The Clean Paper

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

Egy MI-ügynök önállóan vitt végig teljes betegeseteket — szimulátorban, múltbeli kórlapokon

A MIRA az orvosi MI új típusa: ahelyett, hogy egyetlen kérdésre válaszolna, egy teljes esetet dolgoz fel egy szimulált kórházi nyilvántartásban — anamnézist vesz fel, vizsgálatokat rendel és értelmez, diagnózist állít fel, majd utasításokat ír. Egy nyilvános adatbázis 574 retrospektív esetén, nyolc előre kiválasztott diagnózisban a szerzők szerint diagnosztikai pontosságban felülmúlta az orvosokat, döntései pedig nagyrészt irányelv-konformak és gyógyszerbiztonsági szempontból megfelelőek voltak. De minden minősítés számít: sandboxban, múltbeli kórlapokon, kizárólag szöveggel futott; előnyének nagy része egyértelmű vizsgálati eredményekkel járó állapotokon jelent meg; körülbelül kétszer annyi vérvizsgálati információt használt, mint az orvosok; és több kimenetelt ahhoz pontoztak, amit az eredeti kórlap rögzített. A valódi előrelépés egy olyan ügynök, amely az egész munkafolyamatban cselekszik — nem annak bizonyítéka, hogy egy gép most már jobban diagnosztizál az orvosnál. A szerzők is ezt mondják: az általánosíthatóságot, biztonságot és irányítást prospektív, valós vizsgálatokban kell még tesztelni.

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

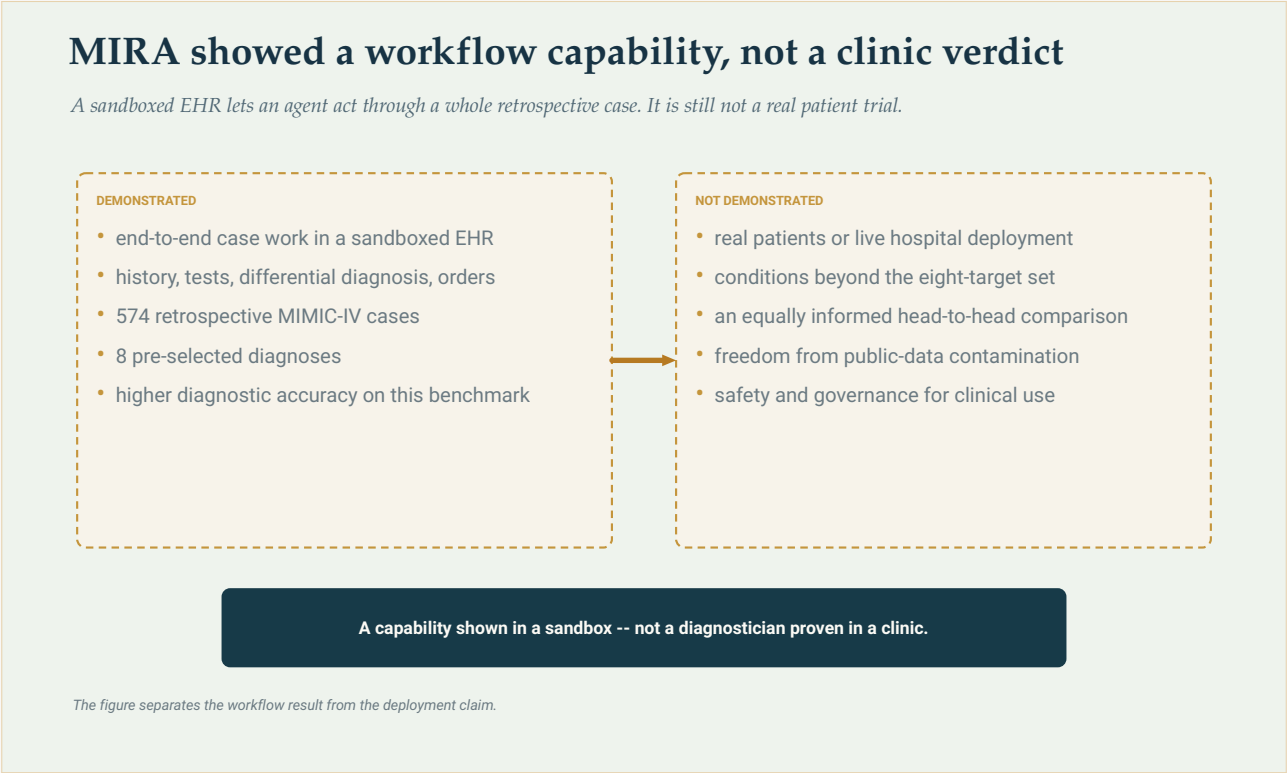
Egy kérdés megválaszolása nem ugyanaz, mint egy teljes eset végigvitele

