



The CLEAN PAPER

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

Ein KI-Agent bearbeitete ganze Patientenfälle selbstständig — in einem Simulator, auf historischen Akten

MIRA ist eine neue Art medizinischer KI: Statt eine einzelne Frage zu beantworten, bearbeitet der Agent einen ganzen Fall innerhalb einer simulierten elektronischen Patientenakte — Anamnese, Anordnung und Interpretation von Tests, Diagnose und schließlich klinische Anordnungen. Auf 574 retrospektiven Fällen aus einer öffentlichen Datenbank mit acht vorab ausgewählten Diagnosen berichten die Autoren eine höhere diagnostische Genauigkeit als bei Ärzten und weitgehend leitlinienkonforme, medikamentös sichere Entscheidungen. Jeder einschränkende Zusatz in diesem Satz zählt: Das System lief in einer Sandbox auf historischen Akten und nur per Text; ein großer Teil des Vorsprungs entstand bei Erkrankungen mit eindeutigen Testergebnissen; es nutzte etwa doppelt so viele Blutparameter wie die Ärzte; und mehrere Ergebnisse wurden daran gemessen, was in der ursprünglichen Akte dokumentiert war. Der wirkliche Fortschritt ist ein Agent, der über den gesamten Workflow hinweg handelt — kein Beweis, dass eine Maschine nun besser diagnostiziert als ein Arzt. Die Autoren sagen das selbst: Generalisierbarkeit, Sicherheit und Governance müssen noch prospektiv in der realen Versorgung untersucht werden.

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

Eine Frage zu beantworten ist nicht dasselbe wie einen ganzen Fall zu führen

Die meisten Geschichten über medizinische KI handeln von einem Modell, das *antwortet*. Man gibt ihm eine Fallvignette oder Prüfungsfrage, es liefert eine Diagnose oder einen Absatz mit Empfehlungen und erzielt eine gute Punktzahl. Nützlich, aber eng begrenzt: ein kluger Konsiliararzt, der die Patientenakte nie selbst anfasst.

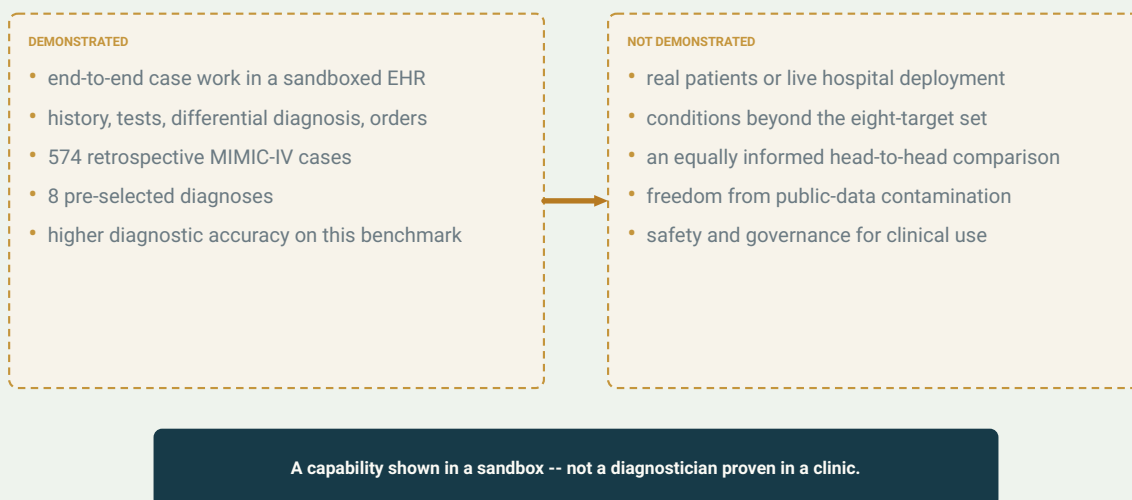
MIRA wurde für etwas anderes gebaut, und genau dieser Unterschied ist die eigentliche Nachricht. Statt eine einzelne Frage zu beantworten, bearbeitet das System einen ganzen Fall: Es liest die Patientenakte, entscheidet, welche Anamneseinformationen noch fehlen, ordnet Labor-, Bildgebungs- und mikrobiologische Untersuchungen an, liest die Ergebnisse, grenzt die Differenzialdiagnose ein und schreibt anschließend die daraus folgenden Anordnungen — Rezepte, eine stationäre Aufnahme, eine Überweisung zur Operation. Es handelt dabei *innerhalb* einer elektronischen Patientenakte, wie ein Kliniker, statt nur freien Text zu erzeugen, den ein Mensch anschließend übertragen muss.

