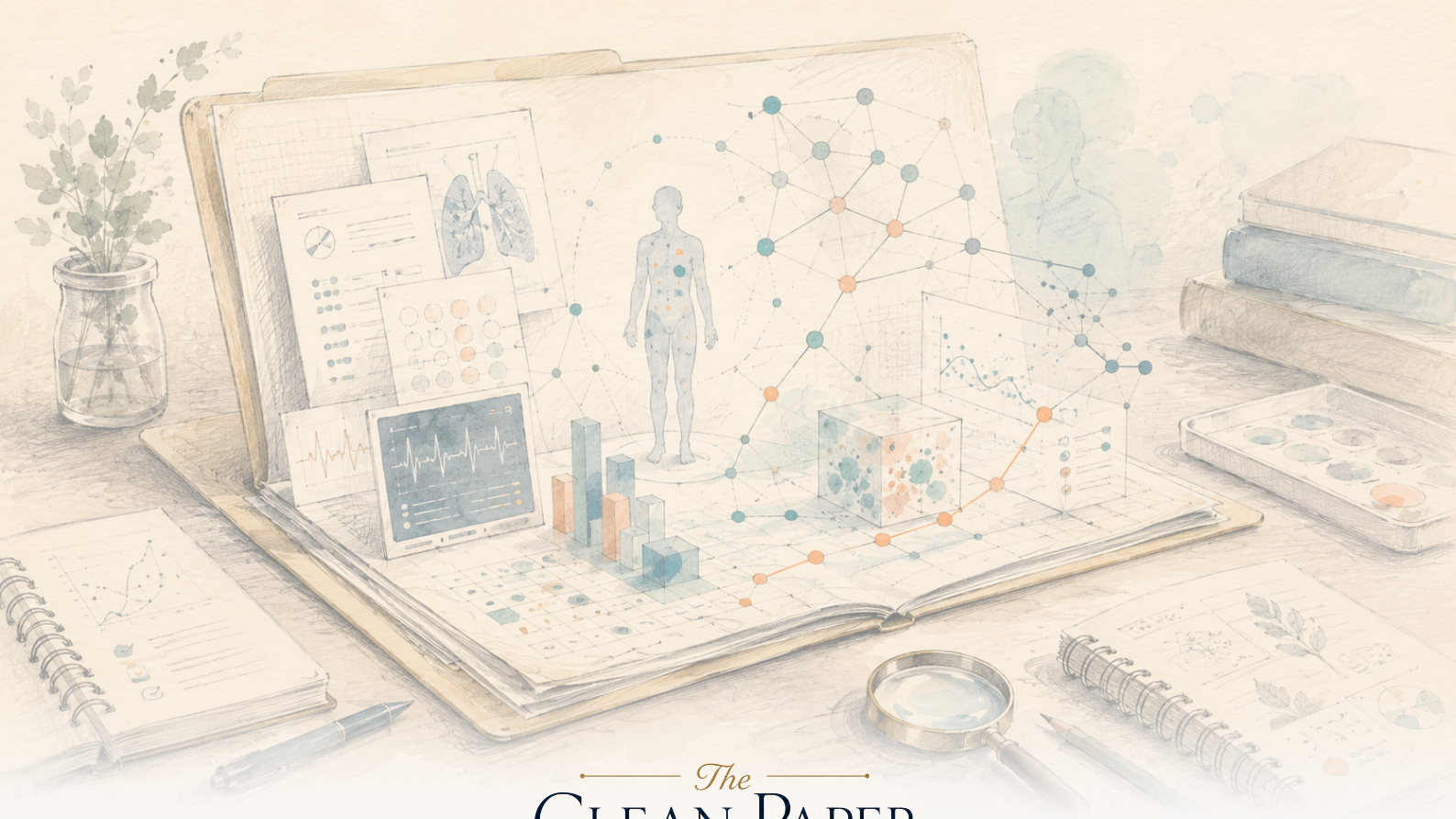

The Clean Paper

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

Agent AI samostatně zpracovával celé případy pacientů — v simulátoru nad staršími záznamy

