The Clean Paper

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

En AI-agent håndterede hele patientforløb på egen hånd — i en simulator, baseret på tidligere journaler

MIRA er en ny slags medicinsk AI: I stedet for at besvare ét spørgsmål håndterer den et helt forløb i en simuleret hospitalsjournal — optager sygehistorie, bestiller og læser undersøgelser, når frem til en diagnose og skriver ordinationerne. På 574 retrospektive forløb fra en offentlig database, inden for otte på forhånd udvalgte diagnoser, rapporterer forfatterne, at den overgik læger i diagnostisk nøjagtighed og traf beslutninger, som i vid udstrækning fulgte retningslinjerne og var lægemiddelsikre. Hvert forbehold i den sætning vejer tungt: Den kørte i et sandkassemiljø med tidligere journaler, udelukkende i tekst; en stor del af forspringet kom ved tilstande med entydige testresultater; den bestilte cirka dobbelt så mange blodprøver som lægerne; og flere resultater blev bedømt op mod det, der var registreret i den oprindelige journal. Det virkelige fremskridt er en agent, der handler gennem hele arbejdsgangen — ikke bevis for, at en maskine nu stiller bedre diagnoser end en læge. Forfatterne siger det selv: Generaliserbarhed, sikkerhed og styring kræver stadig prospektive studier i den virkelige verden.

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

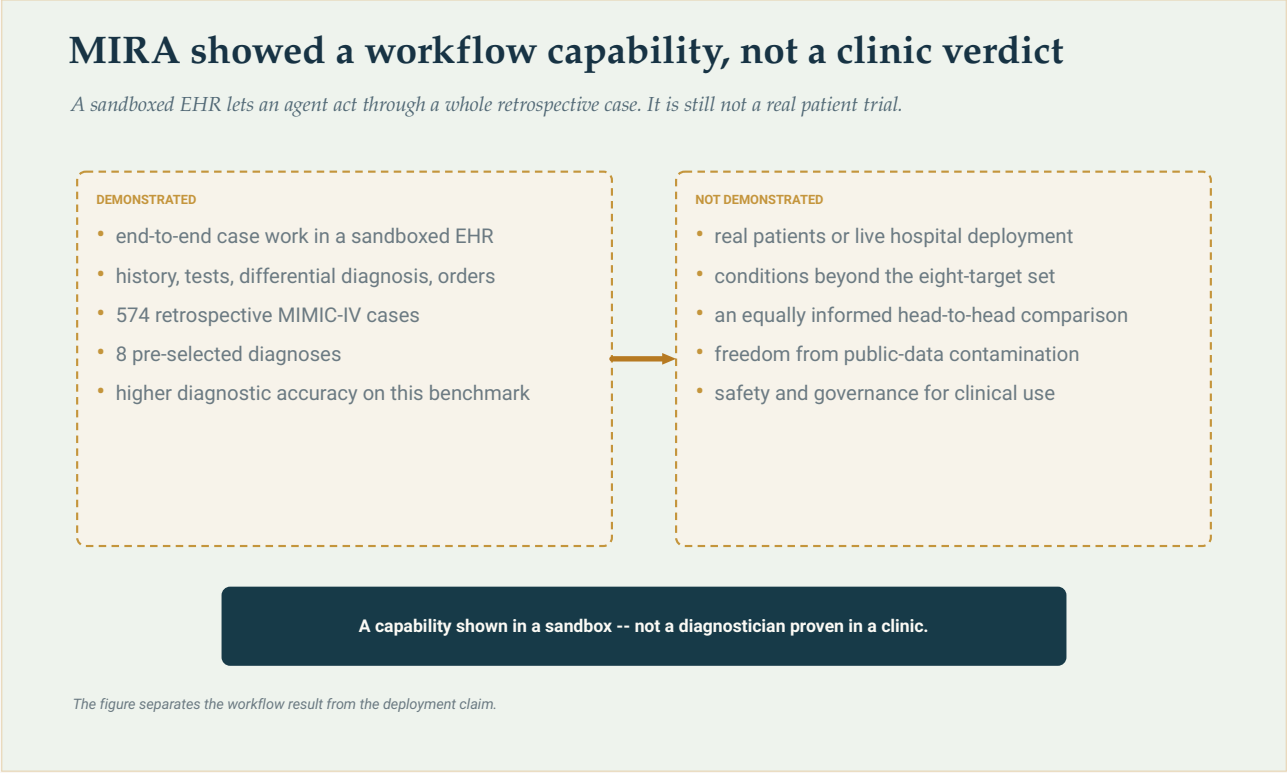
At besvare et spørgsmål er ikke det samme som at håndtere hele sagen

De fleste historier om medicinsk AI handler om en model, der *svarer*. Man giver den en sygehistorie eller et eksamensspørgsmål, den giver en diagnose eller et afsnit med råd tilbage og scorer godt. Nyttigt, men snævert: en skarp konsulent, der aldrig rører journalen.