Az orvosi MI-ről szóló történetek többsége egy olyan modellről szól, amely *válaszol*. Kap egy esetleírást vagy vizsgakérdést, visszaad egy diagnózist vagy néhány bekezdésnyi tanácsot, és jó pontszámot ér el. Hasznos, de szűk szerep: egy okos tanácsadó, aki soha nem nyúl a betegkartonhoz.

A MIRA-t valami másra építették, és éppen ez a különbség az igazi hír. Ahelyett, hogy egyetlen kérdésre válaszolna, egy teljes esetet visz végig: elolvassa a beteg kórlapját, eldönti, milyen anamnézisre van még szüksége, laboratóriumi, képalkotó és mikrobiológiai vizsgálatokat rendel, értelmezi az eredményeket, szűkíti a differenciáldiagnózist, majd megírja a következő utasításokat — gyógyszerrendelést, felvételt, sebészeti beutalást. Mindezt az elektronikus egészségügyi nyilvántartáson *belül* végrehajtott műveletekkel teszi, ahogy egy klinikus, nem pedig úgy, hogy szabad szöveget ír, amelyet valakinek kézzel kell átvezetnie.

Ez valódi lépés: a tanácsot adó rendszertől a teljes munkafolyamatban cselekvő rendszer felé. És pontosan az a fajta lépés, amelyet a sajtó könnyen „az MI felülmúlja az orvosokat” mondattá lapít. Ezért érdemes pontosan megnézni, mit teszteltek — és hol.

Egy mondatban: a MIRA önállóan vitt végig teljes eseteket, és ezen a benchmarkon diagnosztikai pontosságban felülmúlta az orvosokat — de mindezt sandboxban, retrospektív kórlapokon, nyolc előre kiválasztott diagnózis körében, kizárólag szöveges kommunikációval tette, és előnyének jelentős része azokon a betegségeken jelent meg, amelyeknél a vizsgálati eredmények egyértelműek voltak. Az előrelépés valós. Az eredménytábla viszont nem a klinika.



A MIRA egy sandboxolt EHR-ben demonstrálta a teljes, elejétől végéig tartó munkafolyamat kezelését. Nem demonstrálta a valós klinikai bevezetést, a széles diagnosztikai lefedettséget vagy az orvosokkal való, azonos információkon alapuló igazságos közvetlen összehasonlítást.

Original Aurora diagram — The Clean Paper CC BY 4.0

Mit csináltak a szerzők

A MIRA — Medical Intelligence for Reasoning and Action — az OpenAI modelljeire épülő autonóm ügynök: a beszélgetést és a műveleteket végző rész **GPT-4o**-t használ, míg egy külön tervezési lépés az OpenAI **o1** érvelési modelljét. Ezek egy egyedi, szabványkompatibilis keretrendszerben működnek — HL7 FHIR-re építve, arra az adatstandardra, amelyet valódi kórházak használnak az egészségügyi adatok mozgatására —, több mint nyolcvanezer lehetséges klinikai művelettel. Ebben a sandboxban a MIRA lekérheti a beteg anamnézisét, labor-, képalkotó és mikrobiológiai vizsgálatokat rendelhet és értelmezhet, differenciáldiagnózist állíthat fel, valamint kezelési tervet fogalmazhat meg — gyógyszereket írhat elő, műtéti beavatkozást ütemezhet, kórházi felvételt tervezhet. Egy részletet érdemes rögtön kiemelni: a MIRA válaszait pontozó automatizált „bírák" maguk is GPT-4o-modellek voltak — ugyanabba a modellcsaládba tartoznak, mint amelyet teszteltek —; miután egy szakmai lektor ezt kifogásolta, a szerzők egy szakvizsgázott orvos ellenőrzésével egészítették ki az értékelést.

