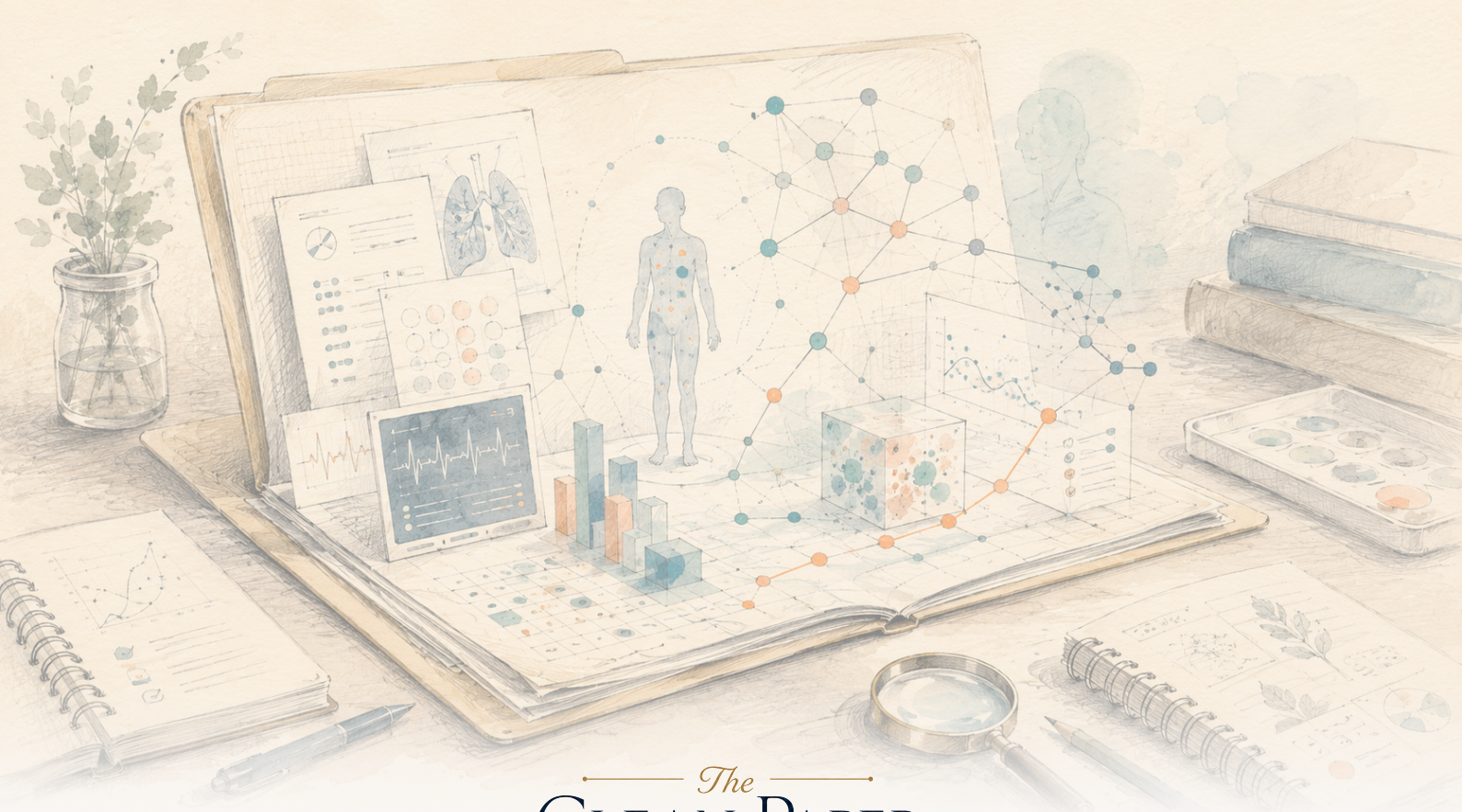


The Clean Paper

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

En AI-agent handlade hela patientfall på egen hand — i en simulator, utifrån tidigare journaler