Das ist ein echter Schritt: von einem System, das berät, zu einem System, das über den gesamten Arbeitsablauf hinweg handelt. Und es ist genau die Art von Schritt, die in der Berichterstattung leicht zu „KI schlägt Ärzte“ plattgedrückt wird. Deshalb lohnt es sich, genau hinzusehen, was getestet wurde — und wo.

Die Ein-Satz-Version: MIRA bearbeitete vollständige Fälle selbstständig und war auf diesem Benchmark bei der diagnostischen Genauigkeit besser als Ärzte — allerdings in einer Sandbox, auf retrospektiven Akten, über acht vorab ausgewählte Diagnosen hinweg, ausschließlich per Text; einen großen Teil seines Vorsprungs erzielte es zudem bei Erkrankungen mit besonders eindeutigen Testergebnissen. Der Fortschritt ist real. Die Rangliste ist nicht die Klinik.

MIRA showed a workflow capability, not a clinic verdict

A sandboxed EHR lets an agent act through a whole retrospective case. It is still not a real patient trial.



The figure separates the workflow result from the deployment claim.

MIRA demonstrierte einen durchgängigen klinischen Arbeitsablauf innerhalb einer abgeschotteten EHR-Sandbox. Nicht demonstriert wurden ein Einsatz in der realen Versorgung, breite diagnostische Abdeckung oder ein fairer Direktvergleich mit Ärzten bei identischem Informationsstand.

Original Aurora diagram — The Clean Paper CC BY 4.0

Was die Autoren gemacht haben

MIRA — Medical Intelligence for Reasoning and Action — ist ein autonomer Agent auf Basis von OpenAI-Modellen: Der Teil, der kommuniziert und handelt, läuft mit **GPT-4o**, ein separater Planungsschritt verwendet das Reasoning-Modell **o1** von OpenAI. Diese Modelle sitzen in einem eigens entwickelten, standardkonformen Framework auf Basis von HL7 FHIR — dem Datenstandard, mit dem reale Krankenhäuser Patientendaten austauschen — und verfügen dort über einen Werkzeugkasten mit mehr als achtzigtausend möglichen klinischen Aktionen. Innerhalb dieser Sandbox kann MIRA die Vorgeschichte eines Patienten abrufen, Labor, Bildgebung und Mikrobiologie anordnen und interpretieren, eine Differenzialdiagnose erstellen und Behandlungspläne formulieren — Medikamente verschreiben, chirurgische Eingriffe planen oder stationäre Aufnahmen vorbereiten. Ein Detail sollte von Anfang an genannt werden: Die automatischen „Gutachter“, die MIRAs Antworten bewerteten, waren selbst GPT-4o — also dieselbe Modellfamilie, die getestet wurde. Nachdem ein Peer Reviewer dies beanstandet hatte, ergänzten die Autoren die Bewertung durch einen Facharzt.

Der Testsatz stammte aus **MIMIC-IV**, einer großen, öffentlich verfügbaren und de-identifizierten Datenbank elektronischer Gesundheitsakten. Aus rund 300.000 Patienten, die zwischen 2008 und 2019 am Beth Israel Deaconess Medical Center behandelt wurden, wählten die Autoren **574**

Patientenfälle mit **acht Zieldiagnosen** aus — abdominale Erkrankungen und internistische Notfälle, darunter Appendizitis, Pankreatitis, Pneumonie und Harnwegsinfektion. Bei jedem Fall startete MIRA mit einem begrenzten Informationsstand und musste Schritt für Schritt entscheiden, was als Nächstes zu tun war — strukturell ähnlich einer realen Abklärung, nur anhand einer historischen Akte, deren tatsächliches Ende bereits dokumentiert ist.

Anschließend wurde seine Leistung auf denselben Fällen mit Ärzten verglichen.

Was „autonomer Agent in einer EHR-Sandbox“ tatsächlich bedeutet

Drei Wörter leisten hier sehr viel Arbeit, deshalb lohnt es sich, sie auseinanderzunehmen.

Agent bedeutet, dass das System kein Chatbot ist, der nur auf einen Prompt antwortet. Es arbeitet in einer Schleife: aktuellen Zustand ansehen, eine Aktion auswählen (diesen Test anordnen, nach jener Information fragen), Ergebnis sehen, nächste Aktion wählen — bis Diagnose und Plan stehen. Die „Intelligenz“ ist ein Sprachmodell; die *Handlungsfähigkeit* entsteht durch die Software darum herum, die Modelltext in erlaubte EHR-Aktionen übersetzt und die Ergebnisse wieder zurückspielt.

EHR-Sandbox bedeutet eine simulierte, abgeschottete Kopie eines Systems für elektronische Patientenakten, nicht ein aktives Krankenhaus-System. MIRA kann einen Test „anordnen“ und das Ergebnis erhalten, das dieser Patient tatsächlich bekam, weil der Fall historisch ist und die Antwort bereits in der Akte steht. Keine seiner Aktionen betrifft einen echten Patienten oder eine echte Klinik.

