The Clean Paper

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

Agent AI samodzielnie prowadził całe przypadki pacjentów — w symulatorze, na dawnych dokumentacjach

MIRA to nowy rodzaj medycznej AI: zamiast odpowiadać na pojedyncze pytanie, prowadzi cały przypadek w symulowanej dokumentacji szpitalnej — zbiera wywiad, zleca i odczytuje badania, ustala rozpoznanie i zapisuje zlecenia. W 574 retrospektywnych przypadkach z publicznej bazy danych, obejmujących osiem wcześniej wybranych diagnoz, według autorów osiągnęła większą trafność diagnostyczną niż lekarze, a jej decyzje były w znacznej mierze zgodne z wytycznymi i bezpieczne pod względem leków. Każde zastrzeżenie w tym zdaniu ma znaczenie: system działał w środowisku testowym, na dawnych dokumentacjach i wyłącznie na tekście; znaczna część przewagi dotyczyła chorób z jednoznacznymi wynikami badań; zlecał około dwa razy więcej badań krwi niż lekarze; kilka wyników oceniano zaś względem tego, co zapisano w pierwotnej dokumentacji. Rzeczywistym postępem jest agent działający w całym procesie pracy — nie dowód, że maszyna diagnozuje teraz lepiej od lekarza. Sami autorzy podkreślają, że uogólnienie wyników, bezpieczeństwo i nadzór nadal wymagają prospektywnych badań w rzeczywistym świecie.

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

Odpowiedź na pytanie to nie to samo co prowadzenie całego przypadku

Większość historii o medycznej AI dotyczy modelu, który *odpowiada*. Dostaje opis przypadku albo pytanie egzaminacyjne, zwraca diagnozę lub akapit porady i osiąga dobry wynik. To użyteczne, ale wąskie zastosowanie: błyskotliwy konsultant, który nigdy nie pracuje z dokumentacją.

MIRA powstała po to, by robić coś innego, i właśnie ta różnica jest prawdziwą nowością. Zamiast odpowiadać na jedno pytanie prowadzi cały przypadek: czyta kartę pacjenta, ustala, jakich danych z wywiadu nadal potrzebuje, zleca badania laboratoryjne, obrazowe i mikrobiologiczne, odczytuje wyniki, zawęża diagnostykę różnicową, a następnie zapisuje wynikające z niej zlecenia — recepty, przyjęcie do szpitala, skierowanie na operację. Robi to, podejmując działania *wewnątrz* elektronicznej dokumentacji medycznej, tak jak lekarz, zamiast generować swobodny tekst, który człowiek musi przepisać.

To rzeczywisty krok: od systemu doradzającego do systemu działającego w całym toku pracy. Jest to także dokładnie ten rodzaj postępu, który w przekazie medialnym spłaszcza się do hasła „AI przewyższa lekarzy". Warto więc precyzyjnie powiedzieć, co i gdzie sprawdzono.

Wersja jednozdaniowa: MIRA samodzielnie prowadziła całe przypadki i w tym benchmarku przewyższyła lekarzy pod względem trafności diagnozy — ale robiła to w środowisku izolowanym, na danych retrospektywnych, w obrębie ośmiu z góry wybranych rozpoznań, komunikując się wyłącznie tekstowo, a dużą część przewagi osiągnęła w chorobach dających najbardziej jednoznaczne wyniki badań. Postęp jest rzeczywisty. Tabela wyników nie jest kliniką.

MIRA pokazała sprawność w procesie pracy, nie wynik kliniczny

Środowisko testowym pozwala agentowi przejść przez cały retrospektywny przypadek. To nadal nie jest badanie z udziałem rzeczywistych pacjentów.

Rysunek oddziela wynik dotyczący procesu pracy od twierdzeń o wdrożeniu.

MIRA zademonstrowała zdolność prowadzenia całego procesu w izolowanym systemie elektronicznej dokumentacji medycznej. Nie wykazała gotowości do rzeczywistego zastosowania klinicznego, szerokiego zakresu diagnostycznego ani sprawiedliwego porównania z lekarzami dysponującymi równie obszernymi informacjami.

