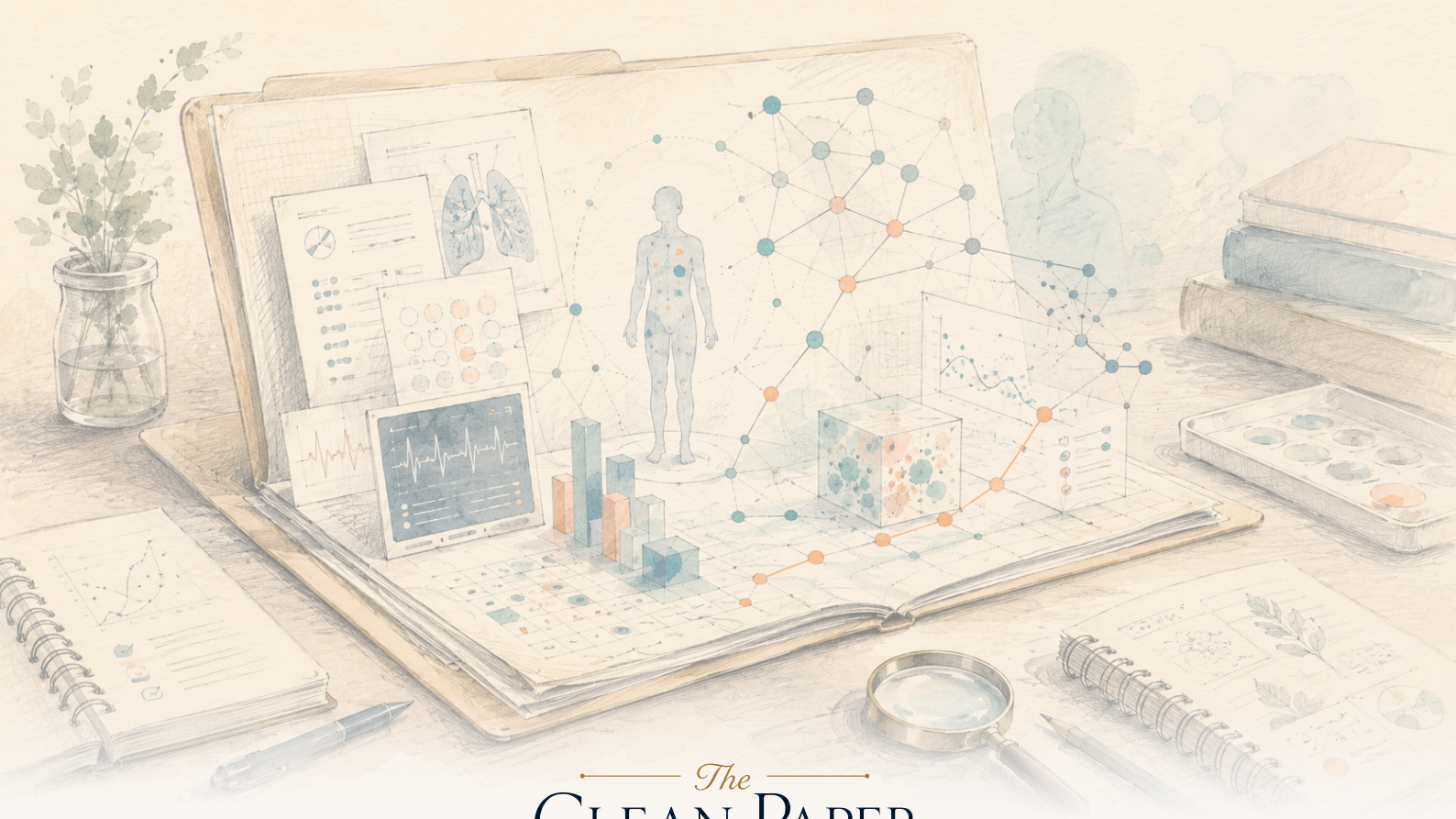


The Clean Paper

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

Agent AI samostatne spracoval celé prípady pacientov — v simulátore a z minulých záznamov