A tesztkészlet a **MIMIC-IV** adatbázisból származott, amely egy nagy, nyilvánosan hozzáférhető, anonimizált EHR-adatbázis. A Beth Israel Deaconess Medical Centerben 2008 és 2019 között kezelt mintegy 300 000 beteg közül a szerzők **574 esetet** választottak ki **nyolc cél-diagnózisban** — hasi kórképek és belgyógyászati sürgősségi állapotok, köztük appendicitis, pancreatitis, pneumonia és húgyúti fertőzés. Minden esetben a MIRA korlátozott kezdeti képből indult, majd lépésről lépésre kellett eldöntenie, mi legyen a következő teendő — ugyanabban a formában, mint egy valódi kivizsgálás, csak egy olyan kórlapon futtatva, amelynek valódi végkimenetele már ismert.

Teljesítményét ezután ugyanazokon az eseteken orvosokéval hasonlították össze.

Mit jelent valójában az, hogy „autonóm ügynök egy sandboxolt EHR-ben"

Három szó végez itt rengeteg munkát, ezért érdemes kibontani őket.
Az **ügynök** azt jelenti, hogy a rendszer nem egyszerű chatbot, amely egy promptot megválaszol. Ciklusban működik: megnézi az aktuális állapotot, választ egy műveletet (rendelje meg ezt a vizsgálatot, kérdezzen rá arra az anamnesztikus adatra), megkapja az eredményt, majd újabb műveletet választ — amíg el nem jut egy diagnózisig és tervig. Az „intelligencia" a nyelvi modell; az *ágencia* a köré épített keretrendszer, amely a modell szövegét engedélyezett EHR-műveletekké alakítja, majd visszaadja neki az eredményeket.
A **sandboxolt EHR** egy orvosi nyilvántartó rendszer szimulált, elzárt másolatát jelenti, nem élő kórházi rendszert. A MIRA „megrendelhet" egy vizsgálatot, és megkaphatja azt az eredményt, amelyet a beteg valóban kapott, mert az eset történeti, és az eredmény már szerepel a kórlapban. Semmi, amit tesz, nem érint valódi beteget vagy valódi klinikát.
Az **autonóm** azt jelenti, hogy az esetet emberi közbeavatkozás nélkül végigviszi — *a sandboxon*

> *belül.* Ez **nem** azt jelenti, hogy felügyelet nélkül betegeket bíztak rá. A kettő nagyon különböző állítás, és csak az elsőt tesztelték.
> Még egy fontos részlet a sandboxról, mert könnyű félreérteni: a MIRA *megrendelhet* bármilyen vizsgálatot, de a rendszer csak akkor tud eredményt visszaadni, ha azt a vizsgálatot **valóban elvégezték** ennél a betegnél a valós életben. Ha olyan vérvizsgálatot kér, amelyet a beteg kapott, megkapja a valódi értékeket; ha olyan képalkotást kér, amelyet senki nem rendelt meg, „N/A — nem volt elvégezhető" választ kap, soha nem kitalált számot. Az anamnézis ugyanígy működik: ha a kórlap nem tartalmazza azt, amire a MIRA rákérdez, a szimulált beteg egyszerűen azt mondja, hogy nem tudja. A MIRA tehát nem varázsolhat elő egy olyan döntő vizsgálatot, amelyet soha nem végeztek el — csak abból dolgozhat, amit a valódi kivizsgálás történetesen tartalmazott.
> A megkülönböztetés azért fontos, mert a főcímben szereplő szó — „autonóm" — a legkönnyebben olvasható úgy, mintha a második dolgot jelentené.

Mit találtak

A benchmarkon **a MIRA diagnosztikai pontosságban felülmúlta az orvosokat** — de a főcím három külön számot rejt, amelyeket könnyű összemosni, ezért érdemes szétválasztani őket.