Co zrobili autorzy

MIRA — Medical Intelligence for Reasoning and Action — jest autonomicznym agentem opartym na modelach OpenAI: część prowadząca rozmowę i wykonująca działania korzysta z **GPT-4o**, a osobny etap planowania używa modelu rozumującego **o1**. Modele działają w specjalnie przygotowanym środowisku zgodnym ze standardami, zbudowanym na HL7 FHIR — standardzie danych używanym przez szpitale do wymiany dokumentacji — i dysponują zestawem ponad osiemdziesięciu tysięcy możliwych działań klinicznych. W tym środowisku MIRA może zbierać wywiad pacjenta, zlecać i interpretować badania laboratoryjne, obrazowe i mikrobiologiczne, tworzyć diagnostykę różnicową oraz układać plany leczenia — przepisywać leki, planować zabiegi chirurgiczne i przyjęcia. Jeden szczegół warto zaznaczyć od początku: automatycznymi „sędziami" oceniającymi odpowiedzi MIRA były również modele GPT-4o — ta sama rodzina, którą testowano. Po zwróceniu uwagi na ten problem przez recenzenta autorzy uzupełnili ocenę o przegląd przeprowadzony przez lekarza specjalistę.

Zestaw testowy pochodził z **MIMIC-IV**, dużej, publicznie dostępnej bazy zanonimizowanej elektronicznej dokumentacji medycznej. Spośród około 300 000 pacjentów leczonych w Beth Israel Deaconess Medical Center w latach 2008–2019 autorzy wybrali **574 przypadki** obejmujące **osiem docelowych rozpoznań** — schorzenia jamy brzusznej i stany nagłe z zakresu chorób wewnętrznych, między innymi zapalenie wyrostka robaczkowego, zapalenie trzustki, zapalenie płuc i zakażenie układu moczowego. W każdym przypadku MIRA zaczynała od ograniczonego obrazu i musiała krok po kroku decydować, co zrobić dalej — podobnie jak w prawdziwej diagnostyce, lecz na podstawie dokumentacji, której rzeczywiste zakończenie było już zapisane.

Następnie jej wyniki porównano z wynikami lekarzy w tych samych przypadkach.

> **Co naprawdę znaczy „autonomiczny agent w izolowanym systemie dokumentacji"**
>
> Te trzy określenia niosą bardzo dużo treści, dlatego warto je rozpakować.
> **Agent** oznacza, że system nie jest chatbotem odpowiadającym na polecenie. Działa w pętli: sprawdza bieżący stan, wybiera działanie (zleć to badanie, zapytaj o ten element wywiadu), widzi wynik i wybiera następne działanie — aż dochodzi do diagnozy i planu. „Inteligencją" jest model językowy; *sprawczość* zapewnia otaczająca go infrastruktura, która zamienia tekst modelu w dozwolone operacje na dokumentacji i przekazuje mu ich wyniki.
> **Izolowany system elektronicznej dokumentacji medycznej** oznacza symulowaną, odgrodzoną kopię takiego systemu, a nie system działającego szpitala. MIRA może „zlecić" badanie i otrzymać wynik, który faktycznie uzyskano u danego pacjenta, ponieważ przypadek jest historyczny, a odpowiedź już znajduje się w dokumentacji. Żadne jej działanie nie wpływa na

prawdziwego pacjenta ani klinikę.
Autonomiczny oznacza, że system prowadzi przypadek od początku do końca bez udziału człowieka — *wewnątrz środowiska izolowanego*. Nie oznacza to, że pozwolono mu samodzielnie opiekować się pacjentami bez nadzoru. To bardzo różne twierdzenia i sprawdzono tylko pierwsze.
Jeszcze jedna rzecz o środowisku izolowanym, które łatwo wyobrazić sobie błędnie: MIRA może *zlecić* dowolne badanie, ale system zwróci wynik tylko wtedy, gdy badanie to **rzeczywiście wykonano** u pacjenta. Jeśli poprosi o panel krwi, który pacjent miał zrobiony, dostanie prawdziwe wartości; jeśli poprosi o badanie obrazowe, którego nikt nie zlecił, otrzyma „N/D — nie można było wykonać", nigdy wymyśloną liczbę. Z wywiadem jest podobnie: jeżeli dokumentacja nie zawiera odpowiedzi na pytanie MIRA, symulowany pacjent mówi po prostu, że jej nie zna. MIRA nie może więc przywołać rozstrzygającego badania, którego nigdy nie wykonano — może korzystać wyłącznie z tego, co znalazło się w rzeczywistym toku diagnostycznym.
Rozróżnienie jest ważne, ponieważ słowo z nagłówków — „autonomiczny" — najłatwiej odczytać jako to drugie znaczenie.

