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

ШІ-агент самостійно вів цілі клінічні випадки — у симуляторі, на минулих медичних записах

MIRA — новий тип медичного ШІ: замість відповіді на одне запитання система веде весь випадок усередині симульованої лікарняної картки — збирає анамнез, призначає й читає тести, доходить до діагнозу та формує призначення. На 574 ретроспективних випадках із публічної бази, що охоплювали вісім наперед обраних діагнозів, автори повідомляють про вищу за лікарів точність діагнозу та переважно безпечні, узгоджені з настановами рішення. Але кожне уточнення тут важливе: система працювала в пісочниці на минулих записах і лише в тексті; значна частина переваги припала на стани з однозначними тестами; вона використовувала приблизно вдвічі більше аналізів крові, ніж лікарі; а кілька результатів оцінювали відносно того, що було записано в оригінальній картці. Справжній поступ — агент, який діє в межах усього робочого процесу, а не доказ того, що машина тепер діагностує краще за лікаря. Самі автори наголошують: узагальнюваність, безпека та управління ще потребують проспективних досліджень у реальному світі.

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

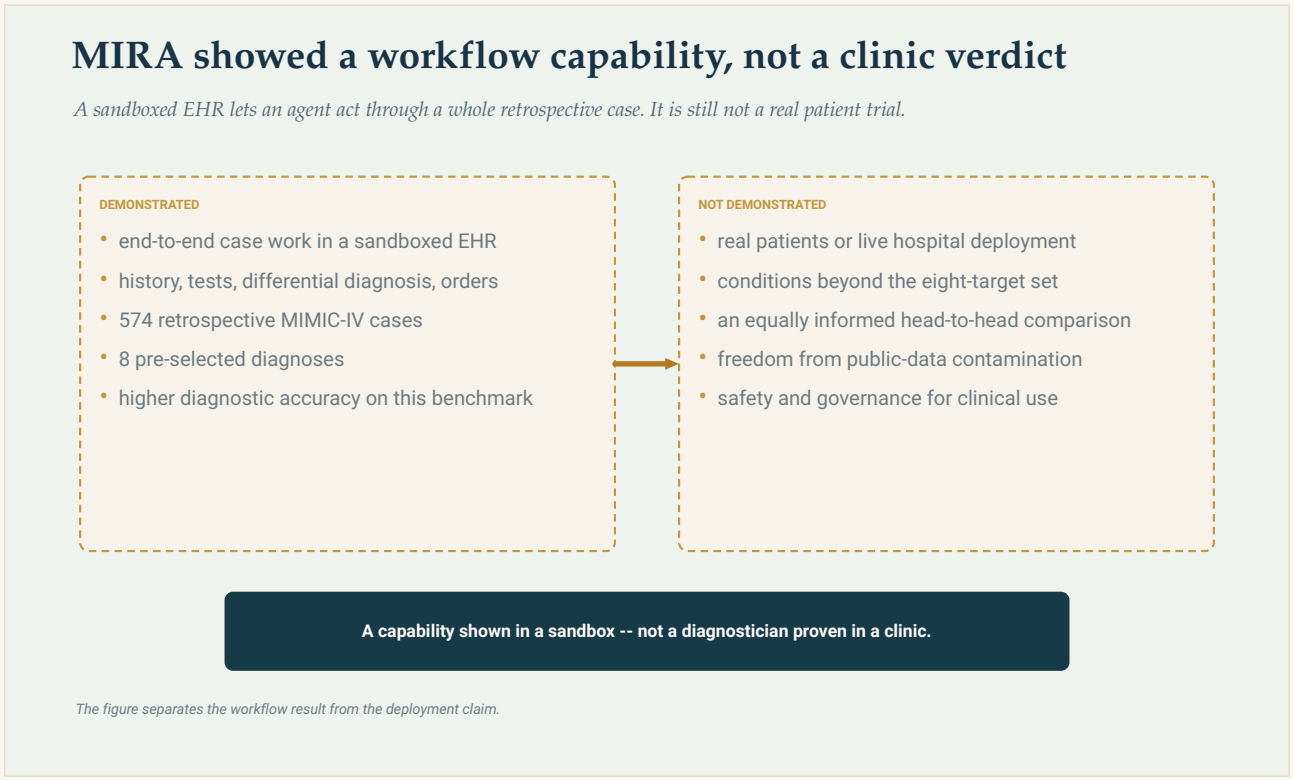
Відповісти на запитання — не те саме, що вести весь клінічний випадок

