІ-агент самастойна вёў цэлыя клінічныя выпадкі — у сімулятары, на мінулых медыцынскіх запісах

MIRA — новы тып медыцынскага ШІ: замест адказу на адно пытанне сістэма вядзе ўвесь выпадак унутры сімуляванай бальнічнай карты — збірае анамнез, прызначае і чытае даследаванні, прыходзіць да дыягназу і фармуе прызначэнні. На 574 рэтраспектыўных выпадках з публічнай базы, якія ахоплівалі восем загадзя выбраных дыягназаў, аўтары паведамляюць пра вышэйшую, чым у лекараў, дакладнасць дыягназу і пераважна бяспечныя рашэнні, узгодненыя з рэкамендацыямі. Але кожнае ўдакладненне тут важнае: сістэма працавала ў пясочніцы на мінулых запісах і толькі з тэкстам; значная частка перавагі прыпала на станы з адназначнымі тэстамі; яна выкарыстоўвала прыкладна ўдвая больш аналізаў крыві, чым лекары; а некалькі вынікаў ацэньвалі адносна таго, што было запісана ў зыходнай карце. Сапраўдны крок наперад — агент, які дзейнічае ў межах усяго працоўнага працэсу, а не доказ таго, што машына цяпер дыягнастуе лепш за лекара. Самі аўтары падкрэсліваюць: абагульняльнасць, бяспека і кіраванне яшчэ патрабуюць праспектыўных даследаванняў у рэальным свеце.

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

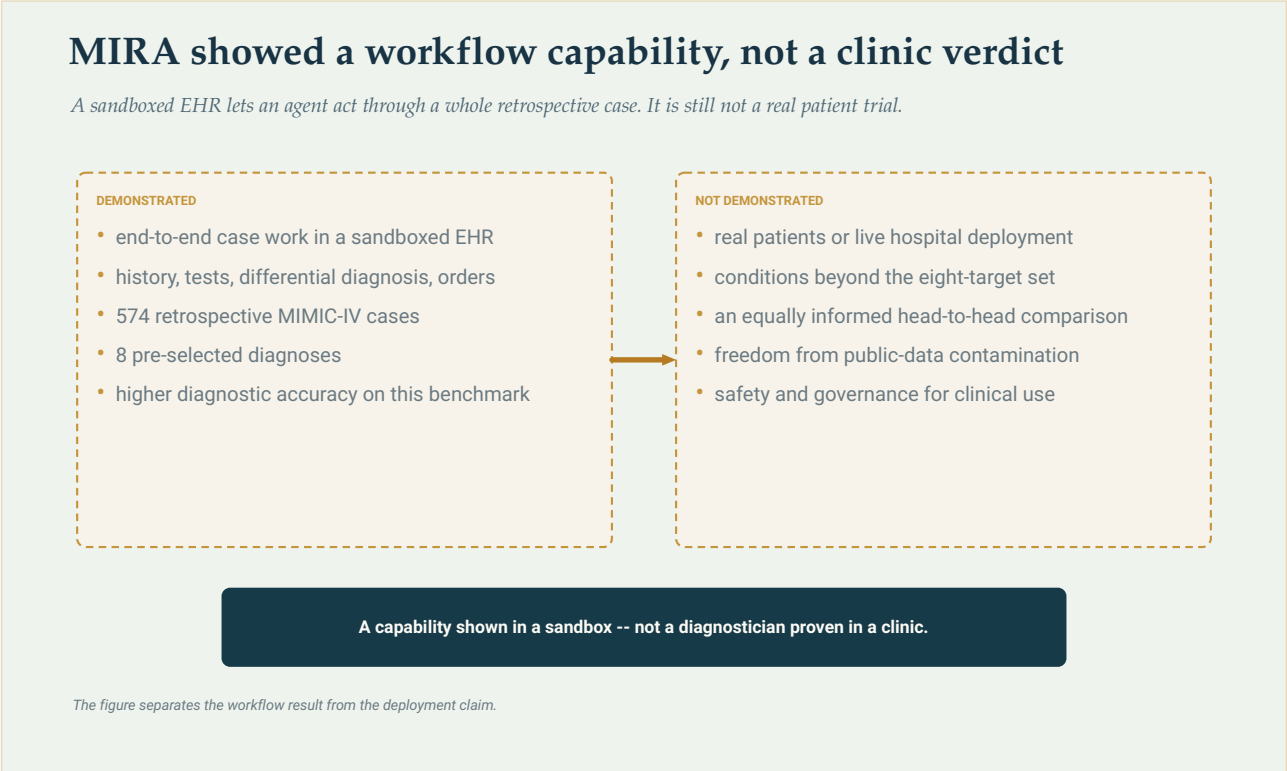
Адказаць на пытанне — не тое самае, што весці ўвесь клінічны выпадак