MIRA je nový druh medicínskej AI: namiesto odpovede na jedinú otázku spracuje celý prípad v simulovanom nemocničnom zázname — odoberie anamnézu, objedná a vyhodnotí vyšetrenia, stanoví diagnózu a zapíše pokyny. V 574 retrospektívnych prípadoch z verejnej databázy, rozdelených medzi osem vopred vybraných diagnóz, podľa autorov prekonala lekárov v diagnostickej presnosti a prijímala prevažne odporúčané rozhodnutia bezpečné z hľadiska liečby. Každé spresnenie v tejto vete má váhu: pracovala v testovacom prostredí s minulými záznamami a iba s textom; veľká časť jej náskoku pochádzala zo stavov s jednoznačnými výsledkami vyšetrení; objednávala približne dvakrát toľko krvných testov ako lekári; a viaceré výsledky sa hodnotili podľa pôvodného záznamu. Skutočným pokrokom je agent, ktorý koná v celom pracovnom postupe — nie dôkaz, že stroj už diagnostikuje lepšie než lekár. Autori to hovoria jasne: zobecniteľnosť, bezpečnosť a správa si stále vyžadujú prospektívne štúdie v reálnom prostredí.

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

Odpovedať na otázku nie je to isté ako viesť prípad

Väčšina príbehov o medicínskej AI je o modeli, ktorý *odpovedá*. Dáte mu krátky opis prípadu alebo skúšobnú otázku, on vám vráti diagnózu alebo odsek rady a dosiahne dobré skóre. Je to užitočné, ale obmedzené: šikovný konzultant, ktorý so samotnou dokumentáciou nepracuje.

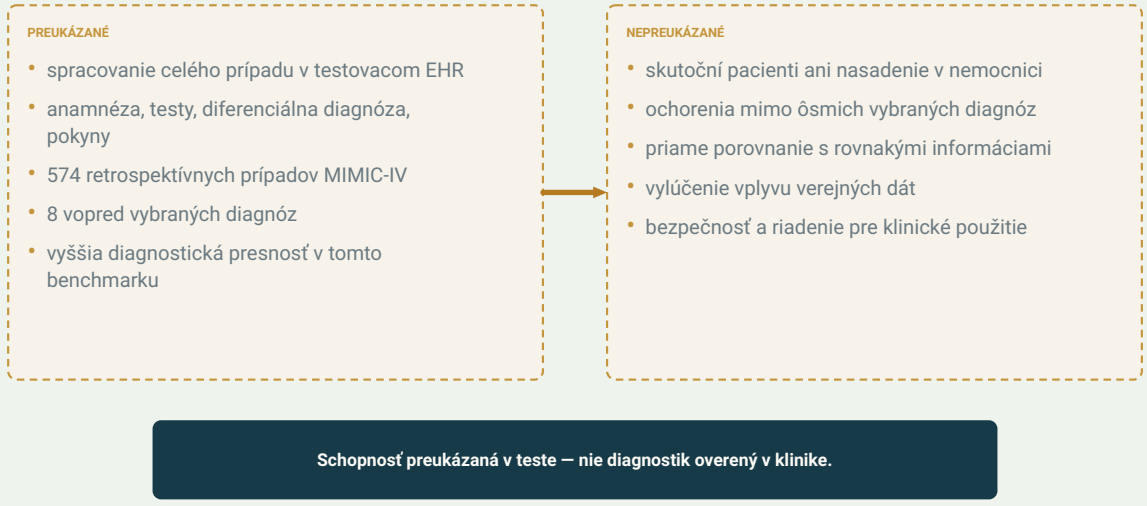
MIRA je navrhnutá na niečo iné a rozdiel spočíva v skutočnom konaní. Namiesto odpovede na jedinú otázku rieši celý prípad: číta záznam pacienta, rozhoduje, ktoré anamnestické údaje ešte potrebuje, objednáva laboratórne, zobrazovacie a mikrobiologické vyšetrenia, číta výsledky, zužuje diferenciálnu diagnostiku a napokon zapisuje potrebné pokyny — predpisy, prijatie do nemocnice či odporúčanie na operáciu. Jednotlivé kroky vykonáva *priamo* v elektronickej zdravotnej dokumentácii ako lekár, namiesto toho, aby iba vytvorila voľný text, ktorý potom prepíše človek.

To je skutočný krok: od systému, ktorý radí, k systému, ktorý prechádza celým pracovným postupom. Je to zároveň presne ten typ pokroku, ktorý médiá zjednodušia na „AI prekonáva lekárov“. Preto stojí za to presne povedať, čo sa testovalo a v akom prostredí.

