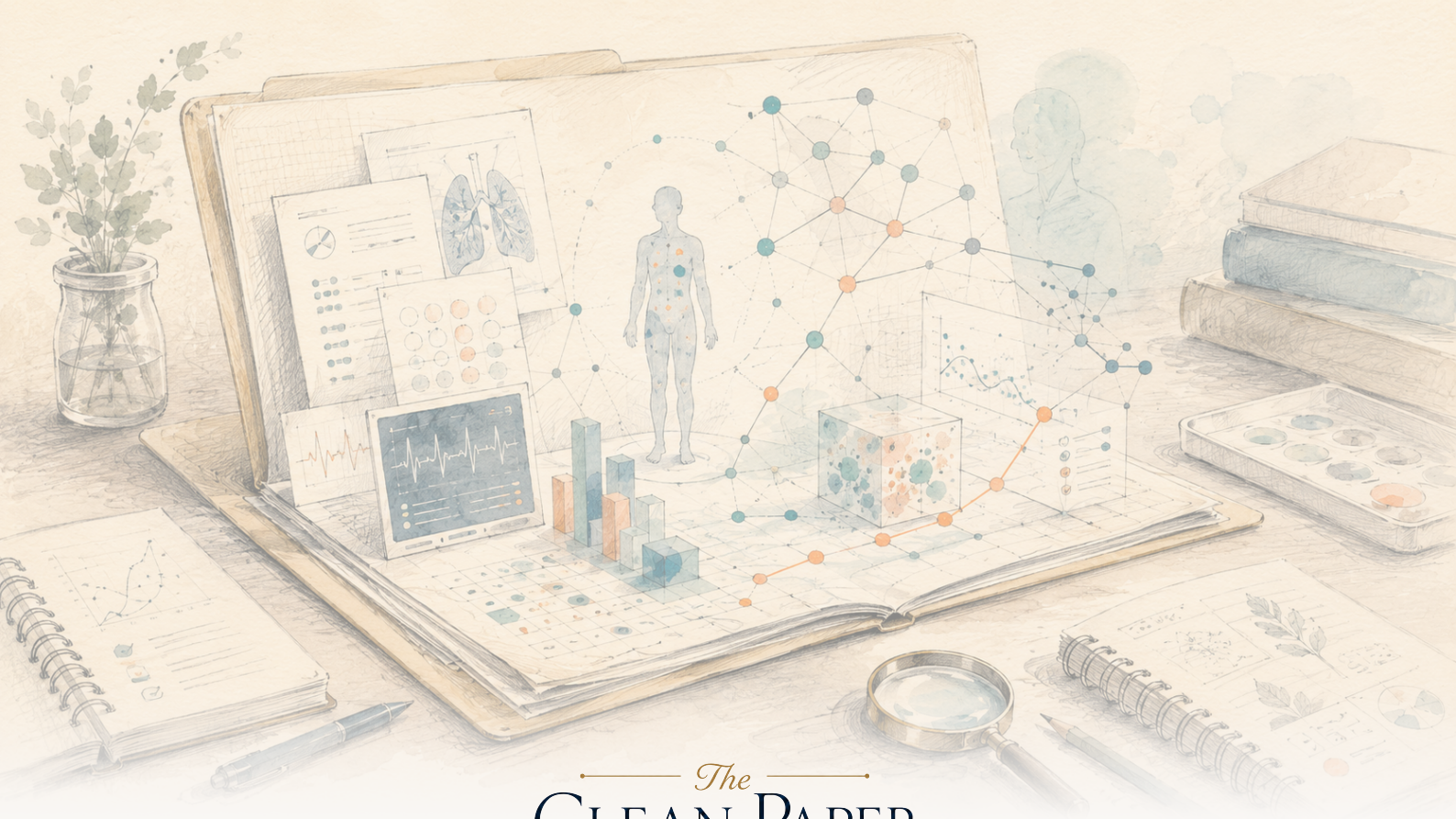


The Clean Paper

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

En KI-agent håndterte hele pasientkasus på egen hånd — i en simulator, basert på tidligere journaler