Большасць гісторый пра медыцынскі ШІ распавядаюць пра мадэль, якая *адказвае*. Ёй даюць клінічную задачу або экзаменацыйнае пытанне, а яна вяртае дыягназ ці абзац рэкамендацый і набірае высокі бал. Гэта карысна, але вузка: нібы разумны кансультант, які ніколі не дакранаецца да медыцынскай карты.

MIRA створана для іншага, і менавіта гэтая розніца тут з'яўляецца сапраўднай навіной. Замест таго каб адказаць на адно пытанне, яна апрацоўвае ўвесь выпадак: чытае карту пацыента, вырашае, якіх даных анамнезу яшчэ не хапае, прызначае лабараторныя, візуалізацыйныя і мікрабіялагічныя даследаванні, чытае вынікі, звужае дыферэнцыяльны дыягназ, а затым фармуе далейшыя прызначэнні — рэцэпты, шпіталізацыю, накіраванне на аперацыю. І робіць гэта, выконваючы дзеянні *ўнутры* электроннай медыцынскай карты, як клініцыст, а не проста генеруючы тэкст, які чалавек павінен перанесці ўручную.

Гэта рэальны крок: ад сістэмы, якая раіць, да сістэмы, якая дзейнічае ў межах усяго працоўнага працэсу. І менавіта такі крок у пераказе лёгка сплюшчваецца да «ШІ перасягнуў лекараў». Таму важна дакладна сказаць, што менавіта правяралі — і дзе.

Версія ў адным сказе: MIRA самастойна вяла выпадкі ад пачатку да канца і на гэтым бенчмарку перасягнула лекараў па дакладнасці дыягназу — але працавала ў ізаляваным сімулятары, на рэтраспектыўных запісах, толькі для васьмі загадзя выбраных дыягназаў, камунікавала выключна тэкстам, а значная частка яе перавагі прыпала на хваробы з найбольш адназначнымі вынікамі тэстаў. Прагрэс рэальны. Табліца вынікаў — не клініка.



MIRA прадэманстравала здольнасць выконваць увесь клінічны працоўны працэс у «пясочніцы» электроннай медыцынскай карты. Яна не прадэманстравала рэальнага клінічнага разгортвання, шырокага дыягнастычнага ахопу або справядлівага параўнання з лекарамі пры аднолькавым аб'ёме інфармацыі.

Што зрабілі аўтары

MIRA — Medical Intelligence for Reasoning and Action — гэта аўтаномны агент на мадэлях OpenAI: частка, якая камунікуе і выконвае дзеянні, працуе на **GPT-4o**, а асобны этап планавання выкарыстоўвае мадэль разважанняў OpenAI **o1**. Яны ўбудаваныя ў спецыяльна створанае, сумяшчальнае са стандартамі асяроддзе на аснове HL7 FHIR — стандарту даных, з дапамогай якога рэальныя бальніцы абменьваюцца медыцынскімі запісамі, — з наборам больш як васьмідзесяці тысяч магчымых клінічных дзеянняў. Унутры гэтай «пясочніцы» MIRA можа атрымаць анамнез пацыента, прызначыць і інтэрпрэтаваць аналізы, візуалізацыю і мікрабіялагічныя тэсты, сфармуляваць дыферэнцыяльны дыягназ і план лячэння — прызначыць лекі, запланаваць аперацыю або шпіталізацыю. Важная дэталь: аўтаматызаванымі «суддзямі», якія ацэньвалі адказы MIRA, былі самі **GPT-4o** — гэта значыць тая ж сям'я мадэляў, якую тэставалі. Пасля таго як рэцэнзент звярнуў увагу на гэтую праблему, аўтары дадалі праверку сертыфікаваным лекарам.

