The Clean Paper

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

Een AI-agent werkte zelfstandig volledige patiëntcasussen af — in een simulator, op oude dossiers

MIRA is een nieuw soort medische AI: in plaats van één vraag te beantwoorden, werkt het een hele casus af in een gesimuleerd ziekenhuisdossier — anamnese, tests aanvragen en lezen, diagnose, opdrachten schrijven. Op 574 retrospectieve casussen uit een publieke database, verdeeld over acht vooraf geselecteerde diagnoses, rapporteerden de auteurs hogere diagnostische nauwkeurigheid dan bij artsen en grotendeels richtlijnconforme, medicatieveilige beslissingen. Maar elke kwalificatie telt: het draaide in een sandbox op oude dossiers, alleen in tekst; veel van het voordeel zat bij aandoeningen met duidelijke testuitslagen; het gebruikte ongeveer twee keer zoveel bloedonderzoek als de artsen; en verschillende uitkomsten werden tegen het oorspronkelijke dossier beoordeeld. De echte vooruitgang is een agent die door de hele workflow handelt — geen bewijs dat een machine nu beter diagnosticeert dan een arts.

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

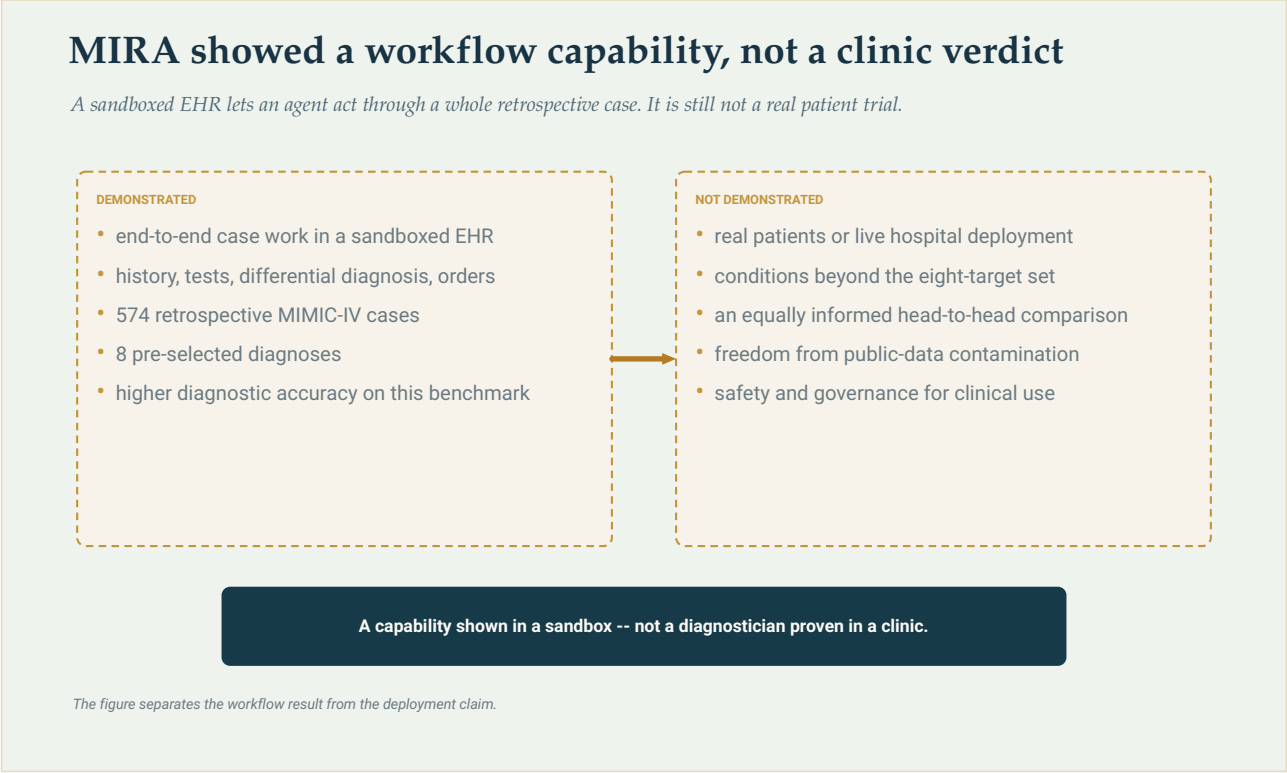
Een vraag beantwoorden is niet hetzelfde als een hele casus afhandelen

De meeste verhalen over medische AI gaan over een model dat *antwoordt*. Je geeft het een casus of examenvraag, het geeft een diagnose of alinea advies terug en scoort goed. Nuttig, maar beperkt: een slimme consultant die nooit zelf in het dossier werkt.