MIRA er en ny type medisinsk KI: I stedet for å besvare ett spørsmål håndterer den en hel sak i en simulert sykehusjournal — tar opp sykehistorie, bestiller og leser undersøkelser, kommer frem til en diagnose og skriver ordinasjonene. På 574 retrospektive kasus fra en offentlig database, innenfor åtte forhåndsvalgte diagnoser, rapporterer forfatterne at den overgikk leger i diagnostisk treffsikkerhet og tok beslutninger som i stor grad fulgte retningslinjene og var legemiddelsikre. Hvert forbehold i den setningen veier tungt: Den kjørte i et sandkassemiljø med tidligere journaler, bare i tekst; mye av forspranget kom ved tilstander med entydige prøveresultater; den bestilte omtrent dobbelt så mange blodprøver som legene; og flere utfall ble vurdert opp mot det som var registrert i originaljournalen. Det virkelige fremskrittet er en agent som handler gjennom hele arbeidsflyten — ikke bevis for at en maskin nå stiller bedre diagnoser enn en lege. Forfatterne sier det selv: Generaliserbarhet, sikkerhet og styring krever fortsatt 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

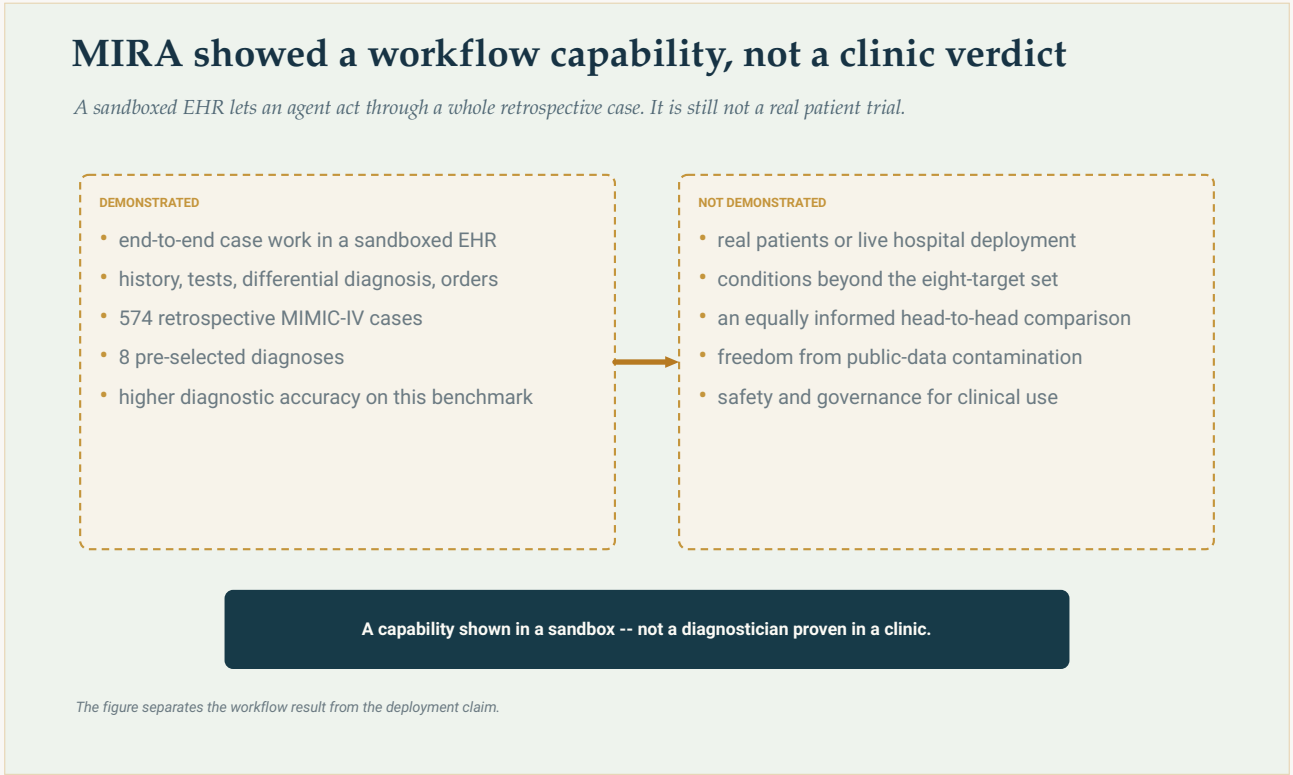
Å besvare et spørsmål er ikke det samme som å håndtere hele saken

De fleste historier om medisinsk KI handler om en modell som *svarer*. Du gir den en kasusbeskrivelse eller en eksamensoppgave, den gir tilbake en diagnose eller et avsnitt med råd, og den oppnår en god poengsum. Nyttig, men snevert: en skarp konsulent som aldri rører journalen.