Co odkryli

W benchmarku **MIRA przewyższyła lekarzy pod względem trafności diagnozy** — ale nagłówek ukrywa trzy różne liczby, które łatwo ze sobą pomieszać, dlatego trzeba je rozdzielić.

Samodzielnie, we wszystkich **574 przypadkach**, MIRA wskazała właściwą diagnozę w **88,9%** przypadków — przy czym punktem odniesienia było rozpoznanie przy wypisie zapisane dla każdego pacjenta w MIMIC-IV. W bezpośrednim porównaniu z ludźmi lekarze nie analizowali wszystkich 574 przypadków, lecz wspólny **podzbiór 311 przypadków**, korzystając z tych samych zapisów i narzędzi co MIRA. W tych samych 311 przypadkach MIRA osiągnęła **87,8%** i została porównana z dwiema osobno zrekrutowanymi grupami. Pierwszą była **grupa starszych stażem lekarzy**: czterech specjalistów z 7–11-letnim doświadczeniem, którzy uzyskali **78,1%**. Drugą była **grupa o mieszanym stażu**: głównie młodsi rezydenci oraz dwóch specjalistów, a więc skład bliższy rzeczywistej obsadzie oddziału ratunkowego; uzyskała **71,1%**. Te same przypadki, te same narzędzia — kolejność była następująca: najpierw AI, potem doświadczeni lekarze, na końcu zespół z przewagą młodszych. Ostrożne sformułowanie samej pracy mówi, że we wszystkich ośmiu chorobach MIRA „konsekwentnie dorównywała, a często przewyższała" lekarzy.

Jej późniejsze decyzje również wypadały dobrze: w dużej mierze były zgodne z wytycznymi, bezpieczne pod względem leków i właściwe w sprawie przyjęcia do szpitala. Autorzy nie odnotowali poważnych błędów lekowych w pięciu kategoriach bezpieczeństwa (interakcje leków, dawkowanie przy zaburzeniach nerek, alergie, ryzyko wydłużenia QT i przepisywanie opioidów), a recepty były poprawne w 467 z 468 przypadków — zaznaczają jednak, że system „nie osiągnął stuprocentowej niezawodności". Przyjęty dosłownie wynik robi wrażenie: agent prowadzi cały przypadek i kończy z trafniejszą diagnozą niż lekarze.

Jednak *charakter* zwycięstwa jest równie ważny jak samo zwycięstwo. Niezależni specjaliści komentujący pracę wskazali, skąd naprawdę pochodziła przewaga. W panelu ekspertów Science Media Centre dr Wei Xing z University of Sheffield zauważył, że główny wynik — przewaga AI nad lekarzami w trafności diagnozy — był **w dużej mierze napędzany przez choroby dające jednoznaczne wyniki badań**, takie jak zapalenie wyrostka robaczkowego i zapalenie trzustki, gdzie rozstrzygający obraz albo parametr laboratoryjny przesądza odpowiedź. W zapaleniu płuc i zakażeniach układu moczowego — dwóch z najczęstszych powodów zgłaszania się na oddział ratunkowy — *zarówno* AI, jak i lekarze wypadli najgorzej, a różnica między nimi była najmniejsza.

Sposób działania obu stron był również asymetryczny, co widać w liczbach z samej pracy. MIRA mocniej polegała na laboratorium: wykorzystywała około **51%** analitów laboratoryjnych dostępnych w rutynowej opiece, wobec około **28%** w przypadku lekarzy specjalistów — mediana wynosiła siedem dodatkowych parametrów krwi na przypadek. Autorzy ostrożnie zaznaczają, że jest to poziom *niższy* niż wyjściowy poziom rutynowej opieki w zbiorze danych, a nie strategia „zlecić wszystko", i nie odnotowują systematycznego wzrostu liczby droższych przekrojowych badań obrazowych. Mimo to większa ilość informacji może sama w sobie podnosić trafność diagnozy — nie jest to więc całkiem równorzędny pojedynek graczy dysponujących taką samą wiedzą, co Xing bezpośrednio wskazał. Lekarz, który mógłby zlecać każde tanie badanie krwi bez uwzględniania kosztu, dyskomfortu czy opóźnienia, również wyglądałby na papierze lepiej.