Jednovetná verzia: MIRA spracovala celé prípady samostatne a podľa tohto benchmarku prekonala lekárov v diagnostickej presnosti — ale urobila to v testovacom prostredí, na retrospektívnych záznamoch, pri ôsmich vopred vybraných diagnózach, komunikovala iba textom a veľkú časť svojej výhody získala pri diagnózach s najjednoznačnejšími výsledkami vyšetrení. Pokrok je skutočný. Výsledok benchmarku však nie je klinická prax.

MIRA preukázala zvládnutie postupu, nie klinickú správnosť

Testovacie prostredie umožní agentovi prejsť celým retrospektívnym prípadom. Stále nejde o štúdiu na skutočných pacientoch.



Obrázok oddeľuje výsledok pracovného postupu od tvrdenia o nasadení.

MIRA predviedla schopnosť prejsť celým pracovným postupom v testovacom EHR. Nepreukázala skutočné klinické nasadenie, široké diagnostické pokrytie ani spravodlivý a rovnako informovaný priamy súboj s lekármi.

Čo autori urobili

MIRA — Medical Intelligence for Reasoning and Action — je autonómny agent založený na modeloch OpenAI: časť, ktorá komunikuje a koná, používa **GPT-4o** a samostatný plánovací krok využíva model uvažovania **o1**. Modely pracujú vo vlastnom rámci založenom na HL7 FHIR, dátovom štandarde, ktorý nemocnice používajú na výmenu zdravotných záznamov, a majú k dispozícii viac než osemdesiattisíc možných klinických úkonov. V tomto testovacom prostredí môže MIRA získavať anamnézu, objednávať a interpretovať laboratórne, zobrazovacie a mikrobiologické vyšetrenia, zostavovať diferenciálnu diagnostiku a formulovať liečebné plány — predpisovať lieky, plánovať chirurgické zákroky a prijatie do nemocnice. Dôležitý detail treba uviesť hneď: automatizovaní „hodnotitelia“ odpovedí MIRA sami používali GPT-4o, teda rovnakú rodinu modelov, aká sa testovala. Keď na tento problém upozornil recenzent, autori hodnotenie doplnili posúdením atestovaného lekára.

Testovacia sada pochádza z **MIMIC-IV**, veľkej verejne dostupnej databázy deidentifikovaných elektronických zdravotných záznamov. Z približne 300 000 pacientov liečených v Beth Israel Deaconess Medical Center v rokoch 2008 až 2019 autori vybrali **574 prípadov pacientov** pokrývajúcich **osem cieľových diagnóz** — brušné a internistické urgentné ochorenia, medzi nimi zápal slepého čreva, pankreatitída, zápal pľúc a infekcie močových ciest. Pre každý prípad MIRA vychádzala z obmedzeného obrazu a musela krok za krokom rozhodnúť, čo urobiť ďalej — podobne ako pri skutočnom vyšetrení, no nad záznamom, ktorého skutočný výsledok je už známy.

Jej výkon bol následne porovnaný s výkonom lekárov pri rovnakých prípadoch.

Čo vlastne znamená “autonómny agent v testovacom EHR”

Tri slová tu robia veľa práce, takže stojí za to ich rozobrať.
Agent znamená, že systém nie je chatbot odpovedajúci na jednu výzvu. Pracuje v cykle: posúdi aktuálny stav, zvolí úkon (objednať vyšetrenie, vyžiadať anamnestický údaj), uvidí výsledok a zvolí ďalší úkon, až kým nedospeje k diagnóze a plánu. „Inteligenciu“ poskytuje jazykový model; *agentný rámec* je podporná vrstva, ktorá premieňa jeho text na povolené operácie v EHR a výsledky mu vracia.
Testovacie EHR znamená simulovanú, uzavretú kópiu systému lekárskych záznamov, nie živý nemocničný systém. MIRA môže “objednať” test a prijať výsledok, ktorý pacient skutočne dostal, pretože prípad je historický a odpoveď je už zaznamenaná. Nič, čo robí, sa nedotkne skutočného pacienta ani skutočnej kliniky.
Autonómny znamená, že dokončí prípad od začiatku do konca bez človeka v slučke — *v testovacom prostredí*. To **neznamená**, že bol vypustený na starostlivosť o pacientov bez dozoru. To sú veľmi odlišné tvrdenia a testované bolo len to prvé.
Ešte jedna vec o testovacom prostredí, pretože je ľahké si ho nesprávne predstaviť: MIRA môže *objednať* akýkoľvek test, ktorý chce, ale systém môže vrátiť výsledok len vtedy, ak bol tento test **skutočne vykonaný** pre tohto pacienta v reálnom živote. Požiadajte o krvný test,