Önmagában, mind az **574 esetben**, a MIRA az esetek **88,9%**-ában nevezte meg a helyes diagnózist — azzal az elbocsátási diagnózissal összevetve, amelyet a MIMIC-IV-ben ténylegesen rögzítettek. Az emberekkel való közvetlen összehasonlításban az orvosok nem mind az 574 esetet dolgozták fel; egy közös **311 esetből álló részhalmazt** kaptak, ugyanazokkal a kórlapokkal és eszközökkel, mint a MIRA. Ezen a 311 eseten a MIRA **87,8%**-ot ért el, és két külön toborzott orvoscsoporttal hasonlították össze. Az első egy **szenior csoport** volt: négy szakvizsgázott orvos 7–11 év tapasztalattal, akik **78,1%**-ot értek el. A második egy **vegyes szenioritású csoport** volt: főként fiatal rezidensek két szakorvossal, ami közelebb áll egy valódi sürgősségi osztály személyzeti összetételéhez; ők **71,1%**-ot értek el. Ugyanazok az esetek, ugyanazok az eszközök — és a sorrend MI, tapasztalt orvosok, majd a főként junior csapat lett. A tanulmány saját óvatos megfogalmazása szerint a MIRA mind a nyolc betegségben „következetesen egyenértékű volt az orvosokkal, és gyakran felül is múlta őket".

A további döntései is jól teljesítettek: nagyrészt összhangban álltak az irányelvekkel, gyógyszerbiztonsági szempontból megfelelőek voltak, és a felvételi döntések is helyesnek bizonyultak. A szerzők öt biztonsági kategóriában — gyógyszerkölcsönhatások, vesefunkcióhoz igazított adagolás, allergiák, QT-kockázat és opioidfelírás — nem találtak nagy súlyosságú gyógyszerelési hibát, és 468 esetből 467-ben helyesnek ítélték a recepteket, miközben megjegyzik, hogy a rendszer „nem érte el a 100%-os megbízhatóságot". Első olvasásra ez feltűnő eredmény: egy ügynök végigviszi az egész esetet, és a diagnózisban az orvosok előtt végez.

De a győzelem *szerkezete* legalább olyan fontos, mint maga a győzelem. A tanulmányt értékelő független szakértők rámutattak, honnan származott valójában az előny. A Science Media Centre szakértői paneljén Dr. Wei Xing (University of Sheffield) megjegyezte, hogy a főcímben szereplő eredményt — az MI nagyobb diagnosztikai pontosságát — **főként azok az állapotok hajtották,**

amelyeknél egyértelmű vizsgálati eredmények voltak, például appendicitis és pancreatitis, ahol egy döntő képalkotó vagy laborérték eldöntheti a kérdést. Pneumonia és húgyúti fertőzés esetén — két nagyon gyakori oknál, amiért az emberek sürgősségi osztályra kerülnek — *mind* az MI, *mind* az orvosok teljesítettek a legrosszabbul, és köztük is itt volt a legkisebb különbség.

Aszimmetria volt abban is, hogyan „játszott" a két oldal, és ez a tanulmány saját számaiban látszik. A MIRA jóval erősebben támaszkodott a laborra: a rutinszerű ellátásban hozzáférhető laborparaméterek körülbelül **51%**-át használta fel, míg a szakvizsgázott orvosok nagyjából **28%**-át — esetenként medián hét további vérparamétert. A szerzők hangsúlyozzák, hogy ez még mindig *kevesebb* volt a teljes adatbázisban látható rutinellátási alapvonalnál, tehát nem „rendeljünk meg mindent" stratégia, és nem találtak szisztematikus növekedést a drágább keresztmetszeti képalkotásban. Ettől függetlenül több információ önmagában is növelheti a diagnosztikai pontosságot — így ez nem teljesen azonos feltételek mellett zajló verseny két egyformán tájékozott fél között, amit Xing közvetlenül is kiemelt. Egy olyan klinikus is élesebbnek tűnne papíron, aki minden olcsó vérvizsgálatot szabadon megrendelhetne anélkül, hogy költséget, kellemetlenséget vagy késlekedést mérlegelne.

Miért nem ugyanaz az, hogy „legyőzte az orvosokat", mint hogy „jobb orvos"

Egy benchmark-győzelmet könnyű túlértelmezni. Három dolog áll a „MIRA magasabb pontszámot ért el" és a „MIRA jobb" között.