Autonom bedeutet, dass MIRA den Fall von Anfang bis Ende ohne Menschen in der Schleife bearbeitet — *innerhalb der Sandbox*. Es bedeutet **nicht**, dass das System unbeaufsichtigt reale Patienten behandeln durfte. Das sind sehr unterschiedliche Behauptungen, und nur die erste wurde getestet.

Noch ein Punkt zur Sandbox, weil man sie sich leicht falsch vorstellt: MIRA darf zwar jeden beliebigen Test *anfordern*, aber das System kann nur dann ein Resultat zurückgeben, wenn dieser Test beim Patienten in der Realität **tatsächlich durchgeführt wurde**. Verlangt MIRA ein Blutpanel, das der Patient bekommen hat, erhält es die echten Werte; verlangt es einen Scan, den damals niemand anordnete, lautet die Antwort „N/A — konnte nicht durchgeführt werden“, niemals eine erfundene Zahl. Bei der Anamnese gilt dasselbe: Enthält die Akte die von MIRA erfragte Information nicht, sagt der simulierte Patient schlicht, dass er es nicht weiß. MIRA kann also nicht den einen entscheidenden Test herbeizaubern, der damals nie gemacht wurde — es kann nur mit dem arbeiten, was die reale Abklärung tatsächlich umfasste.

Diese Unterscheidung ist wichtig, weil gerade das Schlagwort „autonom“ am ehesten so gelesen wird, als sei die zweite Variante getestet worden.

Was sie herausfanden

Auf dem Benchmark **übertraf MIRA die Ärzte bei der diagnostischen Genauigkeit** — aber hinter dieser Schlagzeile stecken drei verschiedene Zahlen, die leicht durcheinandergeraten. Deshalb sollte man sie trennen.

Über alle **574 Fälle** hinweg nannte MIRA in **88,9 %** der Fälle die richtige Diagnose — bewertet anhand der Entlassungsdiagnose, die für den jeweiligen Patienten in MIMIC-IV dokumentiert war. Für den direkten Vergleich mit Menschen bearbeiteten die Ärzte nicht alle 574 Fälle, sondern eine gemeinsame **Teilmenge von 311 Fällen**, mit denselben Akten und Werkzeugen, die auch MIRA zur Verfügung standen. Auf genau diesen 311 Fällen erreichte MIRA **87,8 %** und wurde mit zwei getrennt rekrutierten Ärztegruppen verglichen. Die erste war eine **erfahrene Gruppe**: vier Fachärzte mit 7 bis 11 Jahren Berufserfahrung, die **78,1 %** erreichten. Die zweite war eine **Gruppe mit gemischter Erfahrung**: überwiegend junge Assistenzärzte plus zwei Spezialisten — näher daran, wie eine reale Notaufnahme tatsächlich besetzt ist — mit **71,1 %**. Gleiche Fälle, gleiche Werkzeuge: KI zuerst, erfahrene Ärzte dahinter, die stärker von Junioren geprägte Gruppe an dritter Stelle. Die vorsichtige Formulierung der Arbeit selbst lautet, MIRA sei über alle acht Erkrankungen hinweg „durchgehend gleichwertig und häufig überlegen“ gewesen.

Auch die nachfolgenden Entscheidungen hielten sich gut: weitgehend leitlinienkonform, medikamentös sicher und bei Aufnahmen angemessen. Die Autoren berichten über fünf Sicherheitskategorien hinweg — Arzneimittelinteraktionen, Nieren-Dosierung, Allergien, QT-Risiko und Opioidverordnung — keine Medikationsfehler hohen Schweregrades und korrekte Verordnungen in 467 von 468 Fällen; zugleich weisen sie darauf hin, dass das System „keine 100-prozentige Zuverlässigkeit“ erreichte. Für sich genommen ist das ein bemerkenswertes Ergebnis: ein Agent, der den ganzen Fall bearbeitet und bei der Diagnose vor den Ärzten landet.

Aber die *Form* dieses Vorsprungs ist ebenso wichtig wie der Vorsprung selbst. Unabhängige Fachleute, die die Arbeit beurteilten, wiesen darauf hin, woher der Unterschied tatsächlich kam. Auf dem Expertengremium des Science Media Centre merkte Dr. Wei Xing (University of Sheffield) an, dass die Schlagzeilenzahl — KI schlägt Ärzte bei der Diagnostik — **vor allem von Erkrankungen mit eindeutigen Testergebnissen getragen wurde**, etwa Appendizitis und Pankreatitis, bei denen ein entscheidender Scan oder Laborwert die Frage klärt. Bei Pneumonie und Harnwegsinfektionen — zwei der häufigsten Gründe, aus denen Menschen tatsächlich eine Notaufnahme aufsuchen — schnitten *sowohl* KI als auch Ärzte am schlechtesten ab, und ihr Abstand war am kleinsten.