Більшість історій про медичний ШІ розповідають про модель, яка *відповідає*. Їй дають клінічну задачу або екзаменаційне запитання, а вона повертає діагноз чи абзац рекомендацій і набирає високий бал. Це корисно, але вузько: ніби розумний консультант, який ніколи не торкається медичної картки.

MIRA створена для іншого, і саме ця різниця тут є справжньою новиною. Замість того щоб відповісти на одне запитання, вона опрацьовує весь випадок: читає картку пацієнта, вирішує, яких даних анамнезу ще бракує, призначає лабораторні, візуалізаційні та мікробіологічні дослідження, читає результати, звужує диференційний діагноз, а потім формує подальші призначення — рецепти, госпіталізацію, направлення на операцію. І робить це, виконуючи дії *всередині* електронної медичної картки, як клініцист, а не просто генеруючи текст, який людина має перенести вручну.

Це справжній крок: від системи, яка радить, до системи, яка діє в межах усього робочого процесу. І саме такий крок у висвітленні легко сплющується до «ШІ перевершив лікарів». Тому важливо точно сказати, що саме перевіряли — і де.

Версія в одному реченні: MIRA самостійно вела випадки від початку до кінця й на цьому бенчмарку перевершила лікарів за точністю діагнозу — але працювала в ізольованому симуляторі, на ретроспективних записах, лише для восьми наперед обраних діагнозів, спілкувалася тільки текстом, а значна частина її переваги припала на хвороби з найоднозначнішими результатами тестів. Поступ реальний. Таблиця результатів — це не клініка.



MIRA продемонструвала здатність виконувати весь клінічний робочий процес у «пісочниці» електронної медичної картки. Вона не продемонструвала реального клінічного розгортання, широкого діагностичного охоплення чи справедливого порівняння з лікарями за однакового обсягу інформації.

Що зробили автори

MIRA — Medical Intelligence for Reasoning and Action — це автономний агент на моделях OpenAI: частина, яка спілкується та виконує дії, працює на **GPT-4o**, а окремий етап планування використовує модель міркування OpenAI **o1**. Вони вбудовані у спеціально створене, сумісне зі стандартами середовище на основі HL7 FHIR — стандарту даних, яким реальні лікарні обмінюються медичними записами, — з набором понад вісімдесят тисяч можливих клінічних дій. Усередині цієї «пісочниці» MIRA може отримати анамнез пацієнта, призначити та інтерпретувати аналізи, візуалізацію й мікробіологічні тести, сформувати диференційний діагноз і план лікування — призначити ліки, запланувати операцію або госпіталізацію. Важлива деталь: автоматизованими «суддями», які оцінювали відповіді MIRA, були самі **GPT-4o** — тобто та сама родина моделей, яку тестували. Після того як рецензент звернув увагу на цю проблему, автори додали перевірку сертифікованим лікарем.

Набір тестів взяли з **MIMIC-IV**, великої загальнодоступної деперсоналізованої бази електронних медичних записів. Із приблизно 300 000 пацієнтів, яких лікували в Beth Israel Deaconess Medical Center між 2008 і 2019 роками, автори відібрали **574 клінічні випадки** з **вісьмома цільовими діагнозами** — абдомінальними патологіями та

невідкладними станами внутрішньої медицини, серед яких апендицит, панкреатит, пневмонія й інфекція сечових шляхів. Для кожного випадку MIRA починала з обмеженої картини й мусила крок за кроком вирішувати, що робити далі — за формою це нагадувало реальне обстеження, але відбувалося на картці, реальний фінал якої вже був записаний.

