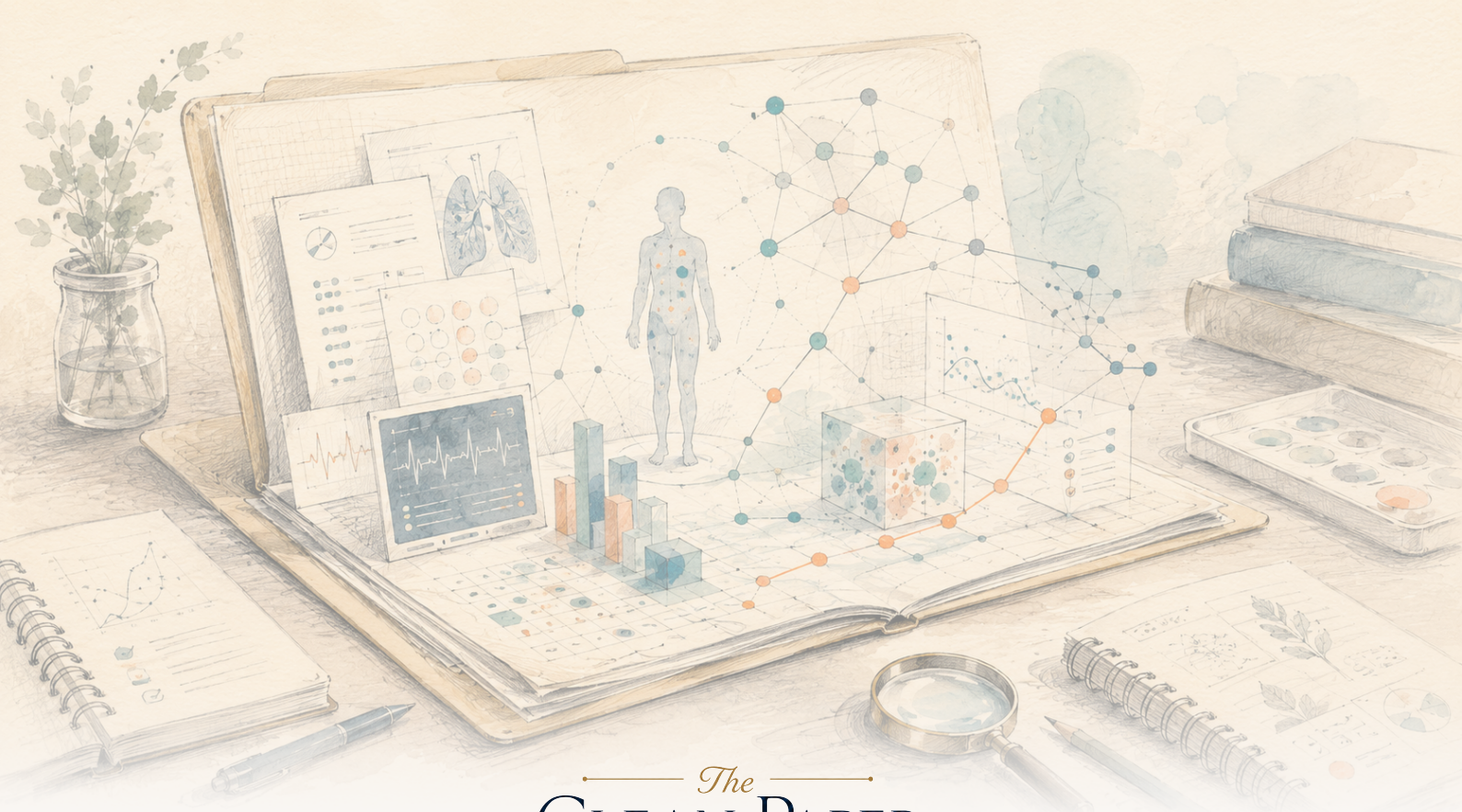


The Clean Paper

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

'n KI-agent het hele pasiëntgevalle op sy eie hanteer — in 'n simulator, op vorige rekords