MIRA er bygget til at gøre noget andet, og netop den forskel er den egentlige nyhed. I stedet for at besvare ét spørgsmål håndterer systemet en hel sag: Det læser patientens journal, afgør, hvilke oplysninger i sygehistorien det stadig har brug for, bestiller laboratorieprøver, billeddiagnostiske og mikrobiologiske undersøgelser, læser resultaterne, indsnævrer en differentialdiagnose og skriver derefter de ordinationer, der følger — recepter, en indlæggelse, en henvisning til operation. Det gør dette ved at udføre handlinger *inde i* en elektronisk patientjournal, sådan som en kliniker gør, i stedet for at producere fritekst, som et menneske skal føre ind.

Det er et reelt skridt: fra et system, der rådgiver, til et system, der handler på tværs af hele arbejdsgangen. Det er også præcis den slags skridt, der i dækningen bliver fladet ud til "AI overgår læger". Derfor er det værd at være præcis om, hvad der blev testet, og hvor.

Versionen i én sætning: MIRA håndterede hele sager på egen hånd og overgik på denne benchmarktest læger i diagnostisk nøjagtighed — men det skete i et sandkassemiljø, på retrospektive journaler, inden for otte på forhånd udvalgte diagnoser, udelukkende med tekstbaseret kommunikation, og en stor del af forspringet kom ved de tilstande, der havde de tydeligste testresultater. Fremskridtet er reelt. Resultattavlen er ikke klinikken.



MIRA demonstrerede evnen til at gennemføre en arbejdsgang fra start til slut i en elektronisk patientjournal i et sandkassemiljø. Det demonstrerede ikke klinisk anvendelse i den virkelige verden, bred diagnostisk dækning eller en retfærdig direkte sammenligning med læger, der havde lige så meget information.

Hvad forfatterne gjorde

MIRA — Medical Intelligence for Reasoning and Action — er en autonom agent bygget på OpenAIs modeller: Den del, der samtaler og handler, kører på **GPT-4o**, og et separat planlægningstrin bruger OpenAIs ræsonneringsmodel **o1**. De indgår i en specialbygget ramme, der følger standarderne — bygget på HL7 FHIR, den datastandard, som virkelige hospitaler bruger til at udveksle journaloplysninger — med en værktøjskasse, der rummer over firs tusind mulige kliniske handlinger. I det sandkassemiljø kan MIRA hente en patients sygehistorie, bestille og fortolke laboratorieprøver, billeddiagnostik og mikrobiologi, opstille en differentialdiagnose og formulere behandlingsplaner — ordinere lægemidler, planlægge kirurgiske indgreb og indlæggelser. Én detalje er værd at fremhæve med det samme: De automatiserede "dommere", der bedømte MIRAs svar, var selv GPT-4o — den samme modelfamilie, som blev testet — hvilket forfatterne supplerede med en speciallæges gennemgang, efter at en fagfælle havde rejst indvendingen.

Testmaterialet kom fra **MIMIC-IV**, en stor, offentligt tilgængelig og afidentificeret database med elektroniske patientjournaler. Blandt cirka 300.000 patienter, der blev behandlet på Beth Israel Deaconess Medical Center mellem 2008 og 2019, udvalgte forfatterne **574 patientforløb** fordelt på **otte måldiagnoser** — mavelidelser og akutmedicinske tilstande, heriblandt blindtarmsbetændelse, bugspytkirtelbetændelse, lungebetændelse og urinvejsinfektion. I hvert forløb begyndte MIRA med et begrænset billede og måtte trin for trin afgøre, hvad der skulle gøres som det næste — samme form som en virkelig udredning, men gennemført op mod en journal, hvor det virkelige udfald allerede var registreret.

Systemets præstation blev derefter sammenlignet med lægers arbejde med de samme forløb.