MIRA je nový druh lékařské AI: místo odpovědi na jedinou otázku zpracuje celý případ v simulovaném nemocničním záznamu — odebere anamnézu, objedná a vyhodnotí vyšetření, dospěje k diagnóze a zapíše pokyny. U 574 retrospektivních případů z veřejné databáze, zahrnujících osm předem vybraných diagnóz, podle autorů překonal lékaře v diagnostické přesnosti a rozhodoval převážně v souladu s doporučeními a bezpečně z hlediska medikace. Každá výhrada v této větě je důležitá: systém běžel v testovacím prostředí, na minulých záznamech a pouze s textem; velká část jeho náskoku pocházela z onemocnění s jednoznačnými výsledky vyšetření; objednal zhruba dvakrát více krevních testů než lékaři; a několik výsledků se hodnotilo podle toho, co obsahovala původní dokumentace. Skutečným pokrokem je agent, který pracuje napříč celým postupem — nikoli důkaz, že stroj nyní diagnostikuje lépe než lékař. Autoři sami uvádějí, že zobecnitelnost, bezpečnost a řízení stále vyžadují prospektivní studie v reálném světě.

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

Odpověď na otázku není totéž jako vést případ

Většina příběhů o lékařské AI je o modelu, který *odpovídá*. Dáte mu krátký popis případu nebo zkušební otázku, on vám vrátí diagnózu nebo odstavec rady a výsledek má dobré hodnocení. Je to užitečné, ale omezené: chytrý konzultant, který se samotnou dokumentací nepracuje.

MIRA je navržena k něčemu jinému a rozdíl spočívá ve skutečném jednání. Místo odpovědi na jedinou otázku zpracuje celý případ: přečte pacientovu dokumentaci, rozhodne, jakou anamnézu ještě potřebuje, objedná laboratorní, zobrazovací a mikrobiologické testy, přečte výsledky, zúží diferenciální diagnózu a poté zapíše potřebné pokyny — předpisy, hospitalizaci, doporučení na operaci. Dělá to tak, že provádí akce *uvnitř* elektronické zdravotní dokumentace, jak to dělá klinik, místo aby vytvářela volný text, který by člověk teprve musel převést do praxe.

To je skutečný krok: od systému, který radí, k systému, který prochází celým pracovním postupem. Je to také přesně ten typ kroku, který média zjednodušují na "AI překonává lékaře." Proto stojí za to přesně říci, co bylo testováno a kde.

Jednovětá verze: MIRA zvládla celé případy samostatně a podle tohoto měřítka překonala lékaře v diagnostické přesnosti — ale dělala to v testovacím prostředí, na retrospektivních záznamech, napříč osmi předem vybranými diagnózami, komunikovala pouze textově a velkou část své výhody získala u diagnóz s nejjednoznačnějšími výsledky vyšetření. Pokrok je skutečný. Výsledek benchmarku ale není klinická praxe.

MIRA prokázala schopnost projít celý pracovní postup v testovacím EHR. Studie neprokázala skutečné klinické nasazení, široké diagnostické pokrytí ani spravedlivý a stejně informovaný přímý souboj s lékaři.

Co autoři udělali

MIRA — Medical Intelligence for Reasoning and Action — je autonomní agent postavený na modelech OpenAI: část, která komunikuje a jedná, běží na **GPT-4o**, a samostatný plánovací krok využívá model uvažování OpenAI **o1**. Tyto modely pracují ve vlastním rámci založeném na standardu HL7 FHIR, který nemocnice používají k výměně zdravotních záznamů, a mají k dispozici více než osmdesát tisíc možných klinických úkonů. V tomto testovacím prostředí může MIRA získat anamnézu pacienta, objednat a interpretovat laboratorní testy, zobrazovací a mikrobiologické vyšetření, vytvořit diferenciální diagnózu a formulovat léčebné plány — předepisovat léky, plánovat chirurgické zákroky, plánovat přijetí. Jeden detail, který stojí za zmínku hned na začátku: automatizované hodnocení odpovědí MIRA prováděly modely GPT-4o — tedy stejná rodina modelů, která byla testována — a autoři jejich hodnocení doplnili posouzením atestovaného lékaře poté, co recenzent tuto obavu vznesl.

Testovací sada pocházela z **MIMIC-IV**, velké, veřejně dostupné databáze deidentifikovaných EHR. Z přibližně 300 000 pacientů léčených v Beth Israel Deaconess Medical Center v letech 2008 až 2019 autoři vybrali **574 případů** pokrývajících **osm cílových diagnóz** — akutní břišní a interní diagnózy, mezi nimiž jsou zánět slepého střeva, pankreatitida, zápal plic a infekce močových cest. Pro každý případ MIRA začínala s omezeným obrazem a musela krok za krokem rozhodnout, co udělat dál — podobně jako při skutečném vyšetření, ovšem nad záznamem, jehož skutečný výsledek už je znám.