MIRA is 'n nuwe soort mediese KI: in plaas daarvan om een vraag te beantwoord, werk dit 'n hele geval binne 'n gesimuleerde hospitaalrekord af — geskiedenis neem, toetse bestel en lees, 'n diagnose bereik, opdragte skryf. Op 574 retrospektiewe gevalle uit 'n openbare databasis, oor agt voorafgekose diagnoses, rapporteer die outeurs dat dit dokters in diagnostiese akkuraatheid oortref en grotendeels riglynkonforme, medikasieveilige besluite geneem het. Elke kwalifiseerder in daardie sin dra gewig: dit het in 'n sandbox op vorige rekords geloop, slegs in teks; 'n groot deel van sy voorsprong het gekom by toestande met duidelike toetsresultate; dit het omtrent twee keer soveel bloedtoetse as die dokters gebruik; en verskeie uitkomste is beoordeel teen wat die oorspronklike rekord aangeteken het. Die werklike vooruitgang is 'n agent wat oor die hele werkvloei optree — nie bewys dat 'n masjien nou beter as 'n dokter diagnoseer nie. Die outeurs sê dit self: veralgemening, veiligheid en bestuur verg nog prospektiewe studies in die werklike wêreld.

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

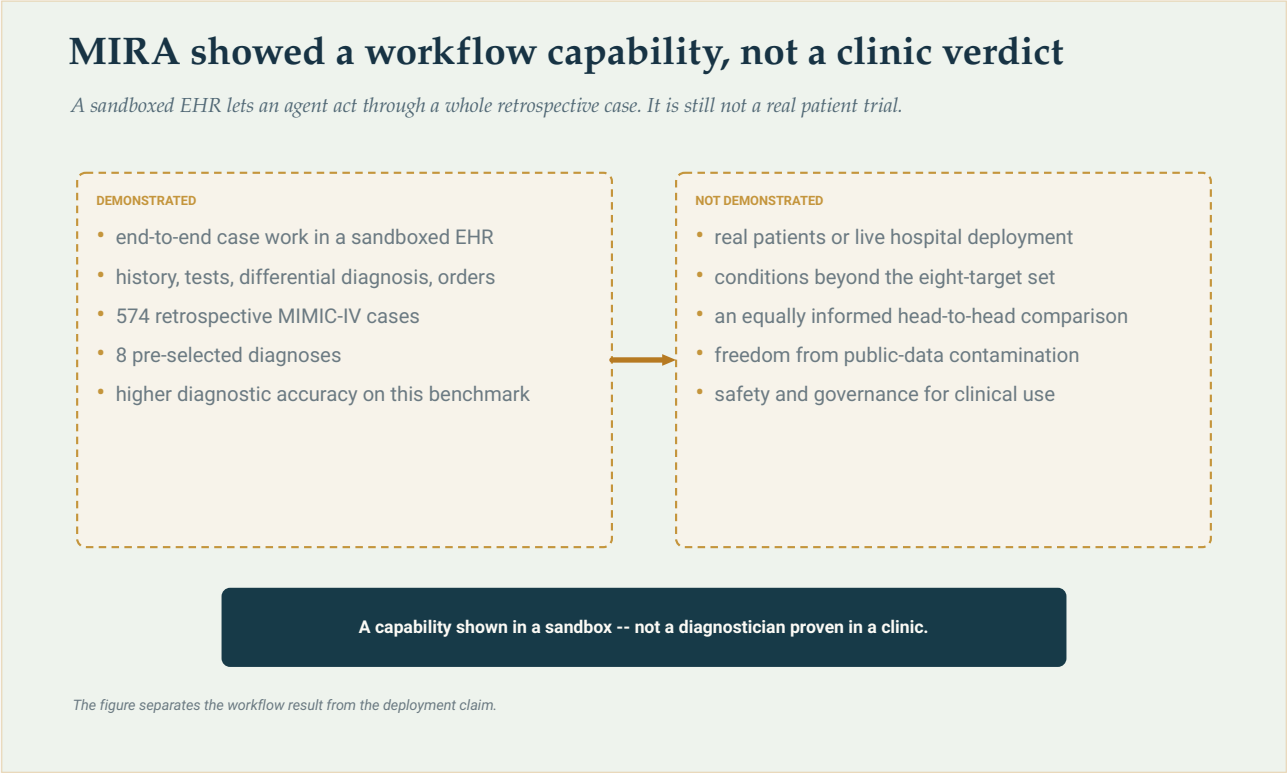
Om ’n vraag te beantwoord is nie dieselfde as om die hele geval te bestuur nie