MIRA er laget for å gjøre noe annet, og nettopp den forskjellen er den egentlige nyheten. I stedet for å besvare ett spørsmål håndterer systemet en hel sak: Det leser pasientjournalen, avgjør hvilke sykehistorieopplysninger det fortsatt trenger, bestiller laboratorieprøver, bildediagnostikk og mikrobiologiske undersøkelser, leser resultatene, snevrer inn en differensialdiagnose og skriver deretter ordinasjonene som følger — resepter, innleggelse, henvisning til kirurgi. Det gjør dette ved å utføre handlinger *inne i* en elektronisk pasientjournal, slik en kliniker gjør, i stedet for å produsere fritekst som et menneske må føre inn.

Det er et reelt skritt: fra et system som gir råd, til et system som handler gjennom hele arbeidsflyten. Det er også akkurat den typen skritt som i mediedekningen flates ut til «KI overgår leger». Derfor er det verdt å være presis om hva som ble testet, og hvor.

Versjonen i én setning: MIRA håndterte hele saker på egen hånd og overgikk på denne referansetesten leger i diagnostisk treffsikkerhet — men det skjedde i et sandkassemiljø, på retrospektive journaler, innenfor åtte forhåndsvalgte diagnoser, med bare tekstbasert kommunikasjon, og mye av forspranget kom ved tilstandene med de tydeligste prøveresultatene. Fremskrittet er reelt. Resultattavlen er ikke klinikken.



MIRA demonstrerte evnen til å gjennomføre en arbeidsflyt fra ende til ende i en elektronisk pasientjournal i et sandkassemiljø. Det demonstrerte ikke klinisk bruk i den virkelige verden, bred diagnostisk dekning eller en rettferdig direkte sammenligning med leger som hadde like mye informasjon.

Original Aurora diagram — The Clean Paper CC BY 4.0

Hva forfatterne gjorde

MIRA — Medical Intelligence for Reasoning and Action — er en autonom agent bygget på OpenAIs modeller: Delen som samtaler og handler, kjører på **GPT-4o**, og et separat planleggingstrinn bruker OpenAIs resonneringsmodell **o1**. Disse inngår i et spesialbygget rammeverk som følger standarder — bygget på HL7 FHIR, datastandarden virkelige sykehus bruker for å utveksle journalopplysninger — med en verktøykasse som rommer over åtti tusen mulige kliniske handlinger. I dette sandkassemiljøet kan MIRA hente en pasients sykehistorie, bestille og tolke laboratorieprøver, bildediagnostikk og mikrobiologi, utarbeide en differensialdiagnose og formulere behandlingsplaner — forskrive legemidler, planlegge kirurgiske inngrep og innleggelser. Én detalj er verdt å fremheve med det samme: De automatiserte «dommerne» som vurderte MIRAs svar, var selv GPT-4o — den samme modellfamilien som ble testet — noe forfatterne supplerte med vurdering fra en spesialistgodkjent lege etter at en fagfelle hadde reist innvendingen.

Testmaterialet kom fra **MIMIC-IV**, en stor, offentlig tilgjengelig og avidentifisert database med elektroniske pasientjournaler. Blant rundt 300 000 pasienter som ble behandlet ved Beth Israel Deaconess Medical Center mellom 2008 og 2019, valgte forfatterne ut **574 pasientkasus** fordelt på **åtte måldiagnoser** — buklidelser og indremedisinske akuttilstander, blant annet appendisitt, pankreatitt, lungebetennelse og urinveisinfeksjon. For hvert kasus startet MIRA med et begrenset bilde og måtte trinn for trinn avgjøre hva som skulle gjøres videre — samme form som en reell utredning, men gjennomført mot en journal der den virkelige avslutningen allerede var registrert.

Ytelsen ble deretter sammenlignet med legers arbeid på de samme kasusene.