Výkon systému byl následně porovnán s lékaři u stejných případů.

Co vlastně znamená „autonomní agent v testovacím EHR“

Tři slova zde dělají hodně práce, takže stojí za to je rozebrat.
Agent znamená, že systém není chatbot, který pouze odpovídá na zadání. Pracuje v opakované smyčce: podívej se na aktuální stav, vyber akci (objednej toto vyšetření, vyžádej si tuto část anamnézy), vidíš výsledek, vyber další akci — dokud nedojde k diagnóze a plánu. „Inteligence“ je jazykový model; *agentní rámec* je podpůrná vrstva kolem něj, která převádí text modelu na povolené operace EHR a výsledky vkládá zpět.
Testovací EHR znamená simulovanou, uzavřenou kopii systému lékařských záznamů, nikoli živý nemocniční systém. MIRA může “objednat” test a obdržet výsledek, který pacient skutečně dostal, protože případ je historický a odpověď je již zaznamenána. Nic, co dělá, se nedotkne skutečného pacienta ani skutečné kliniky.
Autonomní znamená, že dokončí případ od začátku do konce bez člověka ve smyčce — *uvnitř testovacího prostředí*. To neznamená, že byl vypuštěn k péči o pacienty bez dozoru. To jsou velmi odlišná tvrzení a testován byl pouze ten první.
Ještě jedna věc k testovacímu prostředí, protože je snadné si to špatně představit: MIRA může *objednat* jakýkoli test, který chce, ale systém může vrátit výsledek pouze tehdy, pokud byl tento test **skutečně proveden** u tohoto pacienta v reálném životě. Požádejte o krevní test, který pacient podstoupil, a dostanete skutečné hodnoty; požádáte o sken, který nikdo

> neobjednal, a dostanete odpověď "N/A — nebylo možné provést", nikdy nevymyšlené číslo. Anamnéza funguje stejně: pokud záznam neobsahuje to, na co se MIRA ptá, simulovaný pacient jednoduše řekne, že to neví. MIRA tedy nemůže vyvolat jediný rozhodující test, který nikdy nebyl proveden — může fungovat pouze s tím, co zahrnovala skutečná péče.
> Tento rozdíl je důležitý, protože klíčové slovo — „autonomní" — je to, které bude nejčastěji čteno jako druhá věc.

Co našli

Na benchmarku **MIRA překonala lékaře v diagnostické přesnosti** — ale titulek skrývá tři různé čísla, která se snadno rozmazávají, takže stojí za to je oddělit.

Samostatně, ve všech **574 případech**, MIRA uvedla správnou diagnózu v **88,9 %** případů — hodnotila se proti diagnóze při propuštění, která byla skutečně zaznamenána u každého pacienta v MIMIC-IV. U přímého souboje s lidmi lékaři nevyšetřili všech 574 případů; pracovali na sdílené podmnožině **311 případů**, a používali stejné záznamy a nástroje, jaké měla MIRA. Ve stejných 311 případech dosáhla MIRA skóre **87,8 %** a porovnávala se se dvěma zvlášť sestavenými skupinami lékařů. První byla **skupina zkušených lékařů**: čtyři lékaři s certifikací a 7 až 11 lety praxe, kteří dosáhli **78,1 %**. Druhá byla **smíšená skupina podle zkušenosti**: převážně mladší rezidenti plus dva specialisté, blíže tomu, jak je skutečné pohotovostní oddělení obsazeno, kteří dosáhli **71,1 %**. Stejné případy, stejné nástroje — a nejvyšší skóre měla AI, za ní zkušení lékaři a poté tým s převahou juniorů. Samotná práce pečlivě uvádí, že MIRA byla "konzistentně ekvivalentní a často převyšovala" lékaře ve všech osmi nemocech.

Obstála i její navazující rozhodnutí: většinou v souladu s pokyny, bezpečná z hlediska medikace a vhodná při přijetí. Autoři neuvádějí žádné závažné lékové chyby v pěti bezpečnostních kategoriích (interakce léků, dávkování ledvin, alergie, QT riziko a předepisování opioidů) a předpisy jsou správné ve 467 z 468 případů — přičemž poznamenávají, že systém "nedosáhl 100% spolehlivosti." Pokud tato čísla vezmeme doslova, je to pozoruhodný výsledek: agent vede celý případ a v diagnostické přesnosti předčí lékaře.