Die meeste verhale oor mediese KI gaan oor ’n model wat *antwoord*. Jy gee dit ’n gevalbeskrywing of eksamenvraag, dit gee ’n diagnose of ’n paragraaf raad terug, en dit behaal ’n goeie telling. Nuttig, maar beperk: ’n slim konsultant wat nooit aan die rekord raak nie.

MIRA is gebou om iets anders te doen, en daardie verskil is die werklike nuus. In plaas daarvan om een vraag te beantwoord, werk dit deur ’n hele geval: dit lees ’n pasiënt se rekord, besluit watter geskiedenis nog nodig is, bestel laboratorium-, beeld- en mikrobiologietoetse, lees die resultate, vernou ’n differensiële diagnose en skryf dan die opdragte wat volg — voorskrifte, opname, ’n verwysing vir chirurgie. Dit doen dit deur *binne* ’n elektroniese gesondheidsrekord op te tree soos ’n klinikus, eerder as om vrye teks te produseer wat ’n mens moet oorskryf.

Dit is ’n werklike stap: van ’n stelsel wat raad gee na ’n stelsel wat oor die hele werkvloei optree. Dit is ook presies die soort stap wat in beriggewing platgeslaan word tot “KI presteer beter as dokters”. Daarom is dit die moeite werd om presies te wees oor wat getoets is, en waar.

Die een-sin-weergawe: MIRA het hele gevalle op sy eie hanteer en, op hierdie maatstaf, dokters in diagnostiese akkuraatheid oortref — maar dit het dit in ’n afgeskermde toetsomgewing gedoen, op retrospektiewe rekords, oor agt voorafgekose diagnoses, uitsluitlik deur teks, en ’n groot deel van sy voorsprong gekry by toestande met die duidelikste toetsresultate. Die vooruitgang is werklik. Die telbord is nie die kliniek nie.



MIRA het binne ’n afgeskermde EHR-toetsomgewing gewys dat dit ’n hele kliniese werkvloei van begin tot einde kan hanteer. Dit het nie werklike kliniese ontplooiing, breë diagnostiese dekking of ’n billike, ewe goed ingeligte

kop-aan-kop-vergelyking met dokters gedemonstreer nie.

Original Aurora diagram — The Clean Paper CC BY 4.0

Wat die outeurs gedoen het

MIRA — Medical Intelligence for Reasoning and Action — is 'n outonome agent wat op OpenAI se modelle gebou is: die deel wat gesels en optree gebruik **GPT-4o**, en 'n aparte beplanningstap gebruik OpenAI se **o1**-redeneringsmodel. Dit sit binne 'n pasgemaakte, standaardversoenbare raamwerk — gebou op HL7 FHIR, die datastandaard wat werklike hospitale gebruik om rekords rond te skuif — met 'n gereedskapkis van meer as tagtigduisend moontlike kliniese aksies. Binne dié afgeskermde toetsomgewing kan MIRA 'n pasiënt se geskiedenis oproep, laboratorium-, beeld- en mikrobiologietoetse bestel en interpreteer, 'n differensiële diagnose opstel en behandelingsplanne formuleer — medisyne voorskryf, chirurgiese prosedures skeduleer en opnames beplan. Een besonderheid is die moeite werd om vooraf uit te lig: die geoutomatiseerde "beoordelaars" wat MIRA se antwoorde beoordeel het, was self GPT-4o — dieselfde modelfamilie wat getoets is — en die outeurs het dit met 'n oorsig deur 'n gesertifiseerde dokter aangevul nadat 'n ewekniebeoordelaar die probleem uitgewys het.