MIRA is gebouwd om iets anders te doen, en precies dat verschil is het nieuws. In plaats van één vraag te beantwoorden, werkt het een volledige casus af: het leest het patiëntendossier, bepaalt welke anamnese nog ontbreekt, vraagt laboratorium-, beeldvormings- en microbiologische onderzoeken aan, leest de resultaten, vernauwt de differentiaaldiagnose en schrijft vervolgens de opdrachten die daaruit volgen — voorschriften, opname, verwijzing voor chirurgie. Het doet dat door *in* een elektronisch patiëntendossier handelingen uit te voeren, zoals een clinicus, in plaats van vrije tekst te produceren die een mens nog moet overnemen.

Dat is een echte stap: van een systeem dat adviseert naar een systeem dat door de hele workflow heen handelt. Het is ook precies het soort stap dat in berichtgeving wordt platgeslagen tot “AI presteert beter dan artsen”. Daarom is het belangrijk precies te zeggen wat werd getest, en waar.

De versie in één zin: MIRA werkte zelfstandig hele casussen af en scoorde op deze benchmark hoger dan artsen op diagnostische nauwkeurigheid — maar het deed dat in een sandbox, op retrospectieve dossiers, bij acht vooraf gekozen diagnoses, uitsluitend via tekst, en een groot deel van het voordeel kwam bij aandoeningen met de duidelijkste testuitslagen. De vooruitgang is echt. Het scorebord is niet de kliniek.



MIRA demonstreerde een end-to-end workflowcapaciteit in een afgeschermd elektronisch patiëntendossier. Het demonstreerde geen klinische inzet in de echte wereld, brede diagnostische dekking of een eerlijke vergelijking met artsen die over evenveel informatie beschikten.

Wat de auteurs deden

MIRA — Medical Intelligence for Reasoning and Action — is een autonome agent gebouwd op modellen van OpenAI: het deel dat praat en handelt draait op **GPT-4o**, en een aparte planningsstap gebruikt OpenAI's redeneermodel **o1**. Die modellen zitten in een op maat gemaakt, standaardenconform raamwerk op basis van HL7 FHIR, de gegevensstandaard waarmee echte ziekenhuizen dossiers uitwisselen, met een gereedschapskist van meer dan tachtigduizend mogelijke klinische acties. Binnen die sandbox kan MIRA een voorgeschiedenis ophalen, laboratorium-, beeldvormings- en microbiologische onderzoeken bestellen en interpreteren, een differentiaaldiagnose opstellen en behandelplannen formuleren — medicijnen voorschrijven, operaties plannen, opnames organiseren. Eén detail verdient meteen aandacht: de geautomatiseerde "juryleden" die MIRA's antwoorden beoordeelden waren zelf GPT-4o — dezelfde modelfamilie die werd getest. Nadat een peer reviewer daar bezwaar tegen maakte, voegden de auteurs beoordeling door een gecertificeerd arts toe.

De testset kwam uit **MIMIC-IV**, een grote publiek beschikbare, gedeïdentificeerde EHR-database. Uit ongeveer 300.000 patiënten die tussen 2008 en 2019 in Beth Israel Deaconess Medical Center werden behandeld, selecteerden de auteurs **574 patiëntcasussen** verdeeld over **acht doeldiagnoses** — abdominale aandoeningen en internistische spoedgevallen, waaronder appendicitis, pancreatitis, pneumonie en urineweginfectie. Voor elke casus begon MIRA met een beperkt beeld en moest het stap voor stap beslissen wat er vervolgens nodig was — dezelfde vorm als een echte diagnostische work-up, maar uitgevoerd tegen een dossier waarvan de werkelijke afloop al bekend is.

De prestaties werden vervolgens met die van artsen op dezelfde casussen vergeleken.