> ktorý pacient absolvoval, a dostanete skutočné hodnoty; požiadajte o vyšetrenie, ktoré nikto neobjednal, a dostanete odpoveď "N/A — nebolo možné vykonať," nikdy nevymyslené číslo. Anamnéza funguje rovnako: ak záznam neobsahuje to, na čo sa MIRA pýta, simulovaný pacient jednoducho povie, že nevie. MIRA teda nemôže vyvolať jediný rozhodujúci test, ktorý nikdy nebol absolvovaný — môže fungovať len s tým, čo zahŕňala skutočná starostlivosť.
> Tento rozdiel je dôležitý, pretože práve kľúčové slovo — „autonómny" — sa najľahšie chápe v tom druhom zmysle.

Čo našli

V benchmarku **MIRA prekonala lekárov v diagnostickej presnosti** — titulok však skrýva tri odlišné čísla, ktoré sa ľahko pomiešajú, preto ich treba oddeliť.

Vo všetkých **574 prípadoch** MIRA samostatne uviedla správnu diagnózu v **88,9 %** prípadov podľa prepúšťacej diagnózy skutočne zaznamenanej v MIMIC-IV. V priamom porovnaní lekári neriešili všetkých 574 prípadov, ale spoločnú podmnožinu **311 prípadov** s rovnakými záznamami a nástrojmi ako MIRA. V nich MIRA dosiahla **87,8 %** a porovnávala sa s dvoma osobitne zostavenými skupinami. Prvou bola **skupina skúsených lekárov**: štyria atestovaní lekári so 7 až 11 rokmi praxe dosiahli **78,1 %**. Druhá, **skupina s rôznou úrovňou skúseností**, pozostávala prevažne z mladších rezidentov a dvoch špecialistov, takže sa viac podobala obsadeniu skutočného urgentného príjmu; dosiahla **71,1 %**. Pri rovnakých prípadoch a nástrojoch tak skončila prvá AI, po nej skúsení lekári a napokon tím s prevahou mladších lekárov. Práca opatrne uvádza, že výkon MIRA bol pri všetkých ôsmich ochoreniach „konzistentne porovnateľný a často lepší" než výkon lekárov.

Obstáli aj jej následné rozhodnutia: väčšinou boli v súlade s usmerneniami, bezpečné z hľadiska medikácie a vhodné pri prijatí. Autori neuvádzajú žiadne závažné liekové chyby v piatich bezpečnostných kategóriách (interakcie liekov, dávkovanie pri poruche obličiek, alergie, riziko QT a predpisovanie opioidov) a predpisy sú správne v 467 z 468 prípadov — pričom poznamenávajú, že systém "nedosiahol 100% spoľahlivosť." Ak to vezmeme doslova, je to výrazný výsledok: agent, ktorý vedie celý prípad a v diagnostike predstihuje lekárov.

Povaha víťazstva je však rovnako dôležitá ako víťazstvo samo. Nezávislí odborníci upozornili, odkiaľ výhoda pramenila. V stanovisku odborníkov Science Media Centre Dr. Wei Xing (University of Sheffield) poznamenal, že hlavný výsledok — vyššia diagnostická presnosť AI — **ovplyvňovali najmä ochorenia s jednoznačnými výsledkami vyšetrení**, napríklad apendicitída a pankreatitída, pri ktorých rozhodujúce vyšetrenie alebo laboratórna hodnota otázku uzavrú. Pri pneumónii a infekciách močových ciest, dvoch z najčastejších dôvodov skutočných návštev urgentného príjmu, dosiahli *AI aj lekári* najhoršie výsledky a rozdiel medzi nimi bol najmenší.