Dlaczego „pokonała lekarzy" nie znaczy „jest lepszym lekarzem"

Zwycięstwo w benchmarku łatwo nadinterpretować. Między „MIRA uzyskała wyższy wynik" a „MIRA jest lepsza" stoją trzy kwestie.

Po pierwsze, **względem czego ją oceniano**. Profesor Julie Jacko z University of Edinburgh zauważyła, że kilka kluczowych wyników MIRA zdefiniowano względem dokumentacji w bazowym zbiorze danych — system jest więc nagradzany za **odtwarzanie zapisanego postępowania klinicznego**, a niekoniecznie za wykazanie optymalnej opieki. Historyczna karta staje się kluczem odpowiedzi. To rozsądny sposób budowy benchmarku, ale mierzy zgodność z tym, co zrobiono, nie bezwzględną poprawność. Oddając pracy sprawiedliwość, problem najmocniej dotyczy wskaźników *zgodności leczenia* — czy zlecenia MIRA odpowiadały karcie? — podczas gdy główną trafność diagnozy oceniano względem rozpoznania przy wypisie, które jest bliższe rzeczywistemu wynikowi niż echo dokumentacji.

Po drugie, wspomniana już **asymetria informacji**: znacznie więcej badań krwi (około 51% dostępnych analitów wobec 28%) oznacza więcej dowodów. Część różnicy trafności została prawdopodobnie *kupiona*, a nie wywnioskowana.

Po trzecie, **miejsce działania**. Było to środowisko izolowane, retrospektywne przypadki, osiem wybranych wcześniej rozpoznań i wyłącznie tekst. Jak ujął to dr Dominic Oliver z University of Oxford, prawdziwa ocena kliniczna zależy nie tylko od tego, co pacjenci mówią, lecz także jak to mówią — wraz z badaniem fizykalnym, obserwowanym zachowaniem i mową ciała, których agent tekstowy czytający zamkniętą dokumentację nigdy nie widzi. Pacjent, którego problem nie należy

do jednego z ośmiu wybranych rozpoznań, po prostu leży poza zakresem wniosków tego badania.

Recenzenci zgłosili też cichsze zastrzeżenie: **zanieczyszczenie danych treningowych**. MIMIC-IV jest publiczna i szeroko omawiana, więc model językowy trenowany na otwartym internecie mógł już widzieć artykuły, dyskusje przypadków albo same dane. Dr Midhun Parakkal Unni z University of Sheffield wskazał to wprost — jeśli część odpowiedzi znajdowała się w danych treningowych, część wyniku jest pamięcią, nie rozumowaniem klinicznym, a rozdzielić jedno od drugiego może tylko niezależne powtórzenie. Autorzy nie bagatelizują tego problemu: piszą, że ich wyniki „można ostrożnie interpretować jako możliwą górną granicę" i że „mogą zawyżać zdolność generalizacji na inne publiczne przypadki" — praca sama nakłada więc sufit na swój nagłówek.

Żadne z tych zastrzeżeń nie odbiera MIRA znaczenia. Wskazuje właściwy, wyważony odczyt: w retrospektywnym benchmarku agent zdolny działać w całej dokumentacji przewyższył lekarzy przy rozpoznaniach z jednoznacznymi badaniami — częściowo dzięki zlecaniu większej liczby badań, częściowo dzięki zgodności z zapisaną kartą. To rzeczywista demonstracja możliwości, a nie werdykt, że maszyny diagnozują obecnie lepiej od lekarzy.