Először: **mihez viszonyítva pontozták.** Julie Jacko professzor (University of Edinburgh) megjegyezte, hogy a MIRA több kulcsfontosságú kimenetelét az alapul szolgáló adatbázisban dokumentáltakhoz képest határozták meg — vagyis a rendszer jutalmat kap a **rögzített klinikai viselkedés reprodukálásáért**, nem feltétlenül az optimális ellátás demonstrálásáért. A történeti kórlap lesz a megoldókulcs. Ez ésszerű mód egy benchmark felépítésére, de azt méri, mennyire egyezik azzal, amit tettek, nem azt, mi lenne abszolút értelemben helyes. A tanulmány javára írható, hogy ez leginkább a *kezelési egyezést* mérő mutatókat érinti — egyeznek-e a MIRA utasításai a kórlappal? —, míg a fő diagnosztikai pontosságot az elbocsátási diagnózishoz viszonyítják, ami közelebb áll valódi kimenetelhez, mint a dokumentáció puszta visszhangjához.

Másodszor ott van a már említett **információs aszimmetria**: sokkal több vérvizsgálat (az elérhető analitok kb. 51%-a a 28%-kal szemben) több bizonyítékot jelent. A pontossági különbség egy része valószínűleg *megvásárolt* információból származik, nem pusztán jobb érvelésből.

Harmadszor: **hol futott.** Sandboxban, retrospektív eseteken, nyolc előre kiválasztott diagnózisban, kizárólag szöveggel. A valós klinikai értékelés, ahogy Dr. Dominic Oliver (University of Oxford) fogalmazott, nemcsak attól függ, *mit* mondanak a betegek, hanem attól is, *hogyan* mondják — ehhez jön a fizikális vizsgálat, a megfigyelt viselkedés és a testbeszéd, amelyekből egy kész kórlapot olvasó, szöveges ügynök semmit sem lát. És ha a beteg problémája nem tartozik a nyolc kiválasztott diagnózis egyikébe, erről ez a tanulmány egyszerűen nem mond semmit.

Van egy csendesebb aggály is, amelyet a lektorok felvetettek: **adatkontaminació**. A MIMIC-IV nyilvános, és nagyon sokat írtak róla, ezért egy nyílt interneten tanított nyelvi modell találkozhatott már tanulmányokkal, esetmegbeszélésekkel vagy akár magukkal az adatokkal is. Dr. Midhun Parakkal Unni (University of Sheffield) ezt közvetlenül kiemelte — ha a válaszok egy része benne volt a tanítóadatban, a teljesítmény egy része memória, nem klinikai érvelés, és csak független replikáció tudja szétválasztani a kettőt. Figyelemre méltó, hogy a szerzők nem söprik félre a kérdést: azt írják, eredményeik „óvatosan egy lehetséges felső korlátként értelmezhetők", és „túlbecsülhetik a más nyilvános esetekre való általánosíthatóságot" — vagyis maga a tanulmány tesz felső korlátot saját főcímére.

Ettől a MIRA nem lesz kevésbé érdekes. Csak a helyes olvasat lesz kalibráltabb: egy retrospektív benchmarkon egy, a teljes betegkartonon végig cselekvő ügynök felülmúlta az orvosokat azokban a diagnózisokban, amelyeknél tiszta vizsgálati eredmények voltak — részben több vizsgálat megrendelésével, részben a rögzített kórlappal való egyezéssel. Ez valódi képességdemonstráció, nem ítélet arról, hogy a gépek mostantól jobban diagnosztizálnak, mint az orvosok.

Mit *nem* bizonyít

- **Nem** mutatja, hogy a MIRA valódi betegeken működik. Minden eset retrospektív és szimulált volt; egyetlen beteget sem kezelt ténylegesen.
- **Nem** mutatja, hogy nyolc diagnózison túl is működik. Az előre kiválasztott körön kívüli állapotokat — az orvoslás kusza, differenciálatlan többségét — nem tesztelték.
- **Nem** mutatja, hogy biztonságosan bevethető. Maguk a szerzők írják, hogy az általánosíthatóságot, biztonságot és irányítást prospektív, valós környezetben végzett vizsgálatokban kell még tesztelni.
- **Nem** teremt igazságos közvetlen összehasonlítást az orvosokkal. Sokkal több vérvizsgálatot használt (az elérhető analitok kb. 51%-át a 28%-kal szemben), és több kezelési kimenetelt részben a rögzített kórlaphoz, nem pedig egy objektív „legjobb ellátáshoz" pontoztak.
- **Nem** mutatja, hogy az eredmény mentes a memorizálástól. A nyilvános MIMIC-IV adat átfedhetett a modell tanítóadataival; ezt csak független replikáció tudná kizárni.
- **Nem** jelenti azt, hogy „az MI leváltja az orvosokat". Ez olyan döntéstámogatás, amely *cselekedni* is tud a kórlapon; a szakértői értékelések közös következtetése az, hogy a valós használat együttműködés lesz klinikusokkal, akik megtartják a döntési jogkört és hozzáadják mindazt, amit egy szöveges nyilvántartás nem tartalmaz.