Medzi stranami existovala aj informačná asymetria, ktorá sa prejavila priamo v číslach. MIRA sa viac opierala o laboratórne vyšetrenia: použila približne **51 %** laboratórnych analytov dostupných v rutinnej starostlivosti oproti približne **28 %** pri atestovaných lekároch — v mediáne o sedem krvných parametrov na prípad viac. Autori zdôrazňujú, že stále išlo o hodnotu *pod* úrovňou

rutinnej starostlivosti v súbore dát, nie o stratégiu „objednať všetko“, a neuvádzajú systematický nárast drahších prierezových zobrazovacích vyšetrení. Viac informácií však môže samo zvyšovať diagnostickú presnosť, takže nešlo o porovnanie rovnako informovaných účastníkov, ako Xing priamo upozornil. Lekár, ktorý by mohol objednať všetky lacné krvné testy bez zvažovania nákladov, nepohodlia alebo oneskorenia, by na papieri takisto vyzeral lepšie.

Prečo „prekonala lekárov“ neznamená „je lepším lekárom“

Výsledok benchmarku možno ľahko preceniť. Medzi tvrdeniami „MIRA dosiahla vyššie skóre“ a „MIRA je lepšia lekárka“ je rozdiel.

Po prvé, **proti čomu bola hodnotená.** Profesorka Julie Jacko (University of Edinburgh) poznamenala, že niekoľko kľúčových výsledkov MIRA je definovaných vzhľadom na to, čo bolo zdokumentované v základnej databáze — čo znamená, že systém je odmeňovaný za **reprodukciu zaznamenaného klinického správania**, nie nevyhnutne za preukázanie optimálnej starostlivosti. Historická dokumentácia sa stáva kľúčom správnych odpovedí. To je rozumný spôsob, ako si vybudovať referenčný bod, ale meria súhlas s tým, čo bolo urobené, nie správnosť v nejakom absolútnom zmysle. Aby sme boli spravodliví voči práci, najviac to zasahuje do *zarovnania liečby* — zodpovedali pokyny MIRA záznamu? — zatiaľ čo presnosť hlavnej diagnózy sa hodnotí v porovnaní s diagnózou prepustenia, ktorá je bližšie skutočnému výsledku než ozvena dokumentácie.

Po druhé, **asymetria informácií** už spomenutá: oveľa viac krvných testov (asi 51 % dostupných analytov oproti 28 %) je viac dôkazov. Časť rozdielu v presnosti je pravdepodobne *vykúpená* väčším množstvom dát, nie lepším uvažovaním.

Po tretie, **kde sa hodnotenie uskutočnilo.** Išlo o textové testovacie prostredie s retrospektívnymi prípadmi a ôsmimi vopred zvolenými diagnózami. Skutočné klinické hodnotenie, ako uviedol Dr. Dominic Oliver (University of Oxford), závisí nielen od toho, čo pacienti hovoria, ale aj od toho, ako to hovoria — a od fyzikálneho vyšetrenia, pozorovaného správania a reči tela. Nič z toho textový agent čítajúci hotový záznam nevidí. Pacient, ktorého problém nepatrí k ôsmim vybraným diagnózam, je navyše mimo rozsahu tvrdení tejto štúdie.

Recenzenti upozornili aj na menej nápadný problém: **kontamináciu dát.** MIMIC-IV je verejná a často opisovaná databáza, takže jazykový model trénovaný na otvorenom internete mohol vidieť články, diskusie prípadov alebo samotné údaje. Dr. Midhun Parakkal Unni (University of Sheffield) na to priamo upozornil: ak sa niektoré odpovede nachádzali v tréningových dátach, časť výkonu môže byť výsledkom zapamätania, nie klinického uvažovania, a rozlíšiť ich dokáže iba nezávislá replikácia. Autori problém neignorujú: píšu, že výsledky možno „opatrne interpretovať ako možnú hornú hranicu“ a že „môžu nadhodnocovať zovšeobecnenie na iné verejné prípady“. Sami tak stanovujú strop vlastnému titulku.

Nič z toho nerobí MIRA menej zaujímavou. Iba to správne nastavuje mierku hodnotenia: v retrospektívnom benchmarku agent, ktorý mohol pracovať s celým záznamom, prekonal lekárov najmä pri diagnózach s jednoznačnými výsledkami vyšetrení — čiastočne tým, že objednal viac testov, a čiastočne tým, že sa jeho závery zhodovali so zaznamenanou dokumentáciou. To je skutočná demonštrácia

schopností, nie verdikt, že stroje dnes diagnostikujú lepšie ako lekári.