MIRA är en ny sorts medicinsk AI: i stället för att besvara en enda fråga handlägger den ett helt fall i en simulerad sjukhusjournal — tar anamnes, beställer och läser undersökningar, når en diagnos och skriver ordinationerna. På 574 retrospektiva fall från en offentlig databas, inom åtta förvalda diagnoser, rapporterar författarna att den överträffade läkare i diagnostisk träffsäkerhet och fattade beslut som till stor del följde riktlinjer och var läkemedelssäkra. Varje förbehåll i den meningen väger tungt: den kördes i en sandlådemiljö med tidigare journaler, enbart i text; en stor del av försprånget kom vid tillstånd med entydiga testresultat; den beställde ungefär dubbelt så många blodprover som läkarna; och flera utfall bedömdes mot vad originaljournalen hade registrerat. Det verkliga framsteget är en agent som agerar genom hela arbetsflödet — inte ett bevis för att en maskin nu ställer bättre diagnoser än en läkare. Författarna säger det själva: generaliserbarhet, säkerhet och styrning kräver fortfarande prospektiva studier i verklig vård.

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

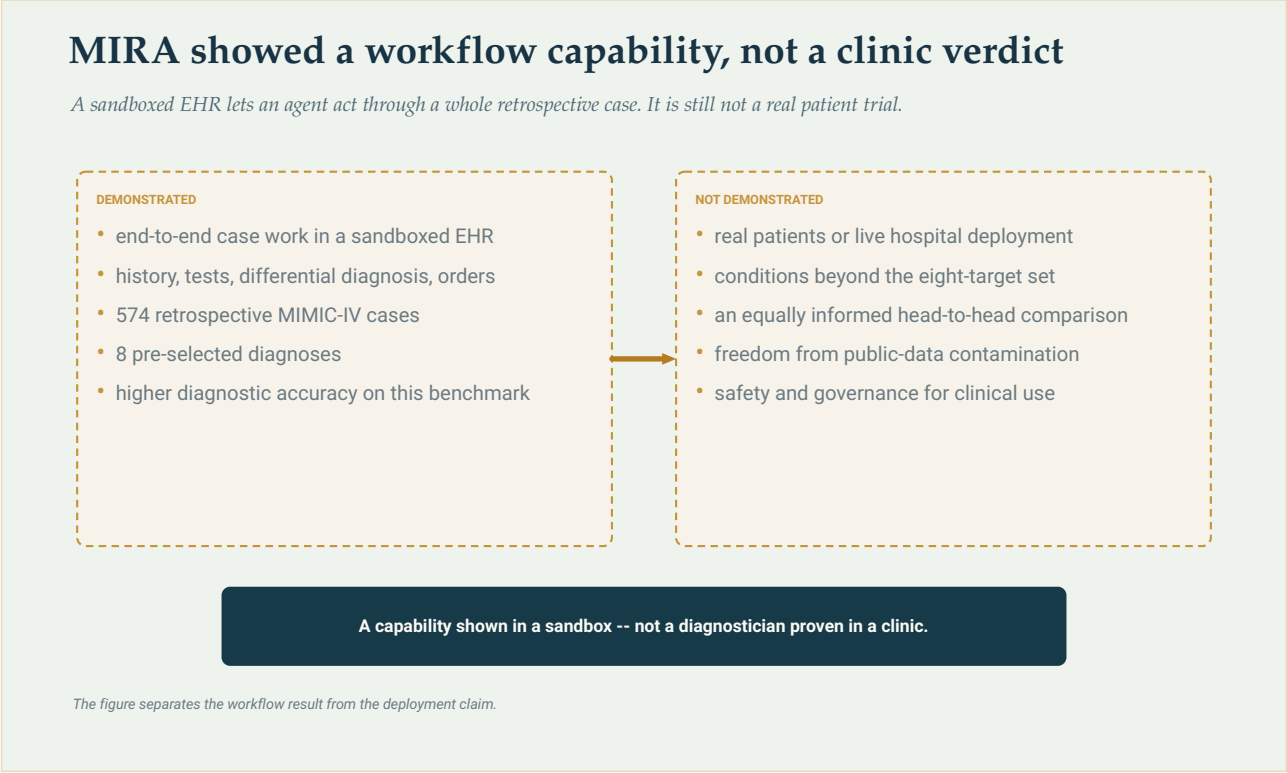
Att besvara en fråga är inte samma sak som att handlägga fallet

De flesta berättelser om medicinsk AI handlar om en modell som *svarar*. Man ger den en fallbeskrivning eller en tentamensfråga, den ger tillbaka en diagnos eller ett stycke med råd och får ett bra resultat. Användbart, men snävt: en klipsk konsult som aldrig rör journalen.