> **Hvad "autonom agent i en elektronisk patientjournal i et sandkassemiljø" egentlig betyder**
>
> Tre ord udfører meget af arbejdet her, så det er værd at forklare dem.
> **Agent** betyder, at systemet ikke er en chatbot, der besvarer en prompt. Det kører i en løkke: ser på den aktuelle tilstand, vælger en handling (bestil denne prøve, spørg efter den oplysning i sygehistorien), ser resultatet, vælger den næste handling — indtil det når frem til en diagnose og en plan. "Intelligensen" er en sprogmodel; *handleevnen* er den ramme omkring den, der omsætter modellens tekst til tilladte operationer i journalsystemet og fører resultaterne tilbage.
> **Elektronisk patientjournal i et sandkassemiljø** betyder en simuleret, afskærmet kopi af et journalsystem, ikke et system på et hospital i drift. MIRA kan "bestille" en undersøgelse og modtage det resultat, som den pågældende patient faktisk fik, fordi forløbet er historisk, og svaret allerede er registreret. Intet af det, systemet gør, berører en virkelig patient eller en

virkelig klinik.
Autonom betyder, at systemet gennemfører forløbet fra start til slut uden et menneske i løkken — *inde i sandkassemiljøet*. Det betyder **ikke**, at det blev sluppet løs for at behandle patienter uden opsyn. Det er vidt forskellige påstande, og kun den første blev testet.
Endnu en ting om sandkassemiljøet, fordi det er let at forestille sig det forkert: MIRA må *bestille* en hvilken som helst undersøgelse, men systemet kan kun give et resultat tilbage, hvis undersøgelsen **faktisk blev udført** på denne patient i virkeligheden. Bed om et blodpanel, patienten fik taget, og du får de virkelige værdier; bed om en scanning, som ingen bestilte, og du får svaret "N/A — kunne ikke udføres", aldrig et opdigtet tal. Sygehistorien fungerer på samme måde: Hvis journalen ikke indeholder det, MIRA spørger om, siger den simulerede patient ganske enkelt, at vedkommende ikke ved det. MIRA kan altså ikke fremtrylle den ene afgørende undersøgelse, som aldrig blev foretaget — det kan kun arbejde med det, den virkelige udredning tilfældigvis omfattede.
Sondringen er vigtig, fordi overskriftsordet — "autonom" — er det, der lettest kan læses som det sidste.

Hvad de fandt

På benchmarktesten **overgik MIRA læger i diagnostisk nøjagtighed** — men overskriften skjuler tre forskellige tal, og de er lette at blande sammen, så det er værd at holde dem adskilt.

På egen hånd, på tværs af alle **574 forløb**, angav MIRA den rigtige diagnose i **88,9 %** af tilfældene — bedømt op mod den udskrivningsdiagnose, der faktisk var registreret for hver patient i MIMIC-IV. I den direkte sammenligning med mennesker arbejdede lægerne ikke med alle 574 forløb; de arbejdede med en fælles **delmængde på 311 forløb** og brugte de samme journaler og værktøjer, som MIRA havde. På de samme 311 forløb scorede MIRA **87,8 %**, og systemet blev målt mod to særskilt rekrutterede grupper. Den første var en **seniorgruppe**: fire speciallæger med 7 til 11 års erfaring, som scorede **78,1 %**. Den anden var en **gruppe med blandet anciennitet**: hovedsageligt yngre læger under specialisering plus to speciallæger, tættere på den faktiske bemanding på en virkelig akutmodtagelse, som scorede **71,1 %**. Samme forløb, samme værktøjer — og rækkefølgen blev AI først, de erfarne læger som nummer to og den juniortunge gruppe som nummer tre. Artiklens egen forsigtige formulering er, at MIRA "konsekvent var på niveau med og ofte overgik" lægerne på tværs af alle otte sygdomme.

Også beslutningerne senere i forløbet holdt niveau: De var i vid udstrækning i overensstemmelse med retningslinjerne, lægemiddelsikre og passende med hensyn til indlæggelse. Forfatterne rapporterer ingen meget alvorlige medicineringsfejl inden for fem sikkerhedskategorier (lægemiddelinteraktioner, dosering ved nyresygdom, allergier, QT-risiko og ordination af opioider) og korrekte ordinationer i 467 af 468 forløb — samtidig med at de bemærker, at systemet "ikke opnåede 100 % pålidelighed". Taget for pålydende er det et opsigtsvækkende resultat: en agent, der håndterer hele forløbet og ender foran lægerne på diagnosen.