Die toetsstel het uit **MIMIC-IV** gekom, 'n groot, openbare, gedeïdentifiseerde EHR-databasis. Uit ongeveer 300 000 pasiënte wat tussen 2008 en 2019 by Beth Israel Deaconess Medical Center behandel is, het die outeurs **574 pasiëntgevalle** gekies wat oor **agt teikendiagnoses** gestrek het — abdominale patologieë en interne-geneeskunde-noodgevalle, onder meer blindedermontsteking, pankreatitis, longontsteking en urienweginfeksie. Vir elke geval het MIRA met 'n beperkte prentjie begin en stap vir stap moes besluit wat om volgende te doen — dieselfde vorm as 'n werklike ondersoek, maar uitgevoer teen 'n rekord waarvan die werklike afloop reeds op lêer was.

Sy prestasie is daarna met dié van dokters op dieselfde gevalle vergelyk.

Wat "outonome agent in 'n afgeskermde EHR-toetsomgewing" eintlik beteken

Drie woorde doen hier baie werk, dus is dit die moeite werd om hulle uit te pak.

Agent beteken die stelsel is nie 'n kletsbot wat bloot 'n opdrag beantwoord nie. Dit loop in 'n siklus: kyk na die huidige toestand, kies 'n aksie (bestel hierdie toets, vra vir daardie geskiedenis), sien die resultaat, kies die volgende aksie — totdat dit by 'n diagnose en plan uitkom. Die "intelligensie" is 'n taalmodel; die *agentskap* is die raamwerk rondom dit wat die model se teks in toegelate EHR-operasies omskep en die resultate weer terugvoer.

Afgeskermde EHR-toetsomgewing (*sandbox*) beteken 'n gesimuleerde kopie van 'n mediese rekordstelsel, nie 'n lewendige hospitaalstelsel nie. MIRA kan 'n toets "bestel" en die uitslag ontvang wat daardie pasiënt werklik gekry het, omdat die geval histories is en die antwoord reeds aangeteken is. Niks wat dit doen raak 'n werklike pasiënt of kliniek nie.

Outonoom beteken dat dit die geval van begin tot einde sonder 'n mens in die lus voltooi — *binne die toetsomgewing*. Dit beteken **nie** dat dit losgelaat is om pasiënte sonder toesig te bestuur

> nie. Dit is baie verskillende aansprake, en net die eerste is getoets.
> Nog een ding oor die toetsomgewing, omdat dit maklik verkeerd voorgestel word: MIRA mag *enige* toets bestel wat dit wil, maar die stelsel kan net 'n uitslag teruggee as daardie toets in die werklike lewe **daadwerklik op hierdie pasiënt gedoen is**. Vra vir 'n bloedpaneel wat die pasiënt ontvang het, en jy kry die werklike waardes; vra vir 'n skandering wat niemand bestel het nie, en jy kry "N/A — kon nie uitgevoer word nie", nooit 'n uitgedinkte getal nie. Geskiedenis werk dieselfde: as die rekord nie bevat wat MIRA vra nie, sê die gesimuleerde pasiënt eenvoudig dat dit nie weet nie. MIRA kan dus nie die een beslissende toets oproep wat nooit gedoen is nie — dit kan net werk met wat die werklike ondersoek toevallig ingesluit het. Die onderskeid maak saak omdat die opskrifwoord — "outonoom" — die een is wat die maklikste as die tweede ding gelees word.

Wat hulle gevind het

Op die maatstaf het **MIRA dokters in diagnostiese akkuraatheid oortref** — maar die opskrif versteek drie verskillende getalle wat maklik deurmekaar raak, dus is dit die moeite werd om hulle apart te hou.