Hva «autonom agent i en elektronisk pasientjournal i et sandkassemiljø» egentlig betyr

Tre ord gjør mye av arbeidet her, så det er verdt å forklare dem.
Agent betyr at systemet ikke er en samtalerobot som besvarer en instruksjon. Det kjører en løkke: ser på den nåværende tilstanden, velger en handling (bestill denne prøven, spør etter den opplysningen i sykehistorien), ser resultatet, velger neste handling — helt til det kommer frem til en diagnose og en plan. «Intelligensen» er en språkmodell; *handleevnen* er rammeverket rundt den som gjør modellens tekst om til tillatte operasjoner i journalsystemet og sender resultatene tilbake.
Elektronisk pasientjournal i et sandkassemiljø betyr en simulert, avskjermet kopi av et journalsystem, ikke et system på et sykehus i drift. MIRA kan «bestille» en undersøkelse og motta resultatet den aktuelle pasienten faktisk fikk, fordi saken er historisk og svaret allerede er registrert. Ingenting det gjør, påvirker en ekte pasient eller en ekte klinikk.
Autonom betyr at systemet fullfører saken fra ende til ende uten et menneske i løkken — *inne i*

sandkassemiljøet. Det betyr **ikke** at det ble sluppet løs for å håndtere pasienter uten tilsyn. Dette er svært forskjellige påstander, og bare den første ble testet.
Én ting til om sandkassemiljøet, fordi det er lett å se det for seg feil: MIRA kan *bestille* hvilken undersøkelse det vil, men systemet kan bare gi tilbake et resultat hvis undersøkelsen **faktisk ble utført** på denne pasienten i virkeligheten. Be om et blodpanel pasienten fikk, og du får de virkelige verdiene; be om en skanning som ingen bestilte, og du får svaret «N/A — kunne ikke utføres», aldri et oppdiktet tall. Sykehistorien fungerer på samme måte: Hvis journalen ikke inneholder det MIRA spør om, sier den simulerte pasienten ganske enkelt at den ikke vet. MIRA kan altså ikke trylle frem den ene avgjørende prøven som aldri ble tatt — det kan bare arbeide med det den virkelige utredningen tilfeldigvis omfattet.
Skillet er viktig fordi overskriftsordet — «autonom» — er det som lettest kan forstås som det siste.

Hva de fant

På referansetesten **overgikk MIRA leger i diagnostisk treffsikkerhet** — men overskriften skjuler tre forskjellige tall, og de er lette å blande sammen, så det er verdt å skille dem.

På egen hånd, på tvers av alle **574 kasus**, oppga MIRA riktig diagnose i **88,9 %** av tilfellene — vurdert opp mot utskrivningsdiagnosen som faktisk var registrert for hver pasient i MIMIC-IV. I den direkte sammenligningen med mennesker arbeidet ikke legene med alle 574 kasusene; de arbeidet med et felles **utvalg på 311 kasus**, med de samme journalene og verktøyene som MIRA hadde. På disse 311 kasusene oppnådde MIRA **87,8 %**, og systemet ble målt mot to separat rekrutterte grupper. Den første var en **seniorgruppe**: fire spesialistgodkjente leger med 7 til 11 års erfaring, som oppnådde **78,1 %**. Den andre var en **gruppe med blandet ansiennitet**: for det meste yngre leger i spesialisering samt to spesialister, nærmere den faktiske bemanningen på et virkelig akuttmottak, som oppnådde **71,1 %**. Samme kasus, samme verktøy — og rekkefølgen ble KI først, de erfarne legene som nummer to og den juniortunge gruppen som nummer tre. Artikkelens egen forsiktige formulering er at MIRA «konsekvent var likeverdig med og ofte overgikk» legene på tvers av alle de åtte sykdommene.

Også beslutningene senere i forløpet holdt mål: De var i stor grad i samsvar med retningslinjene, legemiddelsikre og hensiktsmessige når det gjaldt innleggelse. Forfatterne rapporterer ingen svært alvorlige legemiddelfeil i fem sikkerhetskategorier (legemiddelinteraksjoner, nyredosering, allergier, QT-risiko og forskrivning av opioider) og korrekte ordinasjoner i 467 av 468 kasus — samtidig som de påpeker at systemet «ikke oppnådde 100 % pålitelighet». Tatt for det det er, er dette et oppsiktsvekkende resultat: En agent håndterer hele saken og ender foran legene på diagnosen.