Wat "autonome agent in een sandbox-EHR" werkelijk betekent

Drie woorden doen hier veel werk, dus het loont ze uit te pakken.
Agent betekent dat het systeem geen chatbot is die één prompt beantwoordt. Het doorloopt een lus: kijk naar de huidige toestand, kies een actie (vraag deze test aan, vraag naar dit stukje anamnese), bekijk het resultaat en kies de volgende actie — totdat er een diagnose en plan liggen. De "intelligentie" is het taalmodel; de *agency* zit in de omliggende software die tekst van het model omzet in toegestane EHR-acties en de resultaten weer terugvoert.
Sandboxed EHR betekent een gesimuleerde, afgeschermde kopie van een medisch dossiersysteem, geen live ziekenhuissysteem. MIRA kan een test "aanvragen" en de uitslag krijgen die die patiënt werkelijk had, omdat de casus historisch is en het antwoord al in het dossier staat. Niets wat MIRA doet raakt een echte patiënt of kliniek.
Autonoom betekent dat het de casus van begin tot eind zonder mens in de lus afwerkt — *binnen de sandbox*. Het betekent **niet** dat het zonder toezicht echte patiënten mocht behandelen. Dat

zijn twee totaal verschillende claims, en alleen de eerste werd getest.
Nog één belangrijk detail over de sandbox: MIRA mag *elke* test aanvragen, maar het systeem kan alleen een uitslag teruggeven als die test bij deze patiënt in werkelijkheid **ook echt is uitgevoerd**. Vraagt MIRA een bloedpanel dat de patiënt destijds kreeg, dan ontvangt het de echte waarden; vraagt het een scan die nooit is gemaakt, dan krijgt het "N/A — kon niet worden uitgevoerd", nooit een verzonnen getal. Met anamnese werkt het hetzelfde: als het dossier de gevraagde informatie niet bevat, zegt de gesimuleerde patiënt simpelweg dat hij het niet weet. MIRA kan dus niet magisch die ene beslissende test oproepen die nooit is gedaan — het kan alleen werken met wat de werkelijke work-up toevallig bevatte.
Dat onderscheid is belangrijk omdat het headlinewoord "autonoom" het makkelijkst als die tweede, veel sterkere claim wordt gelezen.

Wat ze vonden

Op de benchmark **presteerde MIRA beter dan artsen op diagnostische nauwkeurigheid** — maar achter die kop zitten drie verschillende getallen die makkelijk door elkaar raken.

Over alle **574 casussen** noemde MIRA zelfstandig in **88,9%** van de gevallen de juiste diagnose, beoordeeld tegen de ontslagdiagnose die in MIMIC-IV voor de patiënt was vastgelegd. Voor de directe vergelijking werkten de artsen niet alle 574 casussen af, maar een gedeelde subset van **311 casussen**, met dezelfde dossiers en hulpmiddelen als MIRA. Op diezelfde 311 casussen scoorde MIRA **87,8%** en werd het vergeleken met twee apart gerekruteerde groepen. De eerste was een **seniorgroep**: vier gecertificeerde artsen met 7 tot 11 jaar ervaring, die **78,1%** haalden. De tweede was een **gemengde ervaringsgroep**: vooral jonge arts-assistenten plus twee specialisten, dichter bij de personeelsmix van een echte spoedeisende hulp; die groep scoorde **71,1%**. Dezelfde casussen, dezelfde tools — AI eerst, ervaren artsen tweede, juniorzwaar team derde. De zorgvuldige formulering van het artikel zelf is dat MIRA over alle acht ziekten "consequent gelijkwaardig aan, en vaak beter dan" de artsen presteerde.

Ook de beslissingen verderop in de keten hielden grotendeels stand: meestal in overeenstemming met richtlijnen, medicatieveilig en passend wat opname betreft. De auteurs melden geen medicatiefouten van hoge ernst in vijf veiligheidsdomeinen — interacties, nierdosering, allergieën, QT-risico en opioïden — en correcte voorschriften in 467 van 468 gevallen, terwijl ze erbij zeggen dat het systeem "geen 100% betrouwbaarheid bereikte". Op het eerste gezicht is dat opvallend: een agent die de hele casus afwerkt en op diagnose voor artsen eindigt.

Maar de *vorm* van die overwinning is minstens zo belangrijk als de overwinning zelf. Onafhankelijke specialisten die het artikel bekeken, wezen erop waar het voordeel vooral vandaan kwam. Op het expertpanel van het Science Media Centre merkte dr. Wei Xing (University of Sheffield) op dat het opvallende resultaat — AI boven artsen op diagnostische nauwkeurigheid — **vooral werd gedragen door aandoeningen met duidelijke testuitslagen**, zoals appendicitis en pancreatitis, waar een beslissende scan of labwaarde de vraag vaak beslecht. Bij pneumonie en urineweginfecties — twee zeer gewone redenen om op een spoedeisende hulp te belanden — presteerden *zowel* AI als artsen het

slechtst en was het verschil het kleinst.