MIRA är byggt för att göra något annat, och just den skillnaden är den verkliga nyheten. I stället för att besvara en enda fråga handlägger systemet ett helt fall: det läser patientens journal, avgör vilka anamnesuppgifter som fortfarande behövs, beställer laboratorie-, bilddiagnostiska och mikrobiologiska undersökningar, läser resultaten, snävar in en differentialdiagnos och skriver sedan de ordinationer som följer — recept, inläggning, remiss till operation. Det gör detta genom att vidta åtgärder *inne i* en elektronisk patientjournal, på samma sätt som en kliniker, i stället för att producera fritext som en människa ska föra in.

Det är ett verkligt steg: från ett system som ger råd till ett system som agerar genom hela arbetsflödet. Det är också precis den sorts steg som i rapporteringen plattas till ”AI överträffar läkare”. Därför är det värt att vara exakt om vad som testades och var.

Versionen i en enda mening: MIRA handlade hela fall på egen hand och överträffade på detta benchmark läkare i diagnostisk träffsäkerhet — men det skedde i en sandlådemiljö, på retrospektiva journaler, inom åtta förvalda diagnoser, med enbart textkommunikation, och en stor del av försprånget kom vid tillstånd med de tydligaste testresultaten. Framsteget är verkligt. Resultattavlan är inte kliniken.



MIRA visade förmåga att genomföra ett arbetsflöde från början till slut i en journalbaserad sandlådemiljö. Det visade inte verklig klinisk användning, bred diagnostisk täckning eller en rättvis direkt jämförelse med läkare som hade tillgång till lika mycket information.

Vad författarna gjorde

MIRA — Medical Intelligence for Reasoning and Action — är en autonom agent som bygger på OpenAI:s modeller: den del som samtalar och agerar körs med **GPT-4o**, och ett separat planeringssteg använder OpenAI:s resonemangsmodell **o1**. De ingår i ett specialbyggt ramverk som följer standarder — byggt på HL7 FHIR, den datastandard som verkliga sjukhus använder för att överföra journaluppgifter — med en verktygslåda som rymmer över åttio tusen möjliga kliniska åtgärder. I den sandlådemiljön kan MIRA hämta patientens anamnes, beställa och tolka laboratorieprover, bilddiagnostik och mikrobiologi, skapa en differentialdiagnos och formulera behandlingsplaner — ordinera läkemedel, planera kirurgiska ingrepp och inläggningar. En detalj som är värd att framhålla från början: de automatiserade ”domare” som poängsatte MIRA:s svar var själva GPT-4o — samma modellfamilj som testades — vilket författarna kompletterade med granskning av en specialistläkare efter att en sakkunnig granskare hade tagit upp invändningen.

Testmaterialet kom från **MIMIC-IV**, en stor, allmänt tillgänglig och avidentifierad databas med elektroniska patientjournaler. Bland omkring 300 000 patienter som behandlades vid Beth Israel Deaconess Medical Center mellan 2008 och 2019 valde författarna ut **574 patientfall** fördelade på **åtta måldiagnoser** — bukåkommor och internmedicinska akuttillstånd, däribland appendicit, pankreatit, pneumoni och urinvägsinfektion. För varje fall utgick MIRA från en begränsad bild och behövde steg för steg avgöra vad som skulle göras härnäst — samma form som en verklig utredning, men genomförd mot en journal där det verkliga slutresultatet redan fanns registrerat.

Prestationen jämfördes sedan med läkares arbete med samma fall.