Mennyire erős a bizonyíték?

Érdemes kettéválasztani az állítást, mert a két felét nagyon különböző erősségű bizonyíték támasztja alá.

Képességdemonstrációként — hogy egy nyelvi modellre épülő ügynök képes egy teljes klinikai esetet végigvinni EHR-sandboxban, az anamnézist, vizsgálatokat, diagnózist és utasításokat egymás után

összekapcsolva — a munka valóban újszerű és meggyőző. Ez az új rész, és erre érdemes figyelni.

Felsőbbrendűségi állításként — hogy a MIRA *jobb az orvosoknál* — a bizonyíték korlátozott, és óvatosan kell olvasni. Egy konkrét retrospektív benchmarkon, nyolc diagnózisban áll, információs aszimmetria mellett (több vizsgálat), történeti kórlapból származó megoldókulccsal és a tanítóadat-kontamináció valós lehetőségével. Ez elég ahhoz, hogy azt mondjuk: „az ügynök lenyűgözően teljesített ezen a benchmarkon". Nem elég ahhoz, hogy azt mondjuk: „az ügynök jobb diagnoszta, mint egy orvos", és a szerzők sem állítják az utóbbit.

A leghasznosabb álláspont tehát sem az, hogy „az MI legyőzi az orvosokat", sem az, hogy „ez csak játék". Inkább ez: egy új típusú orvosi MI — amely *cselekszik* a kórlapon, ahelyett hogy kérdésekre válaszolna — jól teljesített egy nehéz retrospektív benchmarkon, és most ott kell bizonyítania, ahol még nem próbálták: valódi, előre nem kategorizált betegeken, prospektíven és megfelelő irányítás alatt.

Miért fontos

Az elmúlt néhány évben az orvosi MI érdekes kérdése csendben megváltozott. Korábban az volt: „képes-e a modell eltalálni a diagnózist?" — és a modellek egyre tisztább tesztkészleteken újra meg újra igennel válaszoltak. A MIRA egy nehezebb kérdés felé tolja a hangsúlyt: „képes-e a modell *elvégezni a munkát* — adatot gyűjteni, vizsgálatot rendelni, értelmezni, dönteni, cselekedni — egy teljes munkafolyamaton át?" Ez hasznosabb és őszintébb kérdés, mert a nehézség és a kockázat egyaránt a cselekvésnél jelenik meg.

Ezért számít annyira a keretezés. A reflexszerű főcím — *az MI felülmúlja az orvosokat* — az eredmény legkevésbé újszerű és legkevésbé alátámasztott részére mutat. Az igazán új dolog kisebb, de következményesebb: egy ügynök, amely végig tud haladni az egész betegkartonon. Ha ez a képesség megállja a helyét, előbb változtatja meg a **munkafolyamatot**, mint azt, hogy ki irányítja. Az orvos nem tűnik el; először a vizsgálatok rendelése, értelmezése és az eredmények követése — maga a munkafolyamat — az a terület, ahol egy ilyen eszköz megjelenik.

És helyes irányba állítja a bizonyítási terhet. A retrospektív eseteken elért ranglistagyőzelem ok arra, hogy prospektív vizsgálatot indítsunk, nem annak helyettesítője. A szerzők is ezt mondják. A MIRA őszinte olvasata meghívás a következő vizsgálatra — azzal a mércével, amelyet a terület már ismer: betegeken mérve, nem benchmarkokon.