Потім її результати порівняли з роботою лікарів на тих самих випадках.

Що насправді означає «автономний агент у пісочниці EHR»

Тут три слова несуть дуже багато змісту, тому їх варто розібрати.
Агент означає, що система не є чатботом, який відповідає на один запит. Вона працює циклом: дивиться на поточний стан, обирає дію (призначити цей тест, запитати той факт анамнезу), отримує результат, обирає наступну дію — і так до діагнозу та плану. «Інтелект» тут дає мовна модель; *агентність* забезпечує обв'язка навколо неї, яка перетворює текст моделі на дозволені операції в EHR і повертає результати назад.
Пісочниця EHR означає симульовану, ізольовану копію системи медичних записів, а не живу лікарняну систему. MIRA може «призначити» тест і отримати результат, який цей пацієнт справді мав, бо випадок історичний і відповідь уже є в картці. Жодна її дія не зачіпає реального пацієнта чи реальну клініку.
Автономний означає, що система завершує випадок від початку до кінця без людини в циклі — *усередині пісочниці*. Це **не** означає, що їй дозволили без нагляду вести справжніх пацієнтів. Це дуже різні твердження, і перевіряли лише перше.
Ще одна важлива деталь про пісочницю: MIRA може *замовити* будь-який тест, але система здатна повернути результат лише тоді, коли цей тест **справді проводили** цьому пацієнтові. Якщо вона просить аналіз крові, який пацієнт реально здавав, то отримує реальні значення; якщо просить сканування, якого ніхто не призначав, отримує «N/A — неможливо виконати», а не вигадане число. З анамнезом так само: якщо в картці немає відповіді на запитання MIRA, симульований пацієнт просто каже, що не знає. Отже, MIRA не може чарівним чином отримати вирішальний тест, якого в реальному випадку ніколи не робили: вона обмежена тим, що містив фактичний процес обстеження.
Ця різниця важлива, бо саме слово «автономний» найпростіше помилково прочитати у другому сенсі.

Що вони знайшли

На бенчмарку **MIRA перевершила лікарів за точністю діагнозу** — але за гучним формулюванням стоять три різні числа, які легко змішати.

Самостійно, на всіх **574 випадках**, MIRA назвала правильний діагноз у **88,9%** випадків — правильність визначали за виписним діагнозом, реально записаним для кожного

пацієнта в MIMIC-IV. У прямому порівнянні з людьми лікарі працювали не з усіма 574 випадками, а зі спільною **підвибіркою із 311 випадків**, маючи ті самі записи й інструменти, що й MIRA. На цих 311 випадках MIRA отримала **87,8%**. Її порівняли з двома окремо набраними групами. Перша — **старша група**: чотири сертифіковані лікарі з 7–11 роками досвіду, результат **78,1%**. Друга — **група зі змішаним рівнем досвіду**: переважно молодші резиденти плюс два спеціалісти, тобто склад, ближчий до реального відділення невідкладної допомоги, — **71,1%**. Ті самі випадки, ті самі інструменти — і порядок вийшов такий: ШІ перший, досвідчені лікарі другі, команда з переважно молодшими лікарями третя. Обережне формулювання самої статті: для всіх восьми хвороб MIRA була «стабільно еквівалентною лікарям і часто перевершувала їх».

Подальші рішення також виглядали добре: переважно відповідали клінічним настановам, були безпечними щодо ліків і доречними щодо госпіталізації. Автори не зафіксували помилок високої тяжкості в п’яти категоріях медикаментозної безпеки — взаємодії препаратів, дозування при порушенні функції нирок, алергії, ризик подовження QT та призначення опіоїдів — і повідомили про правильні призначення у 467 із 468 випадків, водночас прямо зазначивши, що система «не досягла 100% надійності». Якщо сприймати ці числа буквально, результат вражає: агент веде весь випадок і випереджає лікарів за діагнозом.