Men sejrens *form* betyder lige så meget som sejren. Uafhængige specialister, der læste artiklen,

pegede på, hvor forspringet faktisk kom fra. I Science Media Centres ekspertpanel bemærkede dr. Wei Xing (University of Sheffield), at overskriftstallet — at AI'en slog lægerne i diagnostisk nøjagtighed — **hovedsageligt blev drevet af tilstandene med tydelige testresultater**, såsom blindtarmsbetændelse og bugspytkirtelbetændelse, hvor en afgørende scanning eller laboratorieværdi afklarer spørgsmålet. Ved lungebetændelse og urinvejsinfektioner — to af de hyppigste grunde til, at mennesker faktisk møder op på en akutmodtagelse — klarede *både* AI'en og lægerne sig dårligst, og forskellen mellem dem var mindst.

Der var også en asymmetri i de to siders fremgangsmåde, og den ses i artiklens egne tal. MIRA støttede sig mere til laboratoriet: Det brugte omkring **51 %** af de laboratorieanalytter, der var tilgængelige i rutinebehandlingen, mod cirka **28 %** for speciallægerne — en median på syv flere blodparametre pr. forløb. Forfatterne er omhyggelige med at beskrive dette som *under* datasættets egen baseline for rutinebehandling, ikke som en strategi, hvor alt bestilles, og de rapporterer *ingen* systematisk stigning i dyrere tværsnitsbilleddannelse. Alligevel kan mere information i sig selv give større diagnostisk nøjagtighed — så dette er ikke helt en sammenligning mellem lige velinformerede parter, et hul, Xing påpegede direkte. En kliniker, der frit kunne bestille alle billige blodprøver uden at afveje omkostninger, ubehag eller forsinkelse, ville også tage sig skarpere ud på papiret.

Hvorfor "slog læger" ikke betyder "bedre læge"

Det er let at overfortolke en sejr på en benchmarktest. Tre ting ligger mellem "MIRA scorede højere" og "MIRA er bedre".

For det første: **hvad systemet blev bedømt op imod.** Professor Julie Jacko (University of Edinburgh) bemærkede, at flere af MIRAs centrale resultater er defineret i forhold til det, der blev dokumenteret i det underliggende datasæt — hvilket betyder, at systemet belønnes for at **genskabe den registrerede kliniske adfærd**, ikke nødvendigvis for at demonstrere optimal behandling. Den historiske journal bliver facitlisten. Det er en rimelig måde at opbygge en benchmarktest på, men den måler overensstemmelse med det, der blev gjort, ikke korrekthed i absolut forstand. For at yde artiklen retfærdighed rammer dette hårdest målene for *overensstemmelse i behandlingen* — matchede MIRAs ordinationer journalen? — mens overskriftens diagnostiske nøjagtighed bedømmes op mod udskrivningsdiagnosen, som ligger tættere på et reelt udfald end et ekko af dokumentationen.

For det andet: den allerede nævnte **informationsasymmetri**. Langt flere blodprøver (omkring 51 % af de tilgængelige analytter mod 28 %) giver mere evidens. En del af forskellen i nøjagtighed er sandsynligvis *købt*, ikke ræsonneret frem.

For det tredje: **hvor systemet kørte.** Dette var et sandkassemiljø med retrospektive forløb, inden for otte på forhånd udvalgte diagnoser og udelukkende i tekst. En virkelig klinisk vurdering afhænger, som dr. Dominic Oliver (University of Oxford) udtrykte det, ikke kun af, hvad patienterne siger, men hvordan de siger det — sammen med fysisk undersøgelse, observeret adfærd og kropssprog, som en tekstbaseret agent, der læser en afsluttet journal, aldrig får at se. Og en patient, hvis problem ikke falder inden for en af de otte valgte diagnoser, ligger ganske enkelt uden for det, denne undersøgelse kan udtale sig om.

Der er også en mere stille bekymring, som fagfællerne rejste: **datakontaminering.** MIMIC-IV er offentlig og meget omtalt, så en sprogmodel, der er trænet på det åbne internet, kan allerede have set artikler, diskussioner af patientforløb eller selve dataene. Dr. Midhun Parakkal Unni (University of Sheffield) fremhævede dette direkte — hvis nogle af svarene fandtes i træningsdataene, skyldes en del af præstationen hukommelse, ikke klinisk ræsonnement, og kun uafhængig replikation kan skelne de to. Det er værd at bemærke, at forfatterne ikke affejer punktet: De skriver, at deres resultater "forsigtigt kunne fortolkes som en mulig øvre grænse" og "kan overvurdere generaliserbarheden til andre offentlige patientforløb" — en artikel, der selv lægger loft over sin egen overskrift.