Op sy eie, oor al **574 gevalle**, het MIRA **88,9%** van die tyd die regte diagnose genoem — beoordeel teen die ontslagdiagnose wat werklik vir elke pasiënt in MIMIC-IV aangeteken is. Vir die kop-aan-kop-vergelyking met mense het die dokters nie al 574 gevalle gedoen nie; hulle het 'n gedeelde **311-geval-substel** hanteer, met dieselfde rekords en gereedskap wat MIRA gehad het. Op daardie selfde 311 gevalle het MIRA **87,8%** behaal en is dit teen twee afsonderlik gewerfde groepe gemeet. Die eerste was 'n **senior groep**: vier gesertifiseerde dokters met 7 tot 11 jaar ervaring, wat **78,1%** behaal het. Die tweede was 'n **gemengde senioriteitsgroep**: meestal junior residensiedokters plus twee spesialiste, nader aan hoe 'n werklike noodafdeling beman word, wat **71,1%** behaal het. Dieselfde gevalle, dieselfde gereedskap — en die volgorde was KI eerste, ervare dokters tweede, die junior-swaar span derde. Die artikel se eie versigtige formulering is dat MIRA oor al agt siektes "deurgaans gelykwaardig aan, en dikwels beter as" die dokters was.

Sy daaropvolgende besluite het ook goed gevaar: grotendeels in lyn met riglyne, veilig wat medikasie betref en gepas oor opname. Die outeurs rapporteer geen hoë-erns-medikasiefoute oor vyf veiligheidskategorieë nie (geneesmiddelinteraksies, nierdosering, allergieë, QT-risiko en opioïedvoorskryf) en korrekte voorskrifte in 467 van 468 gevalle — terwyl hulle opmerk dat die stelsel "nie 100% betroubaarheid bereik het nie". Op sigwaarde is dit 'n opvallende resultaat: 'n agent wat die hele geval bestuur en op die diagnose voor dokters uitkom.

Maar die *vorm* van die oorwinning maak net soveel saak as die oorwinning self. Onafhanklike spesialiste wat die artikel gelees het, het uitgewys waar die voorsprong eintlik vandaan gekom het. Op die Science Media Centre se deskundigepaneel het dr. Wei Xing (University of Sheffield) opgemerk dat die opskrifgetal — KI wat dokters in diagnostiese akkuraatheid klop — **hoofsaaklik deur toestande met duidelike toetsresultate aangedryf is**, soos blindedermontsteking en pankreatitis, waar 'n beslissende skandering of laboratoriumwaarde die vraag afhandel. Vir longontsteking en urien-

weginfeksies — twee van die algemeenste redes waarom mense werklik in 'n noodafdeling beland — het *sowel* die KI as die dokters die swakste gevaar, en was die gaping tussen hulle die kleinste.

Daar was ook 'n asimmetrie in hoe die twee kante gespeel het, en dit staan in die artikel se eie getalle. MIRA het sterker op die laboratorium geleun: dit het ongeveer **51%** van die laboratoriumanaliete gebruik wat in roetinesorg beskikbaar was, teenoor sowat **28%** vir die gesertifiseerde dokters — 'n mediaan van sewe meer bloedparameters per geval. Die outeurs is versigtig om dit te raam as *onder* die databasis se eie roetinesorg-basislyn, nie 'n bestel-alles-strategie nie, en rapporteer *geen* sistematiese toename in duurder deursnitbeelding nie. Tog kan meer inligting op sigself hoër diagnostiese akkuraatheid oplewer — dus is dit nie heeltemal 'n gelyke wedstryd tussen ewe goed ingeligte spelers nie, 'n gaping wat Xing direk uitgewys het. 'n Klinikus wat elke goedkoop bloedtoets kon bestel sonder om koste, ongemak of vertraging te weeg, sou op papier ook skerper lyk.

Waarom "dokters geklop" nie "beter dokter" beteken nie

'n Maatstaf-oorwinning is maklik om te oorlees. Drie dinge lê tussen "MIRA het hoër behaal" en "MIRA is beter".