Але *форма* цієї перемоги не менш важлива за сам факт. Незалежні фахівці, які читали статтю, звернули увагу на те, звідки походила перевага. У експертній панелі Science Media Centre доктор Вей Сін (University of Sheffield) зазначив, що головний результат — перевага ШІ за точністю діагнозу — **переважно зумовлений станами з чіткими результатами тестів**, такими як апендицит і панкреатит, де вирішальний знімок або лабораторне значення швидко закриває питання. Для пневмонії та інфекцій сечових шляхів — двох дуже поширених причин звернення до невідкладної допомоги — *і* ШІ, *і* лікарі показали найгірші результати, а розрив між ними був найменшим.

Була й асиметрія в тому, скільки інформації використовували обидві сторони, і вона видно з чисел самої статті. MIRA значно активніше покладалася на лабораторію: використала близько **51%** лабораторних показників, доступних у звичайній практиці, проти приблизно **28%** у сертифікованих лікарів — медіанно на сім додаткових показників крові на один випадок. Автори обережно підкреслюють, що це все одно *менше* за фактичний рівень рутинного використання тестів у датасеті, а не стратегія «замовити все», і не виявили систематичного збільшення дорожчої томографічної візуалізації. Проте більше інформації саме по собі може підвищити точність діагнозу. Тому це не зовсім порівняння двох однаково поінформованих учасників — на що Сін прямо звернув увагу. Лікар, якому дозволити без обмежень замовляти всі дешеві аналізи крові, не зважаючи на вартість, дискомфорт чи затримку, теж виглядатиме точнішим на папері.

Чому «перемогти лікарів» не означає «бути кращим лікарем»

Перемогу на бенчмарку легко переоцінити. Між «MIRA набрала більше» та «MIRA краща» стоять щонайменше три речі.

По-перше, **з чим порівнювали відповіді**. Професорка Джулі Джеко (University of Edinburgh) зауважила, що кілька ключових показників MIRA визначені відносно того, що було задокументовано у вихідному датасеті. Тобто систему винагороджують за **відтворення зафіксованої клінічної поведінки**, а не обов'язково за демонстрацію оптимальної допомоги. Історична картка стає ключем із відповідями. Для бенчмарку це розумний підхід, але він вимірює відповідність тому, що було зроблено, а не абсолютну правильність. Справедливості заради, ця проблема найбільше стосується показників *узгодженості лікування* — чи збігаються призначення MIRA з карткою, — тоді як головну діагностичну точність оцінювали за виписним діагнозом, що ближче до реального результату, ніж просте повторення документації.

По-друге, вже згадана **асиметрія інформації**: значно більше аналізів крові — близько 51% доступних показників проти 28% — означає більше доказів. Імовірно, частину розриву в точності не «виведено міркуванням», а *придбано* додатковими тестами.

По-третє, **де саме система працювала**. Це була пісочниця з ретроспективними випадками, вісьмома наперед відібраними діагнозами й виключно текстовим інтерфейсом. Реальне клінічне оцінювання залежить не тільки від того, *що* каже пацієнт, а й від того, *як* він це каже, а також від фізикального огляду, спостережуваної поведінки та мови тіла — нічого з цього текстовий агент, який читає вже завершену картку, не бачить. А пацієнт, чия проблема не належить до восьми обраних діагнозів, просто лежить поза межами того, про що може говорити це дослідження.

Є й тихіша проблема, яку підняли рецензенти: **забруднення навчальними даними**. MIMIC-IV є публічною базою, про яку багато пишуть, тому мовна модель, навчена на відкритому інтернеті, могла вже бачити статті, розбори випадків або самі дані. Доктор Мідхун Параккал Унні (University of Sheffield) прямо вказав на це: якщо частина відповідей уже була в навчальних даних, то певна частина результату може бути пам'яттю, а не клінічним міркуванням, і відрізнити одне від іншого може лише незалежне повторення. Важливо, що автори не відмахуються від цієї проблеми: вони пишуть, що результати «можна обережно інтерпретувати як можливу верхню межу» і що вони «можуть завищувати узагальнюваність на інші публічні випадки». Рідкісний випадок, коли сама стаття ставить стелю над власним гучним результатом.