Er zat ook asymmetrie in hoeveel informatie beide partijen gebruikten. MIRA leunde sterker op het lab: het gebruikte ongeveer **51%** van de laboratoriumanalyten die in routinezorg beschikbaar waren, tegenover ongeveer **28%** bij de gecertificeerde artsen — mediaan zeven extra bloedparameters per casus. De auteurs benadrukken dat dit nog *onder* de routinezorgbaseline van de dataset lag en dus geen "bestel alles"-strategie was, en ze vonden *geen* systematische toename van duurdere doorsnedebeeldvorming. Toch kan meer informatie op zichzelf al hogere diagnostische nauwkeurigheid opleveren. Dit is dus niet helemaal een vergelijking tussen spelers met dezelfde informatie, een punt dat Xing expliciet maakte. Een arts die onbeperkt elk goedkoop bloedonderzoek mocht bestellen zonder kosten, ongemak of vertraging mee te wegen, zou op papier ook scherper worden.

Waarom "artsen verslaan" niet hetzelfde is als "betere arts"

Een benchmarkwinst is makkelijk te groot te maken. Tussen "MIRA scoorde hoger" en "MIRA is beter" zitten minstens drie dingen.

Ten eerste: **waarmee werd het vergeleken?** Professor Julie Jacko (University of Edinburgh) merkte op dat verschillende belangrijke uitkomsten van MIRA worden gedefinieerd ten opzichte van wat in de onderliggende dataset was gedocumenteerd. Het systeem wordt dus beloond voor **het reproduceren van vastgelegd klinisch gedrag**, niet noodzakelijk voor aantoonbaar optimale zorg. Het historische dossier wordt de antwoordensleutel. Dat is een redelijke manier om een benchmark te bouwen, maar het meet overeenstemming met wat er gebeurde, niet absolute juistheid. Eerlijkheidshalve raakt dit vooral de maten voor *treatment alignment* — kwamen MIRA's opdrachten overeen met het dossier? — terwijl de diagnostische nauwkeurigheid uit de kop tegen de ontslagdiagnose wordt gescoord, wat dichter bij een echte uitkomst ligt dan puur documentatie nadoen.

Ten tweede: de al genoemde **informatieasymmetrie**. Veel meer bloedonderzoek (ongeveer 51% van de beschikbare analyten tegenover 28%) betekent meer bewijs. Een deel van het nauwkeurigheidsverschil is waarschijnlijk *gekocht* met informatie, niet alleen beredeneerd.

Ten derde: **waar het systeem draaide**. Dit was een sandbox met retrospectieve casussen, acht vooraf gekozen diagnoses en alleen tekst. Echte klinische beoordeling hangt, zoals dr. Dominic Oliver (University of Oxford) het formuleerde, niet alleen af van *wat* patiënten zeggen maar ook van *hoe* ze het zeggen — naast lichamelijk onderzoek, geobserveerd gedrag en lichaamstaal, niets waarvan een tekstagent die een afgerond dossier leest iets ziet. En een patiënt met een probleem buiten de acht geselecteerde diagnoses valt simpelweg buiten wat deze studie kan beoordelen.

Daar komt een stillere zorg bij die reviewers noemden: **datacontaminatie**. MIMIC-IV is publiek en er is veel over geschreven, dus een taalmodel dat op het open internet is getraind kan artikelen, casusbesprekingen of mogelijk delen van de data al hebben gezien. Dr. Midhun Parakkal Unni (University of Sheffield) wees hier rechtstreeks op: als sommige antwoorden in de trainingsdata zaten, is een deel van de prestatie geheugen in plaats van klinisch redeneren, en alleen onafhankelijke replicatie kan die twee uit elkaar halen. Opvallend is dat de auteurs het punt niet wegwuiven: zij schrijven dat hun resultaten "voorzichtig als een mogelijke bovengrens" kunnen worden geïnterpreteerd en "gen-

eralisatie naar andere publieke casussen kunnen overschatten". Het artikel zet dus zelf een plafond op zijn eigen headline.