Vad ”autonom agent i en journalbaserad sandlådemiljö” egentligen betyder

Tre ord gör mycket arbete här, så det är värt att reda ut dem.
Agent betyder att systemet inte är en chattbot som besvarar en prompt. Det kör en loop: granskar det aktuella läget, väljer en åtgärd (beställ det här provet, fråga efter den uppgiften i anamnesen), ser resultatet, väljer nästa åtgärd — tills det når en diagnos och en plan. ”Intelligensen” är en språkmodell; *handlingsförmågan* är strukturen runt omkring som omvandlar modellens text till tillåtna operationer i journalsystemet och matar tillbaka resultaten.
Journalbaserad sandlådemiljö betyder en simulerad, avskärmad kopia av ett journalsystem, inte ett system på ett sjukhus i drift. MIRA kan ”beställa” en undersökning och få det resultat som patienten faktiskt fick, eftersom fallet är historiskt och svaret redan finns registrerat. Inget det gör påverkar en verklig patient eller en verklig klinik.
Autonom betyder att systemet genomför fallet från början till slut utan att en människa deltar — *inne i sandlådemiljön*. Det betyder **inte** att det släpptes fritt för att handlägga patienter utan

tillsyn. Det är två helt olika påståenden, och bara det första testades.
Ytterligare en sak om sandlådemiljön, eftersom den är lätt att föreställa sig fel: MIRA får *beställa* vilken undersökning det vill, men systemet kan bara lämna tillbaka ett resultat om undersökningen **faktiskt utfördes** på just den patienten i verkligheten. Fråga efter en blodpanel som patienten fick, så kommer de verkliga värdena; fråga efter en skanning som ingen beställde, så blir svaret ”N/A — kunde inte utföras”, aldrig ett påhittat tal. Anamnesen fungerar på samma sätt: om journalen inte innehåller det som MIRA frågar om säger den simulerade patienten helt enkelt att den inte vet. MIRA kan alltså inte trolla fram det avgörande test som aldrig gjordes — det kan bara arbeta med det som råkade ingå i den verkliga utredningen.
Skillnaden spelar roll eftersom rubrikordet — ”autonom” — är det som lättast kan läsas som det senare.

Vad de fann

På benchmarktestet **överträffade MIRA läkare i diagnostisk träffsäkerhet** — men rubriken döljer tre olika tal, och de är lätta att blanda ihop, så det är värt att hålla isär dem.

På egen hand, sett över samtliga **574 fall**, angav MIRA rätt diagnos i **88,9 %** av fallen — bedömt mot den utskrivningsdiagnos som faktiskt hade registrerats för varje patient i MIMIC-IV. I den direkta jämförelsen med människor arbetade läkarna inte med alla 574 fallen; de arbetade med en gemensam **delmängd på 311 fall**, med samma journaler och verktyg som MIRA. På dessa 311 fall fick MIRA **87,8 %**, och systemet jämfördes med två separat rekryterade grupper. Den första var en **senior grupp**: fyra specialistläkare med 7 till 11 års erfarenhet, som fick **78,1 %**. Den andra var en **grupp med blandad senioritet**: främst yngre ST-läkare samt två specialister, vilket ligger närmare hur en verklig akutmottagning faktiskt är bemannad, och som fick **71,1 %**. Samma fall, samma verktyg — och ordningen blev AI först, de erfarna läkarna tvåa och den grupp som dominerades av yngre läkare trea. Artikelns egen försiktiga formulering är att MIRA ”genomgående var likvärdigt med och ofta överträffade” läkarna inom samtliga åtta sjukdomar.

Även besluten längre fram i kedjan höll måttet: de följde till stor del riktlinjer, var läkemedelssäkra och lämpliga i fråga om inläggning. Författarna rapporterar inga allvarliga läkemedelsfel inom fem säkerhetskategorier (läkemedelsinteraktioner, dosering vid njursjukdom, allergier, QT-risk och opioidförskrivning) och korrekta ordinationer i 467 av 468 fall — samtidigt som de konstaterar att systemet ”inte uppnådde 100 % tillförlitlighet”. Taget för vad det är, är detta ett slående resultat: en agent som driver hela fallet och hamnar före läkarna på diagnosen.