Усе це не робить MIRA менш цікавою. Воно робить правильне прочитання точнішим: на ретроспективному бенчмарку агент, здатний діяти по всій картці, перевершив лікарів насамперед на діагнозах із чіткими тестами — частково замовляючи більше аналізів, частково відтворюючи зафіксований у картці перебіг. Це реальна демонстрація можливості, а не вердикт, що машини тепер діагностують краще за лікарів.

Чого ця робота *не* доводить

- Вона **не** показує, що MIRA працює на реальних пацієнтах. Усі випадки були ретроспективними й симульованими; жодного пацієнта вона не вела.
- Вона **не** показує, що система працює за межами восьми діагнозів. Стани поза наперед вибраним набором — тобто значна частина складної, недиференційованої медицини — не тестувалися.
- Вона **не** показує, що систему безпечно розгортати. Самі автори пишуть, що узагальнюваність, безпеку та управління потрібно перевірити в проспективних дослідженнях у реальному світі.
- Вона **не** встановлює справедливого прямого порівняння з лікарями. MIRA використовувала набагато більше аналізів крові (близько 51% доступних показників проти 28%), а кілька результатів лікування частково оцінювали за тим, що було записано в історичній картці, а не за незалежним еталоном найкращої допомоги.
- Вона **не** доводить відсутність запам'ятовування. Публічні дані MIMIC-IV могли перетинатися з навчальними даними моделі; це може виключити лише незалежна реплікація.
- Вона **не** означає «ШІ замінює лікарів». Це система підтримки рішень, здатна *діяти* по всій картці; консенсус незалежних коментаторів полягає в тому, що реальне використання має відбуватися в партнерстві з клініцистами, які зберігають повноваження й додають усе, чого немає в текстовому записі.

Наскільки переконливі докази?

Тут варто розділити твердження на дві частини, бо сила доказів для них дуже різна.

Як демонстрація можливості — що агент на мовній моделі може провести весь клінічний випадок у пісочниці EHR, послідовно поєднуючи анамнез, тести, діагноз і призначення, — робота справді нова й досить переконлива. Саме це тут нове, і саме на це варто звернути увагу.

Як твердження про перевагу — що MIRA *краща за лікарів* — докази мають чіткі межі й потребують обережного читання. Результат справедливий для конкретного ретроспективного бенчмарку з вісьмома діагнозами, асиметрією інформації (більше тестів), ключем відповідей із історичної картки й реальною можливістю забруднення навчальними даними. Цього достатньо, щоб сказати: «агент вражаюче виступив на цьому бенчмарку». Цього недостатньо, щоб сказати: «агент є кращим діагностом, ніж лікар», — і автори другого не стверджують.

Найкорисніша позиція — не «ШІ переміг лікарів» і не «це лише іграшка». Правильніше так: новий тип медичного ШІ — той, що *діє* по всій картці, а не просто відповідає на запитання, — добре показав себе на складному ретроспективному бенчмарку й тепер має довести себе там, де його ще не випробовували: проспективно, на реальних

недиференційованих пацієнтах і під належним управлінням.

Чому це важливо

За останні кілька років цікаве питання в медичному ШІ непомітно змінилося. Раніше воно звучало: «чи може модель правильно поставити діагноз?» — і моделі дедалі частіше відповідали «так» на дедалі чистіших тестових наборах. MIRA позначає перехід до складнішого питання: «чи може модель *виконувати роботу* — збирати дані, призначати, інтерпретувати, вирішувати й діяти — в межах усього робочого процесу?» Це корисніше й чесніше питання, бо саме в дії живуть і складність, і ризик.

Саме тому рамка важлива. Рефлекторний заголовок — *ШІ перевершив лікарів* — вказує на найменш нову й найслабше підтриману частину результату. Справді нове тут скромніше за формулюванням, але наслідковіше: агент, який здатен рухатися через усю картку. Якщо ця можливість витримає подальшу перевірку, вона змінить **робочий процес** задовго до того, як змінить того, хто ним керує. Лікар не зникає; перше місце, де опиниться такий інструмент, — призначення, інтерпретація, відстеження результатів, тобто сам workflow.