Eerstens, **waartegen dit beoordeel is.** Professor Julie Jacko (University of Edinburgh) het opgemerk dat verskeie van MIRA se sleuteluitkomste gedefinieer word relatief tot wat in die onderliggende databasis aangeteken is — wat beteken dat die stelsel beloon word vir **die reproduksie van die aangetekende kliniese gedrag**, nie noodwendig vir die demonstrasie van optimale sorg nie. Die historiese rekord word die antwoordblad. Dit is 'n redelike manier om 'n maatstaf te bou, maar dit meet ooreenstemming met wat gedoen is, nie korrektheid in 'n absolute sin nie. Om billik teenoor die artikel te wees: dit byt die hardste by die *behandelingsbelyning*-maatstawwe — stem MIRA se opdragte met die rekord ooreen? — terwyl die opskrif-diagnostiese akkuraatheid teen die ontslagdiagnose beoordeel word, wat nader aan 'n werklike uitkoms is as bloot 'n dokumentasie-eggo.

Tweedens, **die reeds genoemde inligtingsasimmetrie**: baie meer bloedtoetse (ongeveer 51% van beskikbare analiete teenoor 28%) beteken meer bewys. 'n Deel van die akkuraatheidsgaping word waarskynlik *gekoop*, nie beredeneer nie.

Derdens, **waar dit geloop het.** Dit was 'n afgeskermde toetsomgewing met retrospektiewe gevalle, oor agt voorafgekose diagnoses, slegs in teks. Werklike kliniese beoordeling, soos dr. Dominic Oliver (University of Oxford) dit gestel het, hang nie net af van wat pasiënte sê nie, maar ook van hoe hulle dit sê — saam met fisiese ondersoek, waargenome gedrag en liggaamstaal, waarvan 'n teksagent wat 'n voltooide rekord lees niks sien nie. En 'n pasiënt wie se probleem nie binne een van die agt gekose diagnoses val nie, is eenvoudig buite wat hierdie studie kan aanspreek.

Daar is ook 'n stiller bekommernis wat beoordelaars geopper het: **datakontaminasie.** MIMIC-IV is openbaar en daar is baie daaroor geskryf, dus kon 'n taalmodel wat op die oop internet opgelei is reeds artikels, gevalbesprekings of die data self gesien het. Dr. Midhun Parakkal Unni (University of Sheffield) het dit direk uitgewys — as sommige antwoorde in die opleidingsdata was, is 'n deel van die prestasie geheue, nie kliniese redenering nie, en net onafhanklike replikasie kan die twee van mekaar onderskei. Opmerklik genoeg wuif die outeurs dit nie weg nie: hulle skryf dat hul resultate

"versigtig as 'n moontlike boonste grens geïnterpreteer kan word" en "veralgemening na ander openbare gevalle kan oorskat" — 'n artikel wat self 'n plafon op sy opskrif plaas.

Niks hiervan maak MIRA minder interessant nie. Dit maak die korrekte lees 'n gekalibreerde een: op 'n retrospektiewe maatstaf het 'n agent wat oor die hele rekord kon optree dokters oortref op diagnoses met duidelike toetsuitslae — deels deur meer toetse te bestel, deels deur met die aangetekende rekord ooreen te stem. Dit is 'n werklike demonstrasie van vermoë, nie 'n uitspraak dat masjiene nou beter as dokters diagnoseer nie.

Wat dit *nie* bewys nie

- Dit wys **nie dat MIRA op werklike pasiënte werk nie**. Elke geval was retrospektief en gesimuleer; geen pasiënt is ooit deur MIRA bestuur nie.
- Dit wys **nie dat dit verder as agt diagnoses werk nie**. Toestande buite die voorafgekose stel — die rommelige, ongedifferensieerde meerderheid van medisyne — is nie getoets nie.
- Dit wys **nie dat dit veilig is om te ontplooi nie**. Die outeurs skryf self dat veralgemening, veiligheid en bestuur nog prospektiewe studies in die werklike wêreld verg.
- Dit **vestig nie 'n billike kop-aan-kop-vergelyking met dokters nie**. Dit het op baie meer bloedtoetse gesteun (ongeveer 51% van beskikbare analiete teenoor 28%), en verskeie behandelingsuitkomste is deels teen die aangetekende rekord beoordeel eerder as teen 'n onafhanklike grondwaarheid van beste sorg.
- Dit wys **nie dat die resultaat vry van memorisering is nie**. Die openbare MIMIC-IV-data kan met die model se opleidingsdata oorvleuel, iets wat onafhanklike replikasie sal moet uitsluit.
- Dit beteken **nie "KI vervang dokters" nie**. Dit is besluitsteun wat oor 'n rekord kan *optree*; die beoordelaars se konsensus is dat werklike gebruik in vennootskap met klinici sal wees, wat gesag behou en alles verskaf wat 'n teksrekord weglaat.