Čo toto *ne* dokazuje

- Neukazuje, že MIRA funguje na skutočných pacientoch. Každý prípad bol retrospektívny a simulovaný; žiadny pacient nebol nikdy liečený týmto systémom.
- Neukazuje, že funguje mimo ôsmich diagnóz. Podmienky mimo predvybraného súboru — chaotickej, nediferencovanej väčšiny medicíny — neboli testované.
- Neukazuje, že je bezpečné nasadiť. Samotní autori píšu, že zobecniteľnosť, bezpečnosť a správa stále potrebujú prospektívne štúdie v reálnom prostredí.
- To **ne** vytvára férový priamy súboj s lekármi. Vychádzal z oveľa väčšieho počtu krvných testov (približne 51 % dostupných analytov oproti 28 %) a niekoľko výsledkov liečby bolo hodnotených čiastočne podľa zaznamenanej dokumentácie, nie podľa základnej najlepšej starostlivosti.
- Neukazuje, že výsledok je bez memorovania. Verejné údaje MIMIC-IV sa môžu prekrývať s trénovaním modelu, čo by nezávislá replikácia musela vylúčiť.
- To **neznamená** "AI nahrádza lekárov." Je to podpora rozhodovania, ktorá môže *konať* naprieč záznamom; konsenzus recenzentov je, že skutočné použitie bude v partnerstve s klinikmi, ktorí si zachovávajú autoritu a poskytujú všetko, čo textový záznam vynecháva.

Aké silné sú dôkazy?

Rozdeľte nárok na dve časti, pretože dôkazy sú pre každú polovicu veľmi odlišné.

Ako demonštráciu schopností — že agent jazykového modelu dokáže viesť celý klinický prípad v rámci EHR sandboxu, prepájať anamnézu, testy, diagnostiku a príkazy od jedného do druhého — je táto práca skutočne nová a pomerne presvedčivá. Toto je nová časť a stojí za to venovať pozornosť.

Ako tvrdenie o nadradenosti — že MIRA je *lepšia ako lekári* — dôkazy sú obmedzené a mali by sa čítať opatrne. Platí na konkrétnom retrospektívnom benchmarku, na ôsmich diagnózach, s informačnou asymetriou (viac testov) a kľúčom odpovedí odvodeným z historickej dokumentácie, a s reálnou možnosťou kontaminácie trénovacích dát. To stačí na to, aby sme povedali, že "agent podal pôsobivý výkon na tomto benchmarku." Nestačí povedať "agent je lepší diagnostik než lekár" a autori netvrdia tú druhú vec.

Najužitočnejší postoj nie je ani "AI poráža lekárov", ani "len hračka". Je to tak: nový druh medicínskej AI — taký, ktorý *pôsobí* naprieč záznamom namiesto odpovedania na otázky — uspel v náročnom retrospektívnom benchmarku a teraz sa musí preukázať tam, kde ešte nebol vyskúšaný: na skutočných, nediferencovaných pacientoch, prospektívne a pod riadnym dohľadom.

Prečo je to dôležité

V medicínskej AI sa zaujímavá otázka už niekoľko rokov nenápadne mení. Kedysi znela „dokáže model správne určiť diagnózu?“ a modely na ňu odpovedali kladne v čoraz čistejších testovacích súboroch. MIRA predstavuje posun k ťažšej otázke: „dokáže model *vykonať úlohu* — zhromažďovať, usporadúvať, interpretovať, rozhodovať a konať — v celom pracovnom postupe?“ Je to užitočnejšia a poctivejšia otázka, pretože práve v konaní ležia ťažkosti aj riziká.

Preto záleží na rámovaní. Reflex v titulkoch — *AI prekonáva lekárov* — poukazuje na najmenej novú a najmenej podporovanú časť výsledku. Skutočne nová vec je menšia a dôležitejšia: agent, ktorý môže prechádzať celým záznamom. Táto schopnosť, ak obstojí, mení **pracovný postup** oveľa skôr, než zmení, kto je za neho zodpovedný. Lekár nie je nahradený; takýto nástroj sa najprv presadí pri objednávaní vyšetrení, interpretácii a dohľadávaní výsledkov.