Набор тэстаў узялі з **MIMIC-IV**, вялікай агульнадаступнай дэперсаналізаванай базы электронных медыцынскіх запісаў. З прыкладна 300 000 пацыентаў, якія лячыліся

ў Beth Israel Deaconess Medical Center паміж 2008 і 2019 гадамі, аўтары адабралі **574 клінічныя выпадкі** з **васьмю мэтавымі дыягназамі** — абдамінальнымі паталогіямі і неадкладнымі станамі ўнутранай медыцыны, сярод якіх апендыцыт, панкрэатыт, пнеўманія і інфекцыя мочавых шляхоў. Для кожнага выпадку MIRA пачынала з абмежаванай карціны і мусіла крок за крокам вырашаць, што рабіць далей — па форме гэта нагадвала рэальнае абследаванне, але адбывалася на карце, рэальны фінал якой ужо быў запісаны.

Затым яе вынікі параўналі з працай лекараў на тых жа выпадках.

Што насамрэч азначае «аўтаномны агент у пясочніцы EHR»

Тут тры словы нясуць вельмі шмат сэнсу, таму іх варта разабраць.

Агент азначае, што сістэма — не чатбот, які адказвае на адзін запыт. Яна працуе цыклам: глядзіць на бягучы стан, выбірае дзеянне (прызначыць гэты тэст, спытаць гэты факт анамнезу), атрымлівае вынік, выбірае наступнае дзеянне — і так да дыягназу і плана. «Інтэлект» тут дае моўная мадэль; *агентнасць* забяспечвае праграмная абвязка вакол яе, якая ператварае тэкст мадэлі ў дазволеныя аперацыі ў EHR і вяртае вынікі назад.

Пясочніца EHR азначае сімуляваную, ізаляваную копію сістэмы медыцынскіх запісаў, а не жывую бальнічную сістэму. MIRA можа «прызначыць» тэст і атрымаць вынік, які ў гэтага пацыента сапраўды быў, бо выпадак гістарычны і адказ ужо ёсць у карце. Ні адно яе дзеянне не закранае рэальнага пацыента або рэальную клініку.

Аўтаномны азначае, што сістэма завяршае выпадак ад пачатку да канца без чалавека ў цыкле — *унутры пясочніцы*. Гэта **не** азначае, што ёй дазволілі без нагляду весці сапраўдных пацыентаў. Гэта вельмі розныя сцвярджэнні, і правяралі толькі першае.

Яшчэ адна важная дэталь пра пясочніцу: MIRA можа *замовіць* любы тэст, але сістэма здольная вярнуць вынік толькі тады, калі гэты тэст **сапраўды праводзілі** дадзенаму пацыенту. Калі яна запытвае аналіз крыві, які пацыент рэальна здаваў, то атрымлівае рэальныя значэнні; калі просіць сканаванне, якога ніхто не прызначаў, атрымлівае «N/A — немагчыма выканаць», а не выдуманую лічбу. З анамнезам гэтак жа: калі ў карце няма адказу на пытанне MIRA, сімуляваны пацыент проста кажа, што не ведае. Таму MIRA не можа чароўным чынам атрымаць вырашальны тэст, якога ў рэальным выпадку ніколі не рабілі: яна абмежаваная тым, што змяшчаў фактычны працэс абследавання.

Гэтае адрозненне важнае, бо слова «аўтаномны» прасцей за ўсё памылкова прачытаць у другім сэнсе.

Што яны выявілі

На бенчмарку **MIRA перасягнула лекараў па дакладнасці дыягназу** — але за гучнай фармулёўкай стаяць тры розныя лічбы, якія лёгка пераблытаць.

Самастойна, на ўсіх **574 выпадках**, MIRA назвала правільны дыягназ у **88.9%** выпадкаў — правільнасць вызначалі па выпісным дыягназе, рэальна запісаным для кожнага пацыента ў MIMIC-IV. У прамым параўнанні з людзьмі лекары працавалі не з усімі 574 выпадкамі, а з агульнай **падвыбаркай з 311 выпадкаў**, маючы тыя ж запісы і інструменты, што і MIRA. На гэтых 311 выпадках MIRA атрымала **87.8%**. Яе параўноўвалі з дзвюма асобна набранымі групамі. Першая — **старэйшая група**: чатыры сертыфікаваныя лекары з 7–11 гадамі досведу, вынік **78.1%**. Другая — **група змешанага ўзроўню досведу**: пераважна малодшыя рэзідэнты плюс два спецыялісты, гэта значыць склад, больш падобны да рэальнага аддзялення неадкладнай дапамогі, — **71.1%**. Тыя ж выпадкі, тыя ж інструменты — і парадак атрымаўся такі: ШІ першы, дасведчаныя лекары другія, каманда з пераважна маладымі лекарамі трэцяя. Асцярожная фармулёўка самога артыкула: для ўсіх васьмі хвароб MIRA была «стабільна эквівалентная лекарам і часта пераўзыходзіла іх».