Dit alles maakt MIRA niet minder interessant. Het maakt de correcte lezing gekalibreerd: op een retrospectieve benchmark presteerde een agent die door het hele dossier kon handelen beter dan artsen bij diagnoses met duidelijke tests — deels door meer tests te gebruiken, deels door overeen te komen met het historische dossier. Dat is een echte capaciteitsdemonstratie, geen oordeel dat machines nu betere diagnostici zijn dan artsen.

Wat dit *niet* bewijst

- Het laat **niet** zien dat MIRA bij echte patiënten werkt. Elke casus was retrospectief en gesimuleerd; geen echte patiënt werd erdoor behandeld.
- Het laat **niet** zien dat het buiten acht diagnoses werkt. Aandoeningen buiten de vooraf gekozen set — de rommelige, ongedifferentieerde meerderheid van de geneeskunde — werden niet getest.
- Het laat **niet** zien dat inzet veilig is. De auteurs schrijven zelf dat generalisatie, veiligheid en governance nog prospectieve real-world studies nodig hebben.
- Het is **geen volledig eerlijke head-to-head met artsen**. MIRA gebruikte veel meer bloedtests (ongeveer 51% van de beschikbare analyten versus 28%), en verschillende behandeluitkomsten werden deels beoordeeld tegen het historische dossier in plaats van tegen objectief optimale zorg.
- Het laat **niet** zien dat memorisatie uitgesloten is. De publieke MIMIC-IV-data kunnen overlappen met trainingsmateriaal van het model; onafhankelijke replicatie is nodig om dat uit te sluiten.
- Het betekent **niet "AI vervangt artsen"**. Het is beslisondersteuning die *kan handelen* in een dossier. Volgens de reviewers ligt reëel gebruik in samenwerking met clinici, die bevoegdheid houden en alles leveren wat een tekstdossier mist.

Hoe sterk is het bewijs?

Splits de claim in twee delen, want het bewijs is heel verschillend.

Als capaciteitsdemonstratie — dat een taalmodelagent een volledige klinische casus in een EHR-sandbox kan uitvoeren, met anamnese, tests, diagnose en opdrachten in één keten — is het werk werkelijk nieuw en redelijk overtuigend. Dit is het nieuwe deel en het deel dat aandacht verdient.

Als superioriteitsclaim — dat MIRA *beter is dan artsen* — is het bewijs begrensd en moet het voorzichtig worden gelezen. Het geldt op één specifieke retrospectieve benchmark, bij acht diagnoses, met informatieasymmetrie (meer tests), een antwoordensleutel uit het historische dossier en een reële mogelijkheid van contaminatie van trainingsdata. Dat is genoeg om te zeggen "de agent presteerde indrukwekkend op deze benchmark". Het is niet genoeg om te zeggen "de agent is een betere diagnosticus dan een arts", en de auteurs claimen dat tweede ook niet.

De nuttigste houding is dus noch "AI verslaat artsen" noch "het is maar speelgoed". Het is: een nieuw

soort medische AI — een systeem dat *handelt* door het dossier in plaats van vragen te beantwoorden — deed het goed op een moeilijke retrospectieve benchmark en moet zich nu bewijzen waar het nog niet is getest: prospectief bij echte, ongedifferentieerde patiënten en onder governance.

Waarom het ertoe doet

De interessante vraag in medische AI is de afgelopen jaren stilletjes veranderd. Eerst was het "kan een model de diagnose goed krijgen?" — en modellen zeiden op steeds zorgvuldigere testsets steeds vaker ja. MIRA markeert de verschuiving naar een moeilijkere vraag: "kan een model *het werk doen* — verzamelen, aanvragen, interpreteren, beslissen, handelen — over een hele workflow?" Dat is een nuttigere en eerlijkere vraag, omdat juist in handelen zowel de moeilijkheid als het risico zit.

Daarom doet de framing ertoe. De automatische kop — *AI presteert beter dan artsen* — wijst naar het minst nieuwe en minst goed ondersteunde deel van het resultaat. Het werkelijk nieuwe is kleiner en belangrijker: een agent die door het hele dossier kan bewegen. Als die capaciteit standhoudt, verandert ze eerst de **workflow**, lang voordat ze verandert wie daar de leiding over heeft. De arts verdwijnt niet; aanvragen, interpreteren, uitslagen najagen — dáár komt zo'n hulpmiddel eerst terecht.