Ale *povaha* vítězství je stejně důležitá jako samotné vítězství. Nezávislí odborníci čtoucí článek ukázali, odkud vlastně ta výhoda pochází. V expertním panelu Science Media Centre Dr. Wei Xing (University of Sheffield) poznamenal, že hlavní číslo — AI překonávající lékaře v diagnostické přesnosti — bylo **většinou způsobeno onemocněními s jednoznačnými výsledky vyšetření**, jako je zánět slepého střeva a pankreatitida, kde rozhodné vyšetření nebo laboratorní hodnota rozhodne otázku. U zápalu plic a infekcí močových cest — dvou z nejčastějších důvodů, proč lidé skutečně přicházejí na pohotovost — *AI i lékaři* si vedli nejhůře a rozdíl mezi nimi byl nejmenší.

Byla zde také asymetrie v tom, jak obě strany hrály, a to se započítává i do čísel samotné práce. MIRA se více opírala o laboratorní vyšetření: čerpala z přibližně **51 %** laboratorních analytů dostupných v rutinní péči, oproti přibližně **28 %** u certifikovaných lékařů — což je medián o sedm parametrů krve na případ více. Autoři pečlivě rámují tuto situaci jako *pod* úrovní rutinní péče zaznamenané v

datové sadě, nikoli strategii „objednat všechno“, a uvádějí *žádný* systematický nárůst nákladnějšího průřezového zobrazování. Přesto více informací samo o sobě může vést k vyšší diagnostické přesnosti — takže nejde o souboj mezi stejně informovanými hráči, což Xing přímo označil. Také klinik, který si může bez omezení objednat levné krevní testy bez ohledu na náklady, nepohodlí či zpoždění, by na papíře vypadal přesnější.

Proč „překonala lékaře“ neznamená „je lepším lékařem“

Výsledek benchmarku lze snadno přecenit. Mezi „MIRA dosáhla vyššího skóre“ a „MIRA je lepší“ stojí tři důležité rozdíly.

Za prvé, **proti čemu byla hodnocena.** Profesorka Julie Jacko (University of Edinburgh) poznamenala, že několik klíčových výsledků MIRA je definováno vzhledem k tomu, co bylo zdokumentováno v základní datové sadě — což znamená, že systém je odměněn za **reprodukci zaznamenaného klinického chování**, nikoli nutně za prokázání optimální péče. Historická dokumentace se stává klíčem správných odpovědí. To je rozumný způsob, jak postavit měřítko, ale měří souhlas s tím, co bylo provedeno, nikoli správnost v absolutním smyslu. Abychom byli k práci spravedliví, nejvíce to zasahuje do metriky *zarovnání léčby* — odpovídaly pokyny MIRA záznamu? — zatímco přesnost hlavní diagnózy je hodnocena vůči diagnóze propuštění, která je blíže skutečnému výsledku než dokumentační ozvěna.

Za druhé, **asymetrie informací** již zmíněná: mnohem více krevních testů (asi 51 % dostupných analytů oproti 28 %) je více důkazů. Část rozdílu v přesnosti pravděpodobně plyne z většího množství dostupných dat, nikoli z lepšího uvažování.

Za třetí, **kde to běželo.** Šlo o testovací prostředí s retrospektivními případy, napříč osmi předem vybranými diagnózami, pouze v textu. Skutečné klinické hodnocení, jak to řekl Dr. Dominic Oliver (University of Oxford), závisí nejen na tom, co pacienti říkají, ale i na tom, jak to říkají — spolu s fyzickým vyšetřením, pozorovaným chováním a řečí těla, což žádný agent, který pouze textově čte hotový záznam, nikdy nevidí. A pacient, jehož potíže nespadají do jedné z osmi vybraných diagnóz, je jednoduše mimo to, co tato studie může vyjádřit.