Czego to *nie dowodzi*

- Badanie **nie pokazuje**, że MIRA działa na prawdziwych pacjentach. Każdy przypadek był retrospektywny i symulowany; system nigdy nie prowadził rzeczywistego pacjenta.
- **Nie pokazuje**, że działa poza ośmioma rozpoznaniami. Nie testowano chorób spoza wcześniej wybranego zestawu — nieuporządkowanej, niejednoznacznej większości medycyny.
- **Nie pokazuje**, że można ją bezpiecznie wdrożyć. Sami autorzy piszą, że generalizacja, bezpieczeństwo i nadzór nadal wymagają prospektywnych badań w rzeczywistym świecie.
- **Nie ustanawia** sprawiedliwego porównania z lekarzami. MIRA wykorzystywała znacznie więcej badań krwi (około 51% dostępnych analitów wobec 28%), a kilka wyników dotyczących leczenia oceniano częściowo względem zapisanej dokumentacji, nie rzeczywiście najlepszego postępowania.
- **Nie pokazuje**, że wynik jest wolny od zapamiętywania. Publiczne dane MIMIC-IV mogą nakładać się na dane treningowe modelu; wykluczenie tego wymagałoby niezależnego powtórzenia.
- **Nie oznacza**, że „AI zastępuje lekarzy". To wspomaganie decyzji zdolne *działać* w całej dokumentacji; zgodnie z opinią recenzentów rzeczywiste użycie będzie polegało na współpracy z klinicystami, którzy zachowują decyzyjność i dostarczają wszystkiego, czego zapis tekstowy nie obejmuje.

Jak mocne są dowody?

Twierdzenie trzeba podzielić na dwie części, ponieważ dowody dla każdej z nich są bardzo różne.

Jako demonstracja możliwości — że agent oparty na modelu językowym potrafi poprowadzić cały przypadek kliniczny w izolowanym systemie dokumentacji, łącząc wywiad, badania, diagnozę i zlecenia od początku do końca — praca jest rzeczywiście nowatorska i dość przekonująca. To właśnie jest

nowe i warte uwagi.

Jako twierdzenie o wyższości — że MIRA jest *lepsza od lekarzy* — dowody mają wyraźne granice i wymagają ostrożności. Wynik dotyczy konkretnego retrospektywnego benchmarku, ośmiu rozpoznań, asymetrii informacji (więcej badań) i klucza odpowiedzi pochodzącego z historycznej dokumentacji, przy realnej możliwości zanieczyszczenia danych treningowych. Wystarczy to, by powiedzieć: „agent wypadł imponująco w tym benchmarku". Nie wystarczy, by stwierdzić: „agent jest lepszym diagnostą niż lekarz", czego zresztą autorzy nie twierdzą.

Najbardziej użyteczne stanowisko nie brzmi ani „AI pokonuje lekarzy", ani „to tylko zabawka". Brzmi ono: nowy rodzaj medycznej AI — działający w całej dokumentacji zamiast odpowiadać na pytania — dobrze poradził sobie w trudnym retrospektywnym benchmarku i teraz musi wykazać swoją wartość tam, gdzie jeszcze go nie próbowano: prospektywnie, u prawdziwych, niewyselekcjonowanych pacjentów i pod nadzorem.

Dlaczego to ma znaczenie

W ostatnich latach interesujące pytanie w medycznej AI po cichu się zmieniło. Kiedyś brzmiało: „czy model potrafi postawić właściwą diagnozę?" — i na coraz czystszych zbiorach testowych modele coraz częściej odpowiadały „tak". MIRA oznacza przejście do trudniejszego pytania: „czy model potrafi *wykonać pracę* — zebrać dane, zlecić, zinterpretować, zdecydować i działać — w całym toku postępowania?". To pytanie jest bardziej użyteczne i uczciwsze, bo właśnie w działaniu mieszczą się zarówno trudność, jak i ryzyko.

Dlatego sposób przedstawienia wyniku ma znaczenie. Odruchowy nagłówek — *AI przewyższa lekarzy* — wskazuje najmniej nowatorską i najsłabiej podpartą część wyniku. Prawdziwa nowość jest mniejsza, lecz ważniejsza: agent potrafiący przejść przez całą dokumentację. Jeśli ta zdolność się potwierdzi, zmieni **tok pracy** na długo przed tym, zanim zmieni osobę, która nim kieruje. Lekarz nie zostaje zastąpiony; pierwszym miejscem działania takiego narzędzia będą zlecenia, interpretacja i pilnowanie wyników — cały proces.

Praca ustawia też ciężar dowodu we właściwym kierunku. Zwycięstwo w rankingu retrospektywnych przypadków jest powodem, by przeprowadzić badanie prospektywne, a nie jego zamiennikiem. Autorzy mówią to wprost. Uczciwy odczyt MIRA jest zaproszeniem do następnego badania — według standardu dobrze znanego tej dziedzinie: na pacjentach, nie na benchmarkach.