Men *formen* på seieren betyr like mye som seieren. Uavhengige spesialister som leste artikkelen, pekte på hvor forspranget faktisk kom fra. I Science Media Centres ekspertpanel påpekte dr. Wei Xing (University of Sheffield) at overskriftstallet — at KI-en slo leger i diagnostisk treffsikkerhet — **i hovedsak ble drevet av tilstandene med tydelige prøveresultater**, som appendisitt og pankreatitt, der en avgjørende skanning eller laboratorieverdi avklarer spørsmålet. For lungebetennelse og urinveisinfeksjoner — to av de vanligste årsakene til at folk faktisk oppsøker en akuttmottakelse — gjorde

både KI-en og legene det dårligst, og forskjellen mellom dem var minst.

Det var også en asymmetri i hvordan de to sidene gikk frem, og den finnes i artikkelens egne tall. MIRA støttet seg mer på laboratoriet: Det brukte rundt **51 %** av laboratorieanalyttene som var tilgjengelige i rutinebehandlingen, mot omtrent **28 %** for de spesialistgodkjente legene — en median på sju flere blodparametere per kasus. Forfatterne er nøye med å beskrive dette som *under* datamaterialets egen grunnlinje for rutinebehandling, ikke som en strategi der alt bestilles, og de rapporterer *ingen* systematisk økning i dyrere tverrsnittsavbildning. Likevel kan mer informasjon i seg selv gi høyere diagnostisk treffsikkerhet — så dette er ikke helt en sammenligning mellom like godt informerte parter, et gap Xing påpekte direkte. En kliniker som fritt kunne bestille alle rimelige blodprøver uten å veie kostnad, ubehag eller forsinkelse, ville også fremstå skarpere på papiret.

Hvorfor «slo leger» ikke betyr «bedre lege»

Det er lett å overtolke en seier på en referansetest. Tre ting ligger mellom «MIRA fikk høyere poengsum» og «MIRA er bedre».

For det første: **hva systemet ble vurdert opp mot.** Professor Julie Jacko (University of Edinburgh) påpekte at flere av MIRAs sentrale utfall er definert i forhold til det som var dokumentert i det underliggende datamaterialet — noe som betyr at systemet belønnes for å **gjenskape den registrerte kliniske atferden**, ikke nødvendigvis for å demonstrere optimal behandling. Den historiske journalen blir fasiten. Det er en rimelig måte å bygge en referansetest på, men den måler samsvar med det som ble gjort, ikke riktighet i absolutt forstand. For å yte artikkelen rettferdighet rammer dette hardest målene for *samsvar i behandlingen* — samsvarte MIRAs ordinasjoner med journalen? — mens den diagnostiske treffsikkerheten i overskriften vurderes opp mot utskrivningsdiagnosen, som ligger nærmere et reelt utfall enn et ekko av dokumentasjonen.

For det andre: den allerede nevnte **informasjonsasymmetrien**. Langt flere blodprøver (rundt 51 % av de tilgjengelige analyttene, mot 28 %) gir mer evidens. Noe av forskjellen i treffsikkerhet er trolig *kjøpt*, ikke resonnert frem.

For det tredje: **hvor systemet kjørte.** Dette var et sandkassemiljø, med retrospektive kasus, innenfor åtte forhåndsvalgte diagnoser og bare i tekst. En virkelig klinisk vurdering avhenger, som dr. Dominic Oliver (University of Oxford) uttrykte det, ikke bare av hva pasientene sier, men hvordan de sier det — sammen med fysisk undersøkelse, observert atferd og kroppsspråk, som en tekstbasert agent som leser en ferdig journal, aldri får se. Og en pasient med et problem som ikke faller innenfor én av de åtte valgte diagnosene, ligger ganske enkelt utenfor det denne studien kan si noe om.

Det finnes også en stillere bekymring som fagfellene tok opp: **datakontaminering.** MIMIC-IV er offentlig og mye omtalt, så en språkmodell som er trent på det åpne internettet, kan allerede ha sett artikler, kasusdiskusjoner eller selve dataene. Dr. Midhun Parakkal Unni (University of Sheffield) påpekte dette direkte — hvis noen av svarene fantes i treningsdataene, skyldes en del av ytelsen hukommelse, ikke klinisk resonnering, og bare uavhengig replikasjon kan skille de to. Det er verdt å merke seg at forfatterne ikke avfeier poenget: De skriver at resultatene deres «forsiktig kan tolkes som en mulig øvre grense» og «kan overvurdere generaliserbarheten til andre offentlige kasus» —

en artikkel som selv setter et tak på sin egen overskrift.