Далейшыя рашэнні таксама выглядалі добра: у асноўным адпавядалі клінічным рэкамендацыям, былі бяспечнымі адносна лекаў і дарэчнымі адносна шпіталізацыі. Аўтары не зафіксавалі памылак высокай цяжкасці ў пяці катэгорыях лекавай бяспекі — узаемадзеянні прэпаратаў, дазаванне пры парушэнні функцыі ныракаў, алергіі, рызыка падаўжэння QT і прызначэнне апіоідаў — і паведамілі пра правільныя прызначэнні ў 467 з 468 выпадкаў, адначасова проста адзначыўшы, што сістэма «не дасягнула 100% надзейнасці». Калі ўспрымаць гэтыя лічбы літаральна, вынік уражвае: агент вядзе ўвесь выпадак і апярэджвае лекараў па дыягназе.

Але *форма* гэтай перамогі не менш важная, чым сам факт. Незалежныя спецыялісты, якія чыталі артыкул, звярнулі ўвагу на тое, адкуль паходзіла перавага. У экспертнай панэлі Science Media Centre доктар Wei Xing (University of Sheffield) адзначыў, што галоўны вынік — перавага ШІ па дакладнасці дыягназу — **у асноўным абумоўлены станамі з выразнымі вынікамі тэстаў**, такімі як апендыцыт і панкрэатыт, дзе вырашальны здымак або лабараторнае значэнне хутка закрывае пытанне. Для пнеўманіі і інфекцый мочавых шляхоў — дзвюх вельмі распаўсюджаных прычын звароту па неадкладную дапамогу — *і* ШІ, *і* лекары паказалі горшыя вынікі, а разрыў паміж імі быў найменшы.

Была і асіметрыя ў тым, колькі інфармацыі выкарыстоўвалі абодва бакі, і яна бачная з лічбаў самога артыкула. MIRA значна больш актыўна абапіралася на лабараторыю: выкарыстоўвала каля **51%** лабараторных паказчыкаў, даступных у звычайнай практыцы, супраць прыкладна **28%** у сертыфікаваных лекараў — медыянна на сем дадатковых паказчыкаў крыві на адзін выпадак. Аўтары асцярожна падкрэсліваюць, што гэта ўсё роўна *менш* за фактычны ўзровень руціннага выкарыстання тэстаў у наборы даных, а не стратэгія «замовіць усё», і не выявілі сістэматычнага павелічэння больш дарагой тамаграфічнай візуалізацыі. Аднак больш інфармацыі само па сабе можа

павысіць дакладнасць дыягназу. Таму гэта не зусім параўнанне двух аднолькава інфармаваных удзельнікаў — на што Xing проста звярнуў увагу. Лекар, якому дазволіць без абмежаванняў прызначаць усе танныя аналізы крыві, не ўлічваючы кошт, дыскамфорт або затрымку, таксама будзе выглядаць больш дакладным на паперы.

Чаму «перамагчы лекараў» не азначае «быць лепшым лекарам»

Перамогу на бенчмарку лёгка пераацаніць. Паміж «MIRA набрала больш» і «MIRA лепшая» стаяць як мінімум тры рэчы.

Па-першае, **з чым параўноўвалі адказы**. Прафесар Julie Jacko (University of Edinburgh) заўважыла, што некалькі ключавых паказчыкаў MIRA вызначаны адносна таго, што было задакументавана ў зыходным наборы даных. Гэта значыць, сістэму ўзнагароджваюць за **узнаўленне зафіксаваных клінічных паводзінаў**, а не абавязкова за дэманстрацыю аптымальнай дапамогі. Гістарычная карта становіцца ключом з адказамі. Для бенчмарку гэта разумны падыход, але ён вымярае адпаведнасць таму, што было зроблена, а не абсалютную правільнасць. Дзеля справядлівасці, гэтая праблема найбольш датычыцца паказчыкаў *узгодненасці лячэння* — ці супадаюць прызначэнні MIRA з картай, — тады як асноўную дыягнастычную дакладнасць ацэньвалі па выпісным дыягназе, што бліжэй да рэальнага выніку, чым простае паўтарэнне дакументацыі.