Hoe sterk is die bewys?

Verdeel die aanspraak in twee, want die bewys verskil baie tussen die twee helftes.

As 'n demonstrasie van vermoë — dat 'n taalmodel-agent 'n hele kliniese geval binne 'n afgeskermde EHR-toetsomgewing kan bestuur en geskiedenis, toetse, diagnose en opdragte van begin tot einde kan ketting — is die werk werklik nuut en redelik oortuigend. Dit is die nuwe deel, en die deel waarop dit die moeite werd is om te let.

As 'n aanspraak op meerderwaardigheid — dat MIRA *beter as dokters* is — is die bewys begrens en moet dit versigtig gelees word. Dit geld op 'n spesifieke retrospektiewe maatstaf, oor agt diagnoses, met 'n inligtingsasimmetrie (meer toetse), 'n antwoordblad wat uit die historiese rekord kom, en 'n werklike moontlikheid van kontaminasie deur opleidingsdata. Dit is genoeg om te sê "die agent het indrukwekkend op hierdie maatstaf gevaar". Dit is nie genoeg om te sê "die agent is 'n beter diagnostikus as 'n dokter" nie, en die outeurs maak nie die tweede aanspraak nie.

Die nuttigste houding is nóg "KI klop dokters" nóg "net 'n speelding". Dit is: 'n nuwe soort mediese KI — een wat oor die rekord *optree* in plaas daarvan om vrae te beantwoord — het goed gevaar op 'n moeilike retrospektiewe maatstaf, en moet hom nou bewys waar dit nog nie getoets is nie: prospektief, op werklike, ongedifferensieerde pasiënte, onder behoorlike bestuur.

Waarom dit saak maak

Vir 'n paar jaar het die interessante vraag in mediese KI stilweg verander. Dit was vroeër "kan 'n model die diagnose reg kry?" — en modelle het op skoner en skoner toetsstelle telkens ja geantwoord. MIRA merk die verskuiwing na 'n moeiliker vraag: "kan 'n model *die werk doen* — versamel, bestel, interpreteer, besluit, optree — oor 'n hele werkvloei?" Dit is 'n nuttiger en eerliker vraag, want optrede is waar sowel die moeilikheid as die risiko lê.

Daarom maak die raamwerk saak. Die opskrifrefleks — *KI presteer beter as dokters* — wys na die minst nuwe en swakste ondersteunde deel van die resultaat. Die werklik nuwe ding is kleiner en meer betekenisvol: 'n agent wat deur die hele rekord kan beweeg. As dié vermoë standhou, verander dit die **werkvloei** lank voordat dit verander wie in beheer daarvan is. Die dokter word nie vervang nie; die bestel, interpreteer, opvolg van resultate — die werkvloei — is waar 'n instrument soos hierdie eerste land.

En dit stel die bewyslas in die regte rigting terug. 'n Ranglys-oorwinning op retrospektiewe gevalle is 'n rede om die prospektiewe proef te doen, nie 'n plaasvervanger daarvoor nie. Die outeurs sê dieselfde. Die eerlike lees van MIRA is 'n uitnodiging tot daardie volgende studie — volgens die standaard wat die veld reeds ken: gemeet op pasiënte, nie op maatstawwe nie.