Men segerns *form* är lika viktig som segern. Oberoende specialister som läste artikeln pekade på var försprånget faktiskt uppstod. I Science Media Centres expertpanel noterade dr Wei Xing (University of Sheffield) att rubriksiffran — att AI:n slog läkarna i diagnostisk träffsäkerhet — **till största delen drevs av tillstånden med tydliga testresultat**, såsom appendicit och pankreatit, där en avgörande skanning eller ett laboratorievärde avgör saken. Vid pneumoni och urinvägsinfektioner — två av de vanligaste orsakerna till att människor faktiskt kommer till en akutmottagning — presterade *både*

AI:n och läkarna sämst, och skillnaden mellan dem var minst.

Det fanns också en asymmetri i hur de två sidorna gick till väga, och den syns i artikelns egna siffror. MIRA lutade sig mer mot laboratoriet: det använde omkring **51 %** av de laboratorieanalyter som fanns tillgängliga i rutinsjukvården, jämfört med ungefär **28 %** för specialistläkarna — en median på sju fler blodparametrar per fall. Författarna är noga med att beskriva detta som *under* datamaterialets egen baslinje för rutinsjukvård, inte som en strategi där allt beställs, och de rapporterar *ingen* systematisk ökning av dyrare tvärsnittsavbildning. Ändå kan mer information i sig ge högre diagnostisk träffsäkerhet — så detta är inte riktigt en jämförelse mellan lika välinformerade parter, ett glapp som Xing lyfte direkt. En kliniker som fritt kunde beställa alla billiga blodprover utan att väga in kostnad, obehag eller fördröjning skulle också se skarpare ut på papperet.

Varför ”slog läkare” inte betyder ”bättre läkare”

Det är lätt att läsa in för mycket i en benchmarkseger. Tre saker ligger mellan ”MIRA fick högre poäng” och ”MIRA är bättre”.

För det första: **vad systemet bedömdes mot.** Professor Julie Jacko (University of Edinburgh) noterade att flera av MIRA:s centrala utfall definieras i förhållande till vad som dokumenterats i det underliggande datamaterialet — vilket innebär att systemet belönas för att **återskapa det registrerade kliniska handlandet**, inte nödvändigtvis för att visa optimal vård. Den historiska journalen blir facit. Det är ett rimligt sätt att bygga ett benchmarktest, men det mäter överensstämmelse med vad som gjordes, inte korrekthet i någon absolut mening. För att göra artikeln rättvisa slår detta hårdast mot måtten på *överensstämmelse i behandlingen* — motsvarade MIRA:s ordinationer journalen? — medan den diagnostiska träffsäkerheten i rubriken bedöms mot utskrivningsdiagnosen, som ligger närmare ett verkligt utfall än ett eko av dokumentationen.

För det andra: den redan nämnda **informationsasymmetrin**. Betydligt fler blodprover (omkring 51 % av de tillgängliga analyterna, jämfört med 28 %) ger mer evidens. En del av skillnaden i träffsäkerhet är sannolikt *köpt*, inte resonerad fram.

För det tredje: **var systemet kördes.** Detta var en sandlådemiljö med retrospektiva fall, inom åtta förvalda diagnoser och enbart i text. Verklig klinisk bedömning beror, som dr Dominic Oliver (University of Oxford) uttryckte det, inte bara på vad patienter säger utan också på hur de säger det — tillsammans med kroppsundersökning, observerat beteende och kroppsspråk, sådant som en textbaserad agent som läser en färdig journal aldrig ser. Och en patient vars problem inte faller inom någon av de åtta valda diagnoserna ligger helt enkelt utanför vad denna studie kan uttala sig om.

Det finns också en mer lågmäld farhåga som granskarna tog upp: **datakontaminering.** MIMIC-IV är offentligt och har diskuterats mycket i skrift, så en språkmodell som tränats på det öppna internet kan redan ha sett artiklar, falldiskussioner eller själva datamaterialet. Dr Midhun Parakkal Unni (University of Sheffield) lyfte detta direkt — om några av svaren fanns i träningsdata är en del av prestationen minne, inte kliniskt resonemang, och bara oberoende replikation kan skilja de två åt. Noterbart är att författarna inte viftar bort saken: de skriver att deras resultat ”försiktigt skulle kunna tolkas som en möjlig övre gräns” och ”kan överskatta generaliserbarheten till andra offentliga fall”

— en artikel som själv sätter ett tak på sin rubrik.