Recenzenti také zmiňovali tišší obavu: **kontaminace dat.** MIMIC-IV je veřejný a hodně se o něm píše, takže jazykový model trénovaný na otevřeném internetu mohl již vidět články, případové diskuse nebo samotná data. Dr. Midhun Parakkal Unni (University of Sheffield) na to přímo upozornil — pokud byly některé odpovědi v tréninkových datech, část výkonu je paměť, nikoli klinické uvažování, a pouze nezávislá replikace může tyto dvě věci rozlišit. Autoři však tento bod nepřehlížejí: píší, že jejich výsledky “by mohly být opatrně interpretovány jako možná horní hranice” a “mohou přeceňovat zobecnění na jiné veřejné případy” — práce tak sama stanovuje horní mez svému titulkovému výsledku.

Nic z toho nečiní MIRA méně zajímavou. Jen to správně kalibruje hodnocení: podle retrospektivního měřítka agent, který mohl působit napříč celým záznamem, překonal lékaře u diagnóz s jednoznačnými výsledky vyšetření — částečně tím, že objednal více testů, částečně tím, že se shodoval se zaznamenanou dokumentací. To je skutečná ukázka schopností, ne verdikt, že stroje nyní diagnostikují lépe než lékaři.

Co to *ne*dokazuje*

- Neukazuje, že MIRA funguje u skutečných pacientů. Každý případ byl retrospektivní a simulovaný; žádný pacient nebyl tímto systémem léčen.
- Neukazuje to, že by fungovalo mimo osm diagnóz. Podmínky mimo předem vybranou sadu — chaotickou, nediferencovanou většinu medicíny — nebyly testovány.
- Neukazuje to, že je bezpečné nasadit. Autoři sami píší, že zobecňování, bezpečnost a správa stále potřebují prospektivní, reálné studie.
- **Nevytváří** spravedlivý přímý souboj s lékaři. Vycházelo z mnohem většího množství krevních testů (asi 51 % dostupných analytů oproti 28 %) a několik výsledků léčby bylo hodnoceno částečně podle zaznamenané dokumentace, nikoli podle prokázané nejlepší péče.
- Neznamená, že výsledek je bez zapamatování. Veřejná data MIMIC-IV se mohou překrývat s tréninkem modelu, což by nezávislá replikace musela vyloučit.
- Neznamená to "AI nahrazuje lékaře." Je to nástroj pro podporu rozhodování, který může *jednat* napříč záznamem; konsenzus recenzentů je, že skutečné využití bude ve spolupráci s kliniky, kteří si zachovávají autoritu a poskytují vše, co textový záznam vynechává.

Jak silné jsou důkazy?

Rozdělte tvrzení na dvě části, protože důkazy jsou pro každou polovinu velmi odlišné.

Jako demonstrace schopností — že agent založený na jazykovém modelu může vést celý klinický případ v testovacím EHR, řetězit anamnézu, testy, diagnózu a příkazy jeden za druhým — je tato práce skutečně nová a poměrně přesvědčivá. Tohle je ta nová část a stojí za to jí věnovat pozornost.

Jako tvrzení o nadřazenosti — že MIRA je *lepší než lékaři* — jsou důkazy omezené a měly by být čteny s opatrností. Drží se konkrétního retrospektivního benchmarku, osmi diagnóz, s informační asymetrií (více testů) a klíčem odpovědí odvozeným z historické dokumentace, a s reálnou možností kontaminace tréninkovými daty. To stačí k tomu, abychom řekli: "Agent na tomto benchmarku podal působivý výkon." Nestačí to k tvrzení "agent je lepší diagnostik než lékař" a autoři netvrdí druhou věc.

Nejužitečnější postoj není ani "AI poráží lékaře", ani "jen hračka". Je to tak: nový druh lékařské AI — taková, která *jedná* napříč záznamy místo odpovídání na otázky — uspěl na tvrdém retrospektivním měřítku a nyní se musí osvědčit tam, kde ještě nebyl vyzkoušen: prospektivně, na skutečných neroztříděných pacientech a pod řádným dohledem.

Proč je to důležité

Několik let se zajímavá otázka v lékařské AI tiše změnila. Dříve to bylo "dokáže model správně diagnostikovat?" — a modely stále odpovídaly ano, na stále čistších testovacích sadách. MIRA znamená přechod k těžší otázce: "Může model *udělat práci* — shromáždit, uspořádat, interpretovat, rozhodovat,

jednat — napříč celým pracovním postupem?" To je užitečnější a upřímnější otázka, protože právě jednání je místem, kde je zároveň obtížnost i riziko.