A to resetuje dôkazné bremeno správnym smerom. Víťazstvo v rebríčku v retrospektívnych prípadoch je dôvodom na uskutočnenie prospektívnej štúdie, nie jeho náhradou. Autori to hovoria. Úprimné čítanie MIRA je pozvánkou k ďalšej štúdii — ktorá sa drží štandardu, ktorý odbor už pozná: merať na pacientoch, nie na benchmarkoch.

Stručné zhrnutie

MIRA je autonómny agent AI pracujúci v testovacej elektronickej zdravotnej dokumentácii: dokáže získavať anamnézu, objednávať a interpretovať laboratórne, zobrazovacie a mikrobiologické vyšetrenia, určovať diagnózu a vytvárať liečebné plány. Pri 574 retrospektívnych prípadoch z verejnej databázy MIMIC-IV a ôsmich vopred zvolených diagnózach prekonala lekárov v diagnostickej presnosti (88,9 % celkovo; 87,8 % oproti 78,1 % v priamom porovnaní) a väčšinou prijímala rozhodnutia v súlade s usmerneniami a bezpečné z hľadiska liekov. Hodnotenie však iba textovo simulovalo historické záznamy; veľká časť výhody pochádzala z ochorení s jednoznačnými výsledkami vyšetrení; MIRA používala oveľa viac krvných testov než lekári (približne 51 % dostupných analytov oproti 28 %); niektoré výsledky liečby sa hodnotili podľa pôvodnej dokumentácie; a verejný súbor dát prináša skutočné riziko kontaminácie tréningových dát, ktoré autori sami označujú za možný dôvod chápať čísla ako hornú hranicu. Skutočným pokrokom je agent konajúci v celom pracovnom postupe namiesto odpovedania na izolované otázky. Autori jasne uvádzajú, že zovšeobecniteľnosť, bezpečnosť a riadenie stále vyžadujú prospektívne štúdie v reálnom prostredí. Správne čítanie preto znie: pôsobivá ukážka schopností, nie dôkaz, že AI diagnostikuje lepšie než lekári, ani systém pripravený na skutočnú kliniku.

Kontrola bez príkras

Čo ukazuje štúdia: Agent jazykového modelu (MIRA) môže prejsť celý klinický prípad od začiatku do konca v sandboxovom EHR — anamnéza, testy, diagnóza, príkazy — a na retrospektívnom bench-

marku 574 prípadov naprieč ôsmimi diagnózami prekonal lekárov v diagnostickej presnosti (88,9 % celkovo; 87,8 % oproti 78,1 % oproti lekárom s certifikáciou) bez závažných chýb pri podávaní liekov.

Čo je pravdepodobné, ale nie dokázané: Že táto schopnosť prináša úžitok pre skutočných, nediferencovaných pacientov; že výhoda v presnosti odráža lepšie uvažovanie, nie viac objednaných testov, zhodu so zaznamenaným záznamom alebo zapamätané verejné údaje.

Čo neukazuje: Že MIRA funguje mimo svojich ôsmich diagnóz; že je bezpečné nasadiť; že prekonáva lekárov v spravodlivom a rovnako informovanom porovnaní; že nahrádza klinikov; že výsledky sú bez kontaminácie trénovacích dát.

Hlavné obmedzenia: Retrospektívne, simulované, iba textové hodnotenie; osem predvybraných diagnóz; informačná asymetria (oveľa viac krvných testov — približne 51 % dostupných analytov oproti 28 % pri krvných testoch, nie pri zobrazovaní); výsledky liečby boli čiastočne hodnotené podľa historického záznamu; a verejný benchmark (MIMIC-IV), ktorý mohol jazykový model vidieť pri tréningu — autori preto výsledok označujú ako možnú hornú hranicu výkonu.

Koľko dôvery by mal mať bežný čitateľ? Vysokú v to, že MIRA je skutočná a nová demonštrácia schopností — agent, ktorý *koná* naprieč záznamom, nie len chatbot. Nižšiu v tvrdenie, že je *lepší diagnostik než lekár*: toto tvrdenie je viazané na jeden retrospektívny benchmark a je zaťažené rozdielnym objednávaním testov, hodnotením podľa záznamu a možnou kontamináciou. Samotný záver autorov je správny — toto si vyžaduje prospektívnu štúdiu v reálnom prostredí, aby to malo význam pre starostlivosť o pacientov.

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