І це правильно змінює тягар доказу. Перемога в таблиці ретроспективних випадків — причина провести проспективне випробування, а не його заміна. Автори кажуть те саме. Чесне прочитання MIRA — це запрошення до наступного дослідження, проведеного за стандартом, який медицина вже знає: на пацієнтах, а не лише на бенчмарках.

Короткий підсумок

MIRA — автономний агент ШІ, що працює в ізольованій електронній медичній картці: він може збирати анамнез, призначати й інтерпретувати лабораторні, візуалізаційні та мікробіологічні дослідження, встановлювати діагноз і формувати план лікування. На 574 ретроспективних випадках із публічної бази MIMIC-IV, що охоплювали вісім наперед вибраних діагнозів, система перевершила лікарів за точністю діагнозу (88,9% загалом; у прямому порівнянні 87,8% проти 78,1%) і переважно ухвалювала рішення, сумісні з настановами й безпечні щодо медикаментів. Але оцінювання було симуляцією на минулих записах і лише в тексті; значна частина переваги припала на стани з однозначними тестами; MIRA використовувала набагато більше аналізів крові, ніж лікарі (близько 51% доступних показників проти 28%); кілька результатів лікування оцінювали відносно того, що було записано в оригінальній картці; а публічність датасету створює реальний ризик забруднення навчальних даних — те, що самі автори називають можливою верхньою межею своїх результатів. Справжній поступ — агент, який діє по всьому робочому процесу, а не відповідає на ізольовані

запитання. Автори прямо пишуть, що узагальнюваність, безпека та управління ще потребують проспективних досліджень у реальному світі. Саме так результат і варто читати: вражаюча демонстрація можливості, а не доказ того, що ШІ діагностує краще за лікаря, і не система, готова до реальної клініки.

Тверезa оцінка

Що показує стаття: Агент на мовній моделі (MIRA) може провести весь клінічний випадок від початку до кінця в пісочниці EHR — анамнез, тести, діагноз, призначення — і на ретроспективному бенчмарку з 574 випадків та восьми діагнозів перевершив лікарів за точністю діагнозу (88,9% загалом; 87,8% проти 78,1% у порівнянні із сертифікованими лікарями), не допустивши помилок високої тяжкості в призначенні ліків.

Що правдоподібно, але не доведено: Що ця можливість принесе користь реальним недиференційованим пацієнтам; що перевага в точності відображає краще міркування, а не більшу кількість призначених тестів, узгодження з історичною карткою чи запам'ятовані публічні дані.

Чого вона не показує: Що MIRA працює за межами восьми діагнозів; що її безпечно розгортати; що вона перемагає лікарів у справедливому порівнянні за однакової інформації; що вона замінює клініцистів; що результати гарантовано вільні від забруднення навчальними даними.

Головні обмеження: Ретроспективне, симульоване й лише текстове оцінювання; вісім наперед обраних діагнозів; асиметрія інформації (значно більше аналізів крові — близько 51% доступних показників проти 28%, причому йдеться про дешеві аналізи крові, а не дорожчу візуалізацію); показники лікування частково оцінювали за історичною карткою; публічний бенчмарк MIMIC-IV, який мовна модель могла бачити під час навчання — і самі автори називають це можливою верхньою межею результативності.

Якої впевненості заслуговує висновок для загального читача? Високої, що MIRA є реальною й новою демонстрацією можливості: це агент, який *діє* по всій картці, а не просто чатбот. Низької щодо твердження, що він *кращий діагност за лікаря*: це обмежено одним ретроспективним бенчмарком і змішане з ефектом більшої кількості тестів, оцінюванням за карткою та можливим забрудненням даних. Власний висновок авторів тут найкращий: перш ніж це щось означатиме для допомоги пацієнтам, потрібне проспективне дослідження в реальному світі.

Джерела

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