Het zet ook de bewijslast de juiste kant op. Een leaderboardwinst op retrospectieve casussen is een reden om een prospectieve trial te doen, geen vervanging ervan. Dat zeggen de auteurs zelf. De eerlijke lezing van MIRA is dus een uitnodiging voor de volgende studie, gemeten volgens de norm die het veld al kent: op patiënten, niet op benchmarks.

Heldere samenvatting

MIRA is een autonome AI-agent die in een gesimuleerd elektronisch patiëntendossier werkt: hij kan anamnese afnemen, lab-, beeldvormings- en microbiologische onderzoeken aanvragen en interpreteren, tot een diagnose komen en behandelopdrachten schrijven. Op 574 retrospectieve casussen uit de publieke MIMIC-IV-database, verdeeld over acht vooraf geselecteerde diagnoses, presteerde het beter dan artsen op diagnostische nauwkeurigheid (88,9% totaal; 87,8% tegenover 78,1% in de directe vergelijking) en nam het grotendeels richtlijnconforme, medicatieveilige beslissingen. Maar de evaluatie was een simulatie op oude dossiers, alleen via tekst; een groot deel van het voordeel kwam bij aandoeningen met duidelijke testuitslagen; MIRA gebruikte veel meer bloedtests dan de artsen (ongeveer 51% van beschikbare analyten versus 28%); verschillende behandeluitkomsten werden beoordeeld tegen wat het oorspronkelijke dossier vermeldde; en de publieke dataset geeft een reëel risico op contaminatie van trainingsdata — iets wat de auteurs zelf een mogelijke bovengrens op hun cijfers noemen. De echte vooruitgang is een agent die door een volledige workflow handelt in plaats van losse vragen te beantwoorden. De auteurs zeggen expliciet dat generalisatie, veiligheid en governance nog prospectief in de echte wereld moeten worden onderzocht. Dat is de juiste lezing: een indrukwekkende capaciteitsdemonstratie, geen bewijs dat AI beter diagnosticeert dan artsen en geen systeem dat klaar is voor een echte kliniek.

Zonder omwegen

Wat het artikel laat zien: Een taalmodelagent (MIRA) kan een volledige klinische casus van begin tot eind uitvoeren in een sandbox-EHR — anamnese, tests, diagnose, opdrachten — en presteerde op een retrospectieve benchmark van 574 casussen bij acht diagnoses beter dan artsen op diagnostische nauwkeurigheid (88,9% totaal; 87,8% versus 78,1% tegen gecertificeerde artsen), zonder medicatiefouten van hoge ernst.

Wat aannemelijk is maar niet bewezen: Dat deze capaciteit voordeel oplevert bij echte, ongedifferentieerde patiënten; dat het nauwkeurigheidsvoordeel beter redeneren weerspiegelt en niet vooral meer aangevraagde tests, overeenstemming met het dossier of gememoriseerde publieke data.

Wat het niet laat zien: Dat MIRA buiten acht diagnoses werkt; dat het veilig kan worden ingezet; dat het artsen verslaat in een eerlijke vergelijking met gelijke informatie; dat het clinici vervangt; dat de resultaten vrij zijn van trainingsdatacontaminatie.

Belangrijkste beperkingen: Retrospectieve, gesimuleerde, tekst-only evaluatie; acht vooraf geselecteerde diagnoses; informatieasymmetrie (veel meer bloedtests — ongeveer 51% van beschikbare analyten versus 28%, vooral goedkope bloedtests en niet beeldvorming); behandeluitkomsten deels beoordeeld tegen het historische dossier; en een publieke benchmark (MIMIC-IV) die het taalmodel mogelijk tijdens training heeft gezien — door de auteurs zelf genoemd als mogelijke bovengrens voor de prestaties.

Hoeveel vertrouwen mag een algemene lezer hierin hebben? Hoog vertrouwen dat MIRA een echte en nieuwe capaciteitsdemonstratie is — een agent die *handelt* door het dossier, niet slechts een chatbot. Laag vertrouwen dat het een *betere diagnosticus dan een arts* is: die claim is beperkt tot één retrospectieve benchmark en wordt verstoord door testgebruik, scoring tegen het dossier en mogelijke datacontaminatie. De bottom line van de auteurs zelf is de juiste: dit heeft prospectief real-world onderzoek nodig voordat het iets over patiëntenzorg betekent.

Bronnen

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