Ingenting av dette gjør MIRA mindre interessant. Det betyr at den riktige lesningen er avstemt: På en retrospektiv referansetest overgikk en agent som kunne handle gjennom hele journalen, leger på diagnosene med tydelige prøveresultater — delvis ved å bestille flere prøver, delvis ved å samsvare med den registrerte journalen. Det er en reell demonstrasjon av en evne, ikke en dom om at maskiner nå stiller bedre diagnoser enn leger.

Hva dette ikke beviser

- Det viser **ikke** at MIRA fungerer på virkelige pasienter. Hvert kasus var retrospektivt og simulert; systemet hadde aldri ansvar for en virkelig pasient.
- Det viser **ikke** at det fungerer utover åtte diagnoser. Tilstander utenfor det forhåndsvalgte settet — det rotete, udifferensierte flertallet i medisinen — ble ikke testet.
- Det viser **ikke** at det er trygt å ta i bruk. Forfatterne skriver selv at generaliserbarhet, sikkerhet og styring fortsatt trenger prospektive studier i den virkelige verden.
- Det etablerer **ikke** en rettferdig direkte sammenligning med leger. Systemet brukte langt flere blodprøver (rundt 51 % av de tilgjengelige analyttene mot 28 %), og flere behandlingsutfall ble delvis vurdert mot den registrerte journalen i stedet for mot en fasit for best mulig behandling.
- Det viser **ikke** at resultatet er fritt for memorering. De offentlige MIMIC-IV-dataene kan overlappe med modellens treningsdata, noe uavhengig replikasjon må utelukke.
- Det betyr **ikke** «KI erstatter leger». Det er beslutningsstøtte som kan *handle* gjennom en journal; ekspertenes samlede syn er at virkelig bruk vil skje i samarbeid med klinikere, som beholder myndigheten og tilfører alt en tekstjournal utelater.

Hvor sterke er bevisene?

Del påstanden i to, fordi bevisene er svært forskjellige for hver halvdel.

Som demonstrasjon av en evne — at en språkmodellagent kan håndtere en hel klinisk sak i et elektronisk journalsystem i et sandkassemiljø og koble sammen sykehistorie, undersøkelser, diagnose og ordinasjoner fra ende til ende — er arbeidet genuint nyskapende og rimelig overbevisende. Dette er den nye delen, og det er den som fortjener oppmerksomhet.

Som påstand om overlegenhet — at MIRA er *bedre enn leger* — er bevisene avgrensede og bør leses med varsomhet. De gjelder på en bestemt retrospektiv referansetest, med åtte diagnoser, en informasjonsasymmetri (flere prøver), en fasit hentet fra den historiske journalen og en reell mulighet for kontaminering av treningsdata. Det er nok til å si «agenten presterte imponerende på denne referansetesten». Det er ikke nok til å si «agenten er bedre til å stille diagnoser enn en lege», og forfatterne hevder ikke det siste.

Det mest nyttige standpunktet er verken «KI slår leger» eller «bare en leke». Det er: En ny type

medisinsk KI — en som *handler* gjennom journalen i stedet for å besvare spørsmål — gjorde det godt på en vanskelig retrospektiv referansetest og må nå bevise seg der den ennå ikke er prøvd: prospektivt, på virkelige, udifferensierte pasienter og under styring.

Hvorfor det er viktig

I noen år har det interessante spørsmålet innen medisinsk KI gradvis endret seg. Før var det «kan en modell stille riktig diagnose?» — og modellene fortsatte å svare ja, på stadig ryddigere testsett. MIRA markerer overgangen til et vanskeligere spørsmål: «kan en modell *gjøre jobben* — innhente, bestille, tolke, beslutte, handle — gjennom en hel arbeidsflyt?» Det er et mer nyttig og et mer ærlig spørsmål, fordi det er i handlingen både vanskeligheten og risikoen ligger.

Derfor betyr innrammingen noe. Overskriftsrefleksen — *KI overgår leger* — peker på den minst nyskapende og svakest underbygde delen av resultatet. Det virkelig nye er mindre og mer betydningsfullt: en agent som kan bevege seg gjennom hele journalen. Hvis denne evnen holder, endrer den **arbeidsflyten** lenge før den endrer hvem som har ansvaret for den. Legen blir ikke erstattet; bestillingen, tolkningen og oppfølgingen av resultater — arbeidsflyten — er der et slikt verktøy først får en rolle.