Па-другое, ужо згаданая **асіметрыя інфармацыі**: значна больш аналізаў крыві — каля 51% даступных паказчыкаў супраць 28% — азначае больш доказаў. Верагодна, частка разрыву ў дакладнасці не «выведзена разважаннем», а *набыта* дадатковымі тэстамі.

Па-трэцяе, **дзе менавіта працавала сістэма**. Гэта была пясочніца з рэтраспектыўнымі выпадкамі, васьмю загадзя адабранымі дыягназамі і выключна тэкставым інтэрфейсам. Рэальная клінічная ацэнка залежыць не толькі ад таго, *што* кажа пацыент, але і ад таго, *як* ён гэта кажа, а таксама ад фізікальнага агляду, назіраных паводзінаў і мовы цела — нічога з гэтага тэкставы агент, які чытае ўжо завершаную карту, не бачыць. А пацыент, чыя праблема не адносіцца да васьмі выбраных дыягназаў, проста знаходзіцца па-за межамі таго, пра што можа казаць гэтае даследаванне.

Ёсць і больш ціхая праблема, узнятая рэцэнзентамі: **забруджванне навучальнымі данымі**. MIMIC-IV — публічная база, пра якую шмат пішуць, таму моўная мадэль, навучаная на адкрытым інтэрнэце, магла ўжо бачыць артыкулы, разборы выпадкаў або самі даныя. Доктар Midhun Parakkal Unni (University of Sheffield) проста ўказаў на гэта: калі частка адказаў ужо была ў навучальных даных, то пэўная доля выніку можа быць памяццю, а не клінічным разважаннем, і адрозніць адно ад другога можа толькі незалежнае паўтарэнне. Важна, што аўтары не адмахваюцца ад гэтай праблемы: яны пішуць, што вынікі «можна асцярожна інтэрпрэтаваць як магчымую верхнюю мяжу» і што яны «могуць завышаць абагульняльнасць на іншыя публічныя выпадкі». Рэдкі

выпадак, калі сам артыкул ставіць столь над уласным гучным вынікам.

Усё гэта не робіць MIRA менш цікавай. Яно робіць правільнае прачытанне дакладнейшым: на рэтраспектыўным бенчмарку агент, здольны дзейнічаць па ўсёй карце, перасягнуў лекараў перадусім на дыягназах з выразнымі тэстамі — часткова замаўляючы больш аналізаў, часткова ўзнаўляючы зафіксаваны ў карце ход. Гэта рэальная дэманстрацыя магчымасці, а не вердыкт пра тое, што машыны цяпер дыягнастуюць лепш за лекараў.

Чаго гэтая праца *не* даказвае

- Яна **не** паказвае, што MIRA працуе на рэальных пацыентах. Усе выпадкі былі рэтраспектыўнымі і сімуляванымі; ніводнага пацыента яна не вяла.
- Яна **не** паказвае, што сістэма працуе па-за васьмю дыягназамі. Станы па-за загадзя выбраным наборам — гэта значыць значная частка складанай, недыферэнцыраванай медыцыны — не тэставаліся.
- Яна **не** паказвае, што сістэму бяспечна разгортваць. Самі аўтары пішуць, што абагульняльнасць, бяспеку і кіраванне трэба праверыць у праспектыўных даследаваннях у рэальным свеце.
- Яна **не** ўстанаўлівае справядлівага прамога параўнання з лекарамі. MIRA выкарыстоўвала значна больш аналізаў крыві (каля 51% даступных паказчыкаў супраць 28%), а некалькі вынікаў лячэння часткова ацэньвалі паводле таго, што было запісана ў гістарычнай карце, а не паводле незалежнага эталона найлепшай дапамогі.
- Яна **не** даказвае адсутнасці запамінання. Публічныя даныя MIMIC-IV маглі перасякацца з навучальнымі данымі мадэлі; выключыць гэта можа толькі незалежная рэплікацыя.
- Яна **не** азначае «ШІ замяняе лекараў». Гэта сістэма падтрымкі рашэнняў, здольная *дзейнічаць* па ўсёй карце; кансэнсус незалежных каментатараў у тым, што рэальнае выкарыстанне павінна адбывацца ў партнёрстве з клініцыстамі, якія захоўваюць паўнамоцтвы і дадаюць усё тое, чаго няма ў тэкставым запісе.