Intet af dette gør MIRA mindre interessant. Det betyder, at den korrekte læsning er afstemt: På en retrospektiv benchmarktest overgik en agent, der kunne handle gennem hele journalen, læger på diagnoserne med tydelige testresultater — dels ved at bestille flere prøver, dels ved at stemme overens med den registrerede journal. Det er en reel demonstration af en evne, ikke en dom om, at maskiner nu stiller bedre diagnoser end læger.

Hvad dette *ikke* beviser

- Det viser **ikke**, at MIRA virker på virkelige patienter. Hvert forløb var retrospektivt og simuleret; ingen patient blev nogensinde behandlet af systemet.
- Det viser **ikke**, at det virker ud over otte diagnoser. Tilstande uden for det på forhånd udvalgte sæt — det rodede, udifferentierede flertal i medicinen — blev ikke testet.
- Det viser **ikke**, at det er sikkert at tage i brug. Forfatterne skriver selv, at generaliserbarhed, sikkerhed og styring stadig kræver prospektive studier i den virkelige verden.
- Det etablerer **ikke** en retfærdig direkte sammenligning med læger. Systemet brugte langt flere blodprøver (omkring 51 % af de tilgængelige analytter mod 28 %), og flere behandlingsresultater blev delvist bedømt op mod den registrerede journal i stedet for op mod en facitliste for bedst mulig behandling.
- Det viser **ikke**, at resultatet er frit for memorering. De offentlige MIMIC-IV-data kan overlappe med modellens træningsdata, hvilket uafhængig replikation vil skulle udelukke.
- Det betyder **ikke** "AI erstatter læger". Det er beslutningsstøtte, der kan *handle* gennem en journal; eksperternes fælles opfattelse er, at virkelig anvendelse vil foregå i partnerskab med klinikere, som beholder myndigheden og tilfører alt det, en tekstjournal udelader.

Hvor stærk er evidensen?

Del påstanden i to, for evidensen er meget forskellig for de to halvdele.

Som demonstration af en evne — at en sprogmodelagent kan håndtere et helt klinisk forløb i en elektronisk patientjournal i et sandkassemiljø og kæde sygehistorie, undersøgelser, diagnose og ordinationer sammen fra start til slut — er arbejdet reelt nyskabende og rimeligt overbevisende. Det er den del, der er ny, og det er den del, der fortjener opmærksomhed.

Som påstand om overlegenhed — at MIRA er *bedre end læger* — er evidensen afgrænset og bør læses med forsigtighed. Den gælder på en bestemt retrospektiv benchmarktest, med otte diagnoser, en informationsasymmetri (flere prøver), en facitliste hentet fra den historiske journal og en reel mulighed for kontaminering af træningsdata. Det er nok til at sige "agenten præsterede imponerende på denne benchmarktest". Det er ikke nok til at sige "agenten er bedre til at stille diagnoser end en læge", og forfatterne hævder ikke det sidste.

Det mest nyttige standpunkt er hverken "AI slår læger" eller "bare et stykke legetøj". Det er: En ny slags medicinsk AI — en, der *handler* gennem journalen i stedet for at besvare spørgsmål — klarede sig godt på en vanskelig retrospektiv benchmarktest og skal nu bevise sit værd dér, hvor den endnu ikke er blevet prøvet: prospektivt, på virkelige, udifferentierede patienter og under styring.

Hvorfor det er vigtigt

I nogle år har det interessante spørgsmål inden for medicinsk AI gradvist ændret sig. Før var det "kan en model stille den rigtige diagnose?" — og modellerne blev ved med at svare ja, på stadig mere ryddelige testsæt. MIRA markerer skiftet til et vanskeligere spørgsmål: "kan en model *udføre arbejdet* — indsamle, bestille, fortolke, beslutte, handle — gennem en hel arbejdsgang?" Det er et mere nyttigt og mere ærligt spørgsmål, fordi det er i handlingen, at både vanskeligheden og risikoen findes.

Derfor betyder indramningen noget. Overskriftsrefleksen — *AI overgår læger* — peger på den mindst nyskabende og svagest underbyggede del af resultatet. Det virkelig nye er mindre og mere betydningsfuldt: en agent, der kan bevæge sig gennem hele journalen. Hvis den evne holder, ændrer den **arbejdsgangen**, længe før den ændrer, hvem der har ansvaret for den. Lægen bliver ikke erstattet; bestillingen, fortolkningen og opfølgningen af resultater — arbejdsgangen — er dér, et værktøj som dette først får en rolle.