Zwięzłe podsumowanie

MIRA jest autonomicznym agentem AI działającym w izolowanym systemie elektronicznej dokumentacji medycznej: może zebrać wywiad, zlecać i interpretować badania laboratoryjne, obrazowe i mikrobiologiczne, stawiać diagnozę oraz zapisywać plany leczenia. W teście obejmującym 574 retrospektywne przypadki z publicznej bazy MIMIC-IV i osiem wcześniej wybranych rozpoznań

przewyższyła lekarzy pod względem trafności diagnozy (88,9% ogółem; w bezpośrednim porównaniu 87,8% wobec 78,1%) i podejmowała decyzje w dużej mierze zgodne z wytycznymi oraz bezpieczne pod względem leków. Ocena była jednak symulacją opartą na dawnych zapisach i wyłącznie tekście; duża część przewagi dotyczyła chorób z jednoznacznymi wynikami badań; MIRA korzystała ze znacznie większej liczby badań krwi niż lekarze (około 51% dostępnych analitów wobec 28%); kilka wyników leczenia oceniano względem tego, co zapisano w oryginalnej dokumentacji; a publiczny zbiór stwarza realne ryzyko zanieczyszczenia danych treningowych — sami autorzy nazywają swoje liczby możliwą górną granicą. Prawdziwym postępem jest agent działający w całym toku pracy, zamiast odpowiadać na pojedyncze pytania. Autorzy wyraźnie zaznaczają, że generalizacja, bezpieczeństwo i nadzór nadal wymagają prospektywnych badań w rzeczywistym świecie. To właściwy odczyt: imponująca demonstracja możliwości, nie dowód, że AI diagnozuje lepiej od lekarzy, ani system gotowy do prawdziwej kliniki.

Kontrola bez upiększeń

Co pokazuje praca: Agent oparty na modelu językowym (MIRA) potrafi przeprowadzić cały przypadek kliniczny od początku do końca w izolowanym systemie dokumentacji — wywiad, badania, diagnozę i zlecenia — a w retrospektywnym benchmarku 574 przypadków i ośmiu rozpoznań przewyższył lekarzy pod względem trafności diagnozy (88,9% ogółem; 87,8% wobec 78,1% w porównaniu z lekarzami specjalistami), bez poważnych błędów lekowych.

Co jest prawdopodobne, ale nieudowodnione: Że zdolność ta przynosi korzyść prawdziwym, niewyselekcjonowanym pacjentom; że przewaga trafności odzwierciedla lepsze rozumowanie, a nie większą liczbę zleconych badań, zgodność z zapisaną dokumentacją lub zapamiętane dane publiczne.

Czego praca nie pokazuje: Że MIRA działa poza ośmioma rozpoznaniami; że można ją bezpiecznie wdrożyć; że pokonuje lekarzy w sprawiedliwym porównaniu stron mających te same informacje; że zastępuje klinicystów; że wyniki są wolne od zanieczyszczenia danych treningowych.

Główne ograniczenia: Retrospektywna, symulowana i wyłącznie tekstowa ocena; osiem wcześniej wybranych rozpoznań; asymetria informacji (znacznie więcej badań krwi — około 51% dostępnych analitów wobec 28%, przy czym chodziło o tanie badania krwi, nie obrazowanie); wyniki leczenia częściowo oceniane względem historycznej dokumentacji; oraz publiczny benchmark MIMIC-IV, który model językowy mógł widzieć podczas treningu — co autorzy sami wskazują jako możliwą górną granicę wyników.

Jak duże zaufanie powinien mieć czytelnik niespecjalista? Wysokie do wniosku, że MIRA jest rzeczywistą i nowatorską demonstracją możliwości — agentem, który *działa* w całej dokumentacji, a nie tylko chatbotem. Niskie do twierdzenia, że jest *lepszym diagnostą niż lekarz*: wynik ogranicza się do jednego retrospektywnego benchmarku i jest zakłócony przez liczbę zlecanych badań, ocenę względem karty oraz możliwe zanieczyszczenie danych. Końcowy wniosek samych autorów jest właściwy — zanim wynik zacznie cokolwiek znaczyć dla opieki nad pacjentem, potrzebne jest prospektywne badanie w rzeczywistym świecie.

Źródła

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