Наколькі пераканаўчыя сведчанні?

Тут варта падзяліць сцвярджэнне на дзве часткі, бо сіла доказаў для іх вельмі розная.

Як дэманстрацыя магчымасці — што агент на моўнай мадэлі можа правесці ўвесь клінічны выпадак у пясочніцы EHR, паслядоўна аб’ядноўваючы анамнез, тэсты, дыягназ і прызначэнні, — праца сапраўды новая і дастаткова пераканаўчая. Менавіта гэта тут новае, і менавіта на гэта варта звярнуць увагу.

Як сцвярджэнне пра перавагу — што MIRA *лепшая за лекараў* — доказы маюць выразныя межы і патрабуюць асцярожнага прачытання. Вынік справядлівы для канкрэтнага рэтраспектыўнага бенчмарку з васьмю дыягназамі, асіметрыяй інфармацыі (больш тэстаў), ключом адказаў з гістарычнай карты і рэальнай магчымасцю забруджвання

навучальнымі данымі. Гэтага дастаткова, каб сказаць: «агент уражліва выступіў на гэтым бенчмарку». Гэтага недастаткова, каб сказаць: «агент — лепшы дыягност, чым лекар», — і другога аўтары не сцвярджаюць.

Найбольш карысная пазіцыя — не «ШІ перамог лекараў» і не «гэта ўсяго толькі цацка». Дакладней так: новы тып медыцынскага ШІ — той, які *дзейнічае* па ўсёй карце, а не проста адказвае на пытанні, — добра паказаў сябе на складаным рэтраспектыўным бенчмарку і цяпер павінен даказаць сябе там, дзе яго яшчэ не выпрабоўвалі: праспектыўна, на рэальных недыферэнцыраваных пацыентах і пры належным кіраванні.

Чаму гэта важна

За апошнія некалькі гадоў цікавае пытанне ў медыцынскім ШІ незаўважна змянілася. Раней яно гучала: «ці можа мадэль правільна паставіць дыягназ?» — і мадэлі ўсё часцей адказвалі «так» на ўсё больш чыстых тэставых наборах. MIRA абазначае пераход да больш складанага пытання: «ці можа мадэль *выконваць працу* — збіраць даныя, прызначаць, інтэрпрэтаваць, вырашаць і дзейнічаць — у межах усяго працоўнага працэсу?» Гэта больш карыснае і сумленнае пытанне, бо менавіта ў дзеянні жывуць і складанасць, і рызыка.

Менавіта таму рамка важная. Рэфлекторны загаловак — *ШІ перасягнуў лекараў* — паказвае на найменш новую і найслабей падтрыманаю частку выніку. Сапраўды новае тут сціплейшае паводле фармулёўкі, але важнейшае паводле наступстваў: агент, здольны рухацца праз усю карту. Калі гэтая магчымасць вытрымае далейшую праверку, яна зменіць **працоўны працэс** задоўга да таго, як зменіць таго, хто ім кіруе. Лекар не знікае; першае месца, дзе апынецца такі інструмент, — прызначэнні, інтэрпрэтацыя, адсочванне вынікаў, гэта значыць сам працоўны працэс.

І гэта правільна змяняе цяжар доказу. Перамога ў табліцы рэтраспектыўных выпадкаў — прычына правесці праспектыўнае выпрабаванне, а не яго замена. Аўтары кажуць тое ж самае. Сумленнае прачытанне MIRA — гэта запрашэнне да наступнага даследавання, праведзенага паводле стандарту, які медыцына ўжо ведае: на пацыентах, а не толькі на бенчмарках.