Inget av detta gör MIRA mindre intressant. Det innebär att den korrekta läsningen är avvägd: på ett retrospektivt benchmarktest överträffade en agent som kunde agera genom hela journalen läkare på diagnoser med tydliga testresultat — delvis genom att beställa fler prover, delvis genom att stämma överens med den registrerade journalen. Det är en verklig demonstration av en förmåga, inte en dom som säger att maskiner nu ställer bättre diagnoser än läkare.

Vad detta *inte* bevisar

- Det visar **inte** att MIRA fungerar på verkliga patienter. Varje fall var retrospektivt och simulerat; systemet ansvarade aldrig för en verklig patient.
- Det visar **inte** att det fungerar utanför åtta diagnoser. Tillstånd utanför den förvalda uppsättningen — den röriga, odifferentierade majoriteten av medicinen — testades inte.
- Det visar **inte** att det är säkert att införa. Författarna skriver själva att generaliserbarhet, säkerhet och styrning fortfarande behöver prospektiva studier i verklig vård.
- Det fastställer **inte** en rättvis direkt jämförelse med läkare. Systemet använde betydligt fler blodprover (omkring 51 % av de tillgängliga analyterna mot 28 %), och flera behandlingsutfall bedömdes delvis mot den registrerade journalen snarare än mot ett facit för bästa möjliga vård.
- Det visar **inte** att resultatet är fritt från memorering. De offentliga MIMIC-IV-data kan överlappa modellens träningsdata, något som oberoende replikation skulle behöva utesluta.
- Det betyder **inte** ”AI ersätter läkare”. Det är beslutsstöd som kan *agera* genom en journal; granskarnas samlade uppfattning är att verklig användning kommer att ske i partnerskap med kliniker, som behåller befogenheten och tillför allt som en textjournal utelämnar.

Hur starka är beläggen?

Dela påståendet i två, eftersom beläggen ser mycket olika ut för de båda delarna.

Som demonstration av en förmåga — att en språkmodellsagent kan driva ett helt kliniskt fall i en journalbaserad sandlådemiljö och länka samman anamnes, undersökningar, diagnos och ordinationer från början till slut — är arbetet genuint nyskapande och rimligt övertygande. Det här är den nya delen, och det är den som förtjänar uppmärksamhet.

Som påstående om överlägsenhet — att MIRA är *bättre än läkare* — är beläggen begränsade och bör läsas försiktigt. Det gäller på ett specifikt retrospektivt benchmarktest, med åtta diagnoser, en informationsasymmetri (fler prover), ett facit hämtat från den historiska journalen och en reell möjlighet till kontaminering av träningsdata. Det räcker för att säga ”agenten presterade imponerande på detta benchmarktest”. Det räcker inte för att säga ”agenten ställer bättre diagnoser än en läkare”, och författarna hävdar inte det senare.

Den mest användbara hållningen är varken ”AI slår läkare” eller ”bara en leksak”. Den är: en ny sorts

medicinsk AI — en som *agerar* genom journalen i stället för att besvara frågor — gjorde väl ifrån sig på ett svårt retrospektivt benchmarktest och måste nu bevisa sig där den ännu inte har prövats: prospektivt, på verkliga, odifferentierade patienter och under styrning.

Varför det spelar roll

Under några år har den intressanta frågan inom medicinsk AI gradvis förändrats. Förr var den ”kan en modell få diagnosen rätt?” — och modellerna fortsatte att svara ja, på allt mer tillrättalagda testmaterial. MIRA markerar skiftet till en svårare fråga: ”kan en modell *utföra arbetet* — samla in, beställa, tolka, besluta, agera — genom ett helt arbetsflöde?” Det är en mer användbar fråga och en ärligare fråga, eftersom det är i handlandet som både svårigheten och risken finns.

Därför spelar inramningen roll. Rubrikreflexen — *AI överträffar läkare* — pekar mot den minst nyskapande och sämst underbyggda delen av resultatet. Det verkligt nya är mindre och mer betydelsefullt: en agent som kan ta sig igenom hela journalen. Om den förmågan håller ändrar den **arbetsflödet** långt innan den ändrar vem som bestämmer över det. Läkaren ersätts inte; beställandet, tolkandet och bevakningen av resultat — arbetsflödet — är där ett sådant verktyg först får en roll.