Og det stiller bevisbyrden i den rigtige retning igen. En topplacering på retrospektive forløb er en grund til at gennemføre den prospektive undersøgelse, ikke en erstatning for den. Forfatterne siger det samme. Den ærlige læsning af MIRA er en invitation til den næste undersøgelse — holdt op mod den standard, feltet allerede kender: målt på patienter, ikke på benchmarktest.

Kort fortalt

MIRA er en autonom AI-agent, der arbejder i en elektronisk patientjournal i et sandkassemiljø: Den kan optage sygehistorie, bestille og fortolke laboratorieprøver, billeddiagnostik og mikrobiologi, nå frem til en diagnose og skrive behandlingsplaner. Testet på 574 retrospektive forløb fra den offentlige database MIMIC-IV, inden for otte på forhånd udvalgte diagnoser, overgik den læger i diagnostisk nøjagtighed (88,9 % samlet; 87,8 % mod 78,1 % i direkte sammenligning) og traf beslutninger, som i vid udstrækning fulgte retningslinjerne og var lægemiddelsikre. Men evalueringen var en simulering baseret på tidligere journaler, udelukkende i tekst; en stor del af forspringet kom ved tilstande med entydige testresultater; MIRA brugte langt flere blodprøver end lægerne (omkring 51 % af de tilgæn-

gelige analytter mod 28 %); flere behandlingsresultater blev bedømt op mod det, der var registreret i den oprindelige journal; og det offentlige datasæt medfører en reel risiko for kontaminering af træningsdata — hvilket forfatterne selv kalder en mulig øvre grænse for deres tal. Det virkelige fremskridt er en agent, der handler gennem hele arbejdsgangen i stedet for at besvare isolerede spørgsmål. Forfatterne gør det klart, at generaliserbarhed, sikkerhed og styring stadig kræver prospektive studier i den virkelige verden — hvilket er den korrekte læsning: en imponerende demonstration af en evne, ikke bevis for, at AI stiller bedre diagnoser end læger, og ikke et system, der er klar til en virkelig klinik.

Tjek uden omsvøb

Hvad artiklen viser: En sprogmodelagent (MIRA) kan håndtere et helt klinisk forløb fra start til slut i en elektronisk patientjournal i et sandkassemiljø — sygehistorie, undersøgelser, diagnose, ordinationer — og overgik på en retrospektiv benchmarktest med 574 forløb og otte diagnoser læger i diagnostisk nøjagtighed (88,9 % samlet; 87,8 % mod 78,1 % sammenlignet med speciallæger), uden meget alvorlige medicineringsfejl.

Hvad der er plausibelt, men ikke bevist: At denne evne gavner virkelige, udifferentierede patienter; at forspringet i nøjagtighed afspejler bedre ræsonnement snarere end flere bestilte prøver, overensstemmelse med den registrerede journal eller memorerede offentlige data.

Hvad det ikke viser: At MIRA virker uden for sine otte diagnoser; at det er sikkert at tage i brug; at det slår læger i en retfærdig sammenligning, hvor parterne har lige meget information; at det erstatter klinikere; at resultaterne er frie for kontaminering af træningsdata.

Vigtigste begrænsninger: Retrospektiv, simuleret evaluering udelukkende i tekst; otte på forhånd udvalgte diagnoser; en informationsasymmetri (langt flere blodprøver — omkring 51 % af de tilgængelige analytter mod 28 %, i billige blodprøver, ikke billeddiagnostik); behandlingsresultater, der delvist blev bedømt op mod den historiske journal; samt et offentligt benchmarkdatasæt (MIMIC-IV), som en sprogmodel kan have set under træningen — hvilket forfatterne fremhæver som en mulig øvre grænse for præstationen.

Hvor stor tillid bør en almindelig læser have? Høj tillid til, at MIRA er en reel og nyskabende demonstration af en evne — en agent, der *handler* gennem journalen, ikke bare en chatbot. Lav tillid til, at det er en *bedre diagnostiker end en læge*: Den påstand er afgrænset til én retrospektiv benchmarktest og påvirkes af bestillingen af prøver, bedømmelsen op mod journalen og mulig kontaminering. Forfatternes egen konklusion er den rigtige — dette skal undersøges prospektivt i den virkelige verden, før det betyder noget for patientbehandlingen.

Kilder

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