Кароткі вынік

MIRA — аўтаномны ШІ-агент, які працуе ў ізаляванай электроннай медыцынскай карце: ён можа збіраць анамнез, прызначаць і інтэрпрэтаваць лабараторныя, візуалізацыйныя і мікрабіялагічныя даследаванні, ставіць дыягназ і фармаваць план лячэння. На 574 рэтраспектыўных выпадках з публічнай базы MIMIC-IV, якія ахоплівалі восем загадзя выбраных дыягназаў, сістэма перасягнула лекараў па дакладнасці дыягназу (88.9% у цэлым; у прамым параўнанні 87.8% супраць 78.1%) і пераважна прымала

рашэнні, узгодненыя з рэкамендацыямі і бяспечныя адносна лекаў. Але ацэнка была сімуляцыяй на мінулых запісах і толькі ў тэксце; значная частка перавагі прыпала на станы з адназначнымі тэстамі; MIRA выкарыстоўвала значна больш аналізаў крыві, чым лекары (каля 51% даступных паказчыкаў супраць 28%); некалькі вынікаў лячэння ацэньвалі адносна таго, што было запісана ў зыходнай карце; а публічнасць набору даных стварае рэальную рызыку забруджвання навучальнымі данымі — тое, што самі аўтары называюць магчымай верхняй мяжой сваіх вынікаў. Сапраўдны крок наперад — агент, які дзейнічае ў межах усяго працоўнага працэсу, а не адказвае на ізаляваныя пытанні. Аўтары проста пішуць, што абагульняльнасць, бяспека і кіраванне яшчэ патрабуюць праспектыўных даследаванняў у рэальным свеце. Менавіта так вынік і варта чытаць: уражлівая дэманстрацыя магчымасці, а не доказ таго, што ШІ дыягнастуе лепш за лекара, і не сістэма, гатовая да рэальнай клінікі.

Праверка без прыкрас

Што паказвае артыкул: Агент на моўнай мадэлі (MIRA) можа правесці ўвесь клінічны выпадак ад пачатку да канца ў пясочніцы EHR — анамнез, тэсты, дыягназ, прызначэнні — і на рэтраспектыўным бенчмарку з 574 выпадкаў і васьмі дыягназаў перасягнуў лекараў па дакладнасці дыягназу (88.9% у цэлым; 87.8% супраць 78.1% у параўнанні з сертыфікаванымі лекарамі), не дапусціўшы памылак высокай цяжкасці ў прызначэнні лекаў.

Што праўдападобна, але не даказана: Што гэтая магчымасць прынясе карысць рэальным недыферэнцыраваным пацыентам; што перавага ў дакладнасці адлюстроўвае лепшае разважанне, а не большую колькасць прызначаных тэстаў, узгадненне з гістарычнай картай або запомненыя публічныя даныя.

Чаго яна не паказвае: Што MIRA працуе па-за васьмю дыягназамі; што яе бяспечна разгортваць; што яна перамагае лекараў у справядлівым параўнанні пры аднолькавай інфармацыі; што яна замяняе клініцыстаў; што вынікі гарантавана свабодныя ад забруджвання навучальнымі данымі.

Галоўныя абмежаванні: Рэтраспектыўная, сімуляваная і толькі тэкставая ацэнка; восем загадзя выбраных дыягназаў; асіметрыя інфармацыі (значна больш аналізаў крыві — каля 51% даступных паказчыкаў супраць 28%, прычым гаворка пра танныя аналізы крыві, а не даражэйшую візуалізацыю); паказчыкі лячэння часткова ацэньвалі па гістарычнай карце; публічны бенчмарк MIMIC-IV, які моўная мадэль магла бачыць падчас навучання — і самі аўтары называюць гэта магчымай верхняй мяжой выніковасці.

Якой упэўненасці заслугоўвае выснова для звычайнага чытача? Высокай у тым, што MIRA — рэальная і новая дэманстрацыя магчымасці: гэта агент, які *дзейнічае* па ўсёй карце, а не проста чатбот. Нізкай адносна сцвярджэння, што ён *лепшы дыягност, чым лекар*: яно абмежавана адным рэтраспектыўным бенчмаркам і змешана з эфектам большай колькасці тэстаў, ацэнкай па карце і магчымым забруджваннем даных.

Уласная выснова аўтараў тут найлепшая: перш чым гэта будзе нешта значыць для дапамогі пацыентам, патрэбна праспектыўнае даследаванне ў рэальным свеце.

Крыніцы

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