Och det återställer bevisbördan i rätt riktning. En topplacering på retrospektiva fall är ett skäl att genomföra den prospektiva studien, inte en ersättning för den. Författarna säger detsamma. Den ärliga läsningen av MIRA är en inbjudan till nästa studie — bedömd efter den standard som fältet redan känner till: mätt på patienter, inte på benchmarktester.

Kort sammanfattning

MIRA är en autonom AI-agent som arbetar i en elektronisk patientjournal i en sandlådemiljö: den kan ta anamnes, beställa och tolka laboratorieprover, bilddiagnostik och mikrobiologi, nå en diagnos och skriva behandlingsplaner. Testad på 574 retrospektiva fall från den offentliga databasen MIMIC-IV, inom åtta förvalda diagnoser, överträffade den läkare i diagnostisk träffsäkerhet (88,9 % totalt; 87,8 % mot 78,1 % i direkt jämförelse) och fattade beslut som till stor del följde riktlinjer och var läkemedelssäkra. Men utvärderingen var en simulering med tidigare journaler, enbart i text; en stor del av försprånget kom vid tillstånd med entydiga testresultat; MIRA använde betydligt fler blodprover än läkarna (omkring 51 % av de tillgängliga analyterna mot 28 %); flera behandlingsutfall bedömdes mot vad som registrerats i originaljournalen; och det offentliga datamaterialet medför en verklig risk för kontaminering av träningsdata — något som författarna själva kallar en möjlig övre gräns för sina siffror. Det verkliga framsteget är en agent som agerar genom hela arbetsflödet i stället för att besvara isolerade frågor. Författarna är tydliga med att generaliserbarhet, säkerhet och styrning fortfarande kräver prospektiva studier i verklig vård — vilket är den korrekta läsningen: en imponerande demonstration av en förmåga, inte ett bevis för att AI ställer bättre diagnoser än läkare och inte ett system som är redo för en verklig klinik.

Rakt på sak

Vad artikeln visar: En språkmodellsagent (MIRA) kan driva ett helt kliniskt fall från början till slut i en journalbaserad sandlådemiljö — anamnes, undersökningar, diagnos, ordinationer — och överträffade på ett retrospektivt benchmarktest med 574 fall och åtta diagnoser läkare i diagnostisk träffsäkerhet (88,9 % totalt; 87,8 % mot 78,1 % jämfört med specialistläkare), utan några allvarliga läkemedelsfel.

Vad som är rimligt men inte bevisat: Att denna förmåga ger nytta för verkliga, odifferentierade patienter; att försprånget i träffsäkerhet återspeglar bättre resonemang snarare än fler beställda prover, överensstämmelse med den registrerade journalen eller memorerade offentliga data.

Vad det inte visar: Att MIRA fungerar utanför sina åtta diagnoser; att det är säkert att införa; att det slår läkare i en rättvis jämförelse där parterna har lika mycket information; att det ersätter kliniker; att resultaten är fria från kontaminering av träningsdata.

Huvudsakliga begränsningar: Retrospektiv, simulerad utvärdering enbart i text; åtta förvalda diagnoser; en informationsasymmetri (betydligt fler blodprover — omkring 51 % av de tillgängliga analyterna mot 28 %, i billiga blodprover, inte bilddiagnostik); behandlingsutfall som delvis bedömdes mot den historiska journalen; samt ett offentligt benchmarkmaterial (MIMIC-IV) som en språkmodell kan ha sett under träningen — vilket författarna beskriver som en möjlig övre gräns för prestationen.

Hur stor tilltro bör en allmän läsare ha? Hög till att MIRA är en verklig och nyskapande demonstration av en förmåga — en agent som *agerar* genom journalen, inte bara en chattbot. Låg till att det är en *bättre diagnostiker än en läkare*: det påståendet är begränsat till ett retrospektivt benchmarktest och påverkas av provbeställningen, bedömningen mot journalen och möjlig kontaminering. Författarnas egen slutsats är den riktiga — detta behöver studeras prospektivt i verklig vård innan det betyder något för patientvården.

Källor

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