Röviden

A MIRA autonóm MI-ügynök, amely egy sandboxolt elektronikus egészségügyi nyilvántartásban működik: anamnézist vehet fel, labor-, képalkotó és mikrobiológiai vizsgálatokat rendelhet és értelmezhet, diagnózist állíthat fel, és kezelési tervet írhat. A nyilvános MIMIC-IV adatbázis 574 retrospektív esetén, nyolc előre kiválasztott diagnózisban tesztelve diagnosztikai pontosságban

felülmúlta az orvosokat (összesen 88,9%; közvetlen összehasonlításban 87,8% a 78,1%-kal szemben), és döntései nagyrészt összhangban álltak az irányelvekkel és gyógyszerbiztonsági szempontból megfelelőek voltak. Az értékelés azonban múltbeli kórlapokon futó, kizárólag szöveges szimuláció volt; az előny nagy része egyértelmű vizsgálati eredményekkel járó állapotokon jelent meg; a MIRA sokkal több vérvizsgálati információt használt, mint az orvosok (az elérhető analitok kb. 51%-át a 28%-kal szemben); több kezelési kimenetelt az eredeti kórlapban rögzítettekhez viszonyítottak; és a nyilvános adatbázis valós tanítóadat-kontaminációs kockázatot jelent — amit maguk a szerzők is lehetséges felső korlátnak neveznek számaikra. A valódi előrelépés egy olyan ügynök, amely elszigetelt kérdések megválaszolása helyett a teljes munkafolyamatban cselekszik. A szerzők egyértelműek abban, hogy az általánosíthatóságot, biztonságot és irányítást prospektív, valós környezetben végzett vizsgálatokkal kell tesztelni. Ez a helyes olvasat: lenyűgöző képességdemonstráció, nem bizonyíték arra, hogy az MI jobban diagnosztizál az orvosoknál, és nem egy valós klinikára kész rendszer.

Tárgyilagos mérleg

Mit mutat a tanulmány: Egy nyelvi modellre épülő ügynök (MIRA) képes egy teljes klinikai esetet elejétől a végéig végigvinni egy sandboxolt EHR-ben — anamnézis, vizsgálatok, diagnózis, utasítások —, és egy nyolc diagnózist lefedő, 574 esetből álló retrospektív benchmarkon diagnosztikai pontosságban felülmúlta az orvosokat (összesen 88,9%; a közvetlen összehasonlításban 87,8% a szakvizsgázott orvosok 78,1%-ával szemben), nagy súlyosságú gyógyszerelési hiba nélkül.

Mi hihető, de nincs bizonyítva: Hogy ez a képesség valódi, előre nem kategorizált betegeknél is előnnyé válik; hogy a pontossági előny jobb érvelésből származik, nem pedig több megrendelt vizsgálatból, a rögzített kórlappal való egyezésből vagy nyilvános adatok memorizálásából.

Mit nem mutat: Hogy a MIRA a nyolc diagnózison kívül is működik; hogy biztonságosan bevethető; hogy egy igazságos, azonos információkon alapuló összehasonlításban is legyőzi az orvosokat; hogy kiváltja a klinikusokat; vagy hogy az eredmények mentesek a tanítóadat-kontaminációtól.

Fő korlátok: Retrospektív, szimulált, kizárólag szöveges értékelés; nyolc előre kiválasztott diagnózis; információs aszimmetria (jóval több vérvizsgálat — az elérhető analitok kb. 51%-a a 28%-kal szemben, főként olcsó vérvizsgálatok, nem képalkotás); kezelési kimenetelek részben a történeti kórlaphoz pontozva; valamint egy nyilvános benchmark (MIMIC-IV), amelyet a nyelvi modell láthatott a tanítás során — amit a szerzők a teljesítmény lehetséges felső korlátjaként említenek.

Mennyire bízhat benne egy általános olvasó? Nagy bizalommal abban, hogy a MIRA valódi és újszerű képességdemonstráció — egy ügynök, amely *cselekszik* a kórlapon, nem pusztán chatbot. Alacsony bizalommal abban, hogy *jobb diagnoszta, mint egy orvos*: ez az állítás egyetlen retrospektív benchmarkra korlátozódik, és torzíthatja a vizsgálatrendelés, a kórlaphoz való pontozás és az esetleges kontamináció. A szerzők saját végkövetkeztetése a helyes: prospektív, valós vizsgálatokra van szükség, mielőtt ennek betegellátási jelentést tulajdoníthatnánk.

Források

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