Kortliks

MIRA is 'n outonome KI-agent wat binne 'n afgeskermde elektroniese gesondheidsrekord-toetsomgewing werk: dit kan 'n geskiedenis neem, laboratorium-, beeld- en mikrobiologietoetse bestel en interpreteer, 'n diagnose bereik en behandelingsplanne skryf. Op 574 retrospektiewe gevalle uit die openbare MIMIC-IV-databasis, oor agt voorafgekose diagnoses, het dit dokters in diagnostiese akkuraatheid oortref (88,9% algeheel; 87,8% teenoor 78,1% kop aan kop) en grotendeels riglynkonforme, medikasieveilige besluite geneem. Maar die evaluering was 'n simulasie op vorige rekords, slegs in teks; 'n groot deel van die voorsprong het gekom by toestande met duidelike toetsresultate; MIRA het baie meer bloedtoetse as die dokters gebruik (ongeveer 51% van beskikbare analiete teenoor 28%); verskeie behandelingsuitkomste is beoordeel teen wat die oorspronklike rekord aangeteken het; en die openbare databasis skep 'n werklike risiko van opleidingsdatakontaminasie — wat die outeurs self 'n moontlike boonste grens op hul getalle noem. Die werklike vooruitgang is 'n agent wat oor die hele werkvloei optree in plaas daarvan om los vrae te beantwoord. Die outeurs is uitdruklik dat veralgemening, veiligheid en bestuur nog prospektiewe studies in die werklike wêreld verg — die korrekte lees: 'n indrukwekkende demonstrasie van vermoë, nie bewys dat KI beter as dokters

diagnoseer nie, en nie ’n stelsel wat gereed is vir ’n werklike kliniek nie.

Nugtere beoordeling

Wat die artikel wys: ’n Taalmodel-agent (MIRA) kan ’n hele kliniese geval van begin tot einde binne ’n afgeskermde EHR-toetsomgewing bestuur — geskiedenis, toetse, diagnose, opdragte — en het op ’n retrospektiewe maatstaf van 574 gevalle oor agt diagnoses dokters in diagnostiese akkuraatheid oortref (88,9% algeheel; 87,8% teenoor 78,1% teen gesertifiseerde dokters), sonder hoë-erns-medikasiefoute.

Wat aannemelik is maar nie bewys nie: Dat dié vermoë in voordeel vir werklike, ongedifferensieerde pasiënte vertaal; dat die akkuraatheidsvoorsprong beter redenering weerspieël eerder as meer bestelde toetse, ooreenstemming met die aangetekende rekord of gememoriseerde openbare data.

Wat dit nie wys nie: Dat MIRA buite sy agt diagnoses werk; dat dit veilig is om te ontplooi; dat dit dokters in ’n billike, ewe goed ingeligte vergelyking klop; dat dit klinici vervang; dat die resultate vry is van opleidingsdatakontaminasie.

Belangrikste beperkings: Retrospektiewe, gesimuleerde, teks-alleen-evaluering; agt voorafgekose diagnoses; ’n inligtingsasimmetrie (veel meer bloedtoetse — ongeveer 51% van beskikbare analiete teenoor 28%, in laekoste-bloedtoetse, nie beeldvorming nie); behandelingsuitkomste deels teen die historiese rekord beoordeel; en ’n openbare maatstaf (MIMIC-IV) wat ’n taalmodel moontlik tydens opleiding gesien het — iets wat die outeurs as ’n moontlike boonste grens op prestasie uitwys.

Hoeveel vertroue behoort ’n algemene leser hierin te hê? Hoë vertroue dat MIRA ’n werklike en nuwe demonstrasie van vermoë is — ’n agent wat oor die rekord *optree*, nie net ’n kletsbot nie. Lae vertroue dat dit ’n *beter diagnostikus as ’n dokter* is: daardie aanspraak is beperk tot een retrospektiewe maatstaf en verstrengel met toetsbestelling, beoordeling teen die rekord en moontlike kontaminasie. Die outeurs se eie slotsom is die regte een — dit verg prospektiewe studie in die werklike wêreld voordat dit enigiets vir pasiëntsorg beteken.

Bronne

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
```