Proto záleží na rámování. Reflex v titulcích — *AI překonává lékaře* — ukazuje na nejméně novou a nejméně podloženou část výsledku. Opravdu nová věc je menší a zásadnější: agent, který může procházet celým záznamem. Tato schopnost, pokud obstojí, mění **pracovní postup** dávno předtím, než změní, kdo za něj nese odpovědnost. Lékař není nahrazen; takový nástroj se nejprve prosadí v pracovním postupu — při objednávání vyšetření, interpretaci a dohledávání výsledků.

A znovu se tím klade důkazní břemeno správným směrem. Vítězství v benchmarku v retrospektivních případech je důvodem k provedení prospektivní studie, nikoli jeho náhradou. Autoři to říkají přímo. Upřímné čtení MIRA je pozvánkou k další studii — držené podle standardu, který obor již zná: měřeno na pacientech, ne na benchmarkech.

Stručné shrnutí

MIRA je autonomní agent AI, který pracuje v testovacím elektronickém zdravotním záznamu: může pořizovat anamnézu, objednávat a interpretovat laboratorní testy, zobrazovací a mikrobiologické vyšetření, stanovit diagnózu a napsat léčebné plány. Při testu na 574 retrospektivních případech z veřejné databáze MIMIC-IV, u osmi předem vybraných diagnóz, překonala lékaře v diagnostické přesnosti (88,9 % celkově; 87,8 % oproti 78,1 % přímo) a činila převážně rozhodnutí v souladu s doporučeními, která byla bezpečná pro medikaci. Hodnocení však bylo pouze textovou simulací historických záznamů; velká část výhody se objevila u stavů s jednoznačnými výsledky vyšetření; MIRA využívala mnohem více krevních testů než lékaři (asi 51 % dostupných analytů oproti 28 %); několik výsledků léčby se hodnotilo podle původního záznamu; veřejná datová sada navíc představuje skutečné riziko kontaminace tréninkových dat — což autoři sami nazývají možnou horní hranicí svých čísel. Skutečný pokrok je agent, který jedná napříč celým pracovním postupem, místo aby odpovídal na izolované otázky. Autoři jasně říkají, že zobecnitelnost, bezpečnost a správa stále vyžadují prospektivní, reálné studie — což je správné čtení: působivá demonstrace schopností, nikoli důkaz, že AI diagnostikuje lépe než lékaři, a ne systém připravený na skutečnou kliniku.

Kontrola bez příkras

Co práce ukazuje: Jazykový model agenta (MIRA) může provést celý klinický případ od začátku do konce v testovacím EHR — anamnézu, testy, diagnózu, příkazy — a na retrospektivním benchmarku 574 případů napříč osmi diagnózami překonal lékaře v diagnostické přesnosti (88,9 % celkově; 87,8 % oproti 78,1 % u lékařů s certifikací) bez závažných chyb medikace.

Co je pravděpodobné, ale není prokázáno: Že tato schopnost se promítá do přínosu pro skutečné, nediferencované pacienty; že přesnost odráží lepší uvažování spíše než více objednaných testů, shodu se zaznamenanou dokumentací nebo zapamatovaná veřejná data.

Co neukazuje: Že MIRA funguje mimo svých osm diagnóz; že je bezpečné ji nasadit; že překonává lékaře ve spravedlivém, stejně informovaném srovnání; že nahrazuje kliniky; že výsledky jsou bez kontaminace tréninkovými daty.

Hlavní omezení: Retrospektivní, simulované, pouze textové hodnocení; osm předem vybraných diagnóz; informační asymetrie (mnohem více krevních testů — asi 51 % dostupných analytů oproti 28 % v levných krevních testech, nikoli zobrazovacích); výsledky léčby byly částečně hodnoceny podle historické dokumentace; a veřejný benchmark (MIMIC-IV), který mohl být součástí tréninku jazykového modelu — autoři ho označují jako možnou horní hranici výkonu.

Jakou důvěru by měl mít běžný čtenář? Vysokou v to, že MIRA je skutečná a nová demonstrace schopností — agent, který *jedná* napříč záznamem, ne jen chatbot. Nižší v tvrzení, že je *lepším diagnostikem než lékař*: toto tvrzení je vázáno na jeden retrospektivní benchmark a je zatíženo rozdílným objednáváním testů, skórováním proti záznamu a možnou kontaminací. Závěr autorů je správný — vyžaduje to prospektivní studie v reálném světě, než to bude mít význam pro péči o pacienty.

Zdroje

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