Og det stiller bevisbyrden i riktig retning igjen. En topplassering på retrospektive kasus er en grunn til å gjennomføre den prospektive studien, ikke en erstatning for den. Forfatterne sier det samme. Den ærlige lesningen av MIRA er en invitasjon til den neste studien — holdt opp mot standarden feltet allerede kjenner: målt på pasienter, ikke på referansetester.

Kort oppsummert

MIRA er en autonom KI-agent som arbeider i en elektronisk pasientjournal i et sandkassemiljø: Den kan ta opp sykehistorie, bestille og tolke laboratorieprøver, bildediagnostikk og mikrobiologi, komme frem til en diagnose og skrive behandlingsplaner. Testet på 574 retrospektive kasus fra den offentlige databasen MIMIC-IV, innenfor åtte forhåndsvalgte diagnoser, overgikk den leger i diagnostisk treffsikkerhet (88,9 % totalt; 87,8 % mot 78,1 % i direkte sammenligning) og tok beslutninger som i stor grad fulgte retningslinjene og var legemiddelsikre. Men evalueringen var en simulering basert på tidligere journaler, bare i tekst; mye av forspranget kom ved tilstander med entydige prøveresultater; MIRA brukte langt flere blodprøver enn legene (rundt 51 % av de tilgjengelige analyttene mot 28 %); flere behandlingsutfall ble vurdert mot det som var registrert i originaljournalen; og det offentlige datamaterialet medfører en reell risiko for kontaminering av treningsdata — noe forfatterne selv kaller en mulig øvre grense for tallene. Det virkelige fremskrittet er en agent som handler gjennom hele arbeidsflyten i stedet for å besvare isolerte spørsmål. Forfatterne er tydelige på at generaliserbarhet, sikkerhet og styring fortsatt krever prospektive studier i den virkelige verden — som er den riktige lesningen: en imponerende demonstrasjon av en evne, ikke bevis for at KI stiller bedre diagnoser enn leger, og ikke et system som er klart for en virkelig klinikk.

Nøktern vurdering

Hva artikkelen viser: En språkmodellagent (MIRA) kan håndtere en hel klinisk sak fra ende til ende i en elektronisk pasientjournal i et sandkassemiljø — sykehistorie, undersøkelser, diagnose, ordinasjoner — og overgikk på en retrospektiv referansetest med 574 kasus og åtte diagnoser leger i diagnostisk treffsikkerhet (88,9 % totalt; 87,8 % mot 78,1 % sammenlignet med spesialistgodkjente leger), uten svært alvorlige legemiddelfeil.

Hva som er plausibelt, men ikke bevist: At denne evnen gir nytte for virkelige, udifferensierte pasienter; at forspranget i treffsikkerhet gjenspeiler bedre resonnering fremfor flere bestilte prøver, samsvar med den registrerte journalen eller memorerte offentlige data.

Hva det ikke viser: At MIRA fungerer utenfor sine åtte diagnoser; at det er trygt å ta i bruk; at det slår leger i en rettferdig sammenligning der partene har like mye informasjon; at det erstatter klinikere; at resultatene er frie for kontaminering av treningsdata.

Viktigste begrensninger: Retrospektiv, simulert evaluering bare i tekst; åtte forhåndsvalgte diagnoser; en informasjonsasymmetri (langt flere blodprøver — rundt 51 % av de tilgjengelige analyttene mot 28 %, i rimelige blodprøver, ikke bildediagnostikk); behandlingsutfall som delvis ble vurdert mot den historiske journalen; og et offentlig referansemateriale (MIMIC-IV) som en språkmodell kan ha sett under trening — noe forfatterne omtaler som en mulig øvre grense for ytelsen.

Hvor stor tillit bør en vanlig leser ha? Høy tillit til at MIRA er en reell og nyskapende demonstrasjon av en evne — en agent som *handler* gjennom journalen, ikke bare en samtalerobot. Lav tillit til at det er en *bedre diagnostiker enn en lege*: Den påstanden er avgrenset til én retrospektiv referansetest og påvirkes av prøvebestillingen, vurderingen mot journalen og mulig kontaminering. Forfatternes egen konklusjon er den riktige — dette må studeres prospektivt i den virkelige verden før det betyr noe for pasientbehandlingen.

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