Außerdem gab es eine Asymmetrie darin, wie beide Seiten „spielten“, und sie steckt in den eigenen Zahlen der Studie. MIRA stützte sich stärker auf das Labor: Es nutzte etwa **51 %** der im Routinebetrieb verfügbaren Laborparameter, gegenüber ungefähr **28 %** bei den Fachärzten — im Median sieben zusätzliche Blutparameter pro Fall. Die Autoren betonen sorgfältig, dass dies *unter* dem Routineversorgungsniveau des Datensatzes lag und keine Strategie des „alles bestellen“ darstellte; zudem berichten sie *keinen* systematischen Anstieg teurerer Schnittbildgebung. Trotzdem kann mehr Information schon für sich genommen zu höherer diagnostischer Genauigkeit führen. Es ist deshalb kein ganz gleichartiger Wettkampf zwischen gleich informierten Teilnehmern — genau diese Lücke hob Xing hervor. Ein Arzt, der jeden billigen Bluttest bestellen dürfte, ohne Kosten, Unbehagen oder Verzögerung abzuwägen, würde auf dem Papier ebenfalls treffsicherer aussehen.

Warum „Ärzte geschlagen“ nicht „besserer Arzt“ bedeutet

Ein Sieg auf einem Benchmark lässt sich leicht überinterpretieren. Zwischen „MIRA erzielte mehr Punkte“ und „MIRA ist besser“ liegen drei Dinge.

Erstens: **Wogegen wurde bewertet?** Professor Julie Jacko (University of Edinburgh) wies darauf hin, dass mehrere zentrale Ergebnisse MIRAs relativ zu dem definiert sind, was im zugrunde liegenden Datensatz dokumentiert wurde. Das System wird also dafür belohnt, **das aufgezeichnete klinische Verhalten zu reproduzieren**, nicht zwingend dafür, eine optimale Behandlung zu demonstrieren. Die historische Akte wird zum Lösungsschlüssel. Das ist eine vernünftige Art, einen Benchmark aufzubauen, misst aber Übereinstimmung mit dem tatsächlich Vorgefallenen und nicht Richtigkeit in einem absoluten Sinn. Fairerweise betrifft dies besonders die Kennzahlen zur *Übereinstimmung der Behandlung* — passten MIRAs Anordnungen zur Akte? —, während die diagnostische Schlagzeilenzahl gegen die Entlassungsdiagnose bewertet wird, die einem realen Ergebnis näherkommt als bloße Dokumentationsspiegelung.

Zweitens: die bereits erwähnte **Informationsasymmetrie**. Deutlich mehr Bluttests — etwa 51 % der verfügbaren Laborparameter gegenüber 28 % — bedeuten mehr Evidenz. Ein Teil des Genauigkeitsvorsprungs wurde wahrscheinlich *durch zusätzliche Information erkaufte*, nicht allein durch besseres Schlussfolgern.

Drittens: **wo das System lief**. Es war eine Sandbox mit retrospektiven Fällen, acht vorab ausgewählten Diagnosen und ausschließlich Text. Eine reale klinische Beurteilung hängt, wie Dr. Dominic Oliver (University of Oxford) formulierte, nicht nur davon ab, *was* Patienten sagen, sondern auch *wie* sie es sagen — zusammen mit körperlicher Untersuchung, beobachtetem Verhalten und Körpersprache. Nichts davon sieht ein textbasierter Agent, der eine abgeschlossene Akte liest. Und ein Patient, dessen Problem nicht zu einer der acht ausgewählten Diagnosen gehört, liegt schlicht außerhalb dessen, worüber diese Studie etwas aussagen kann.

Hinzu kommt eine leisere Sorge der Gutachter: **Datenkontamination**. MIMIC-IV ist öffentlich und wurde intensiv wissenschaftlich diskutiert. Ein Sprachmodell, das auf dem offenen Internet trainiert wurde, könnte daher bereits Arbeiten, Falldiskussionen oder sogar Teile der Daten selbst gesehen haben. Dr. Midhun Parakkal Unni (University of Sheffield) wies direkt darauf hin: Wenn einige Antworten bereits in den Trainingsdaten lagen, stammt ein Teil der Leistung aus Erinnerung statt klinischem Schlussfolgern — und nur unabhängige Replikation kann beides auseinanderhalten. Bemerkenswert ist, dass die Autoren diesen Punkt nicht wegwischen: Sie schreiben, ihre Ergebnisse könnten vorsichtig als „mögliche Obergrenze“ interpretiert werden und könnten die Generalisierbarkeit auf andere öffentliche Fälle überschätzen — eine Arbeit, die der eigenen Schlagzeile selbst eine Obergrenze setzt.

Nichts davon macht MIRA weniger interessant. Es führt nur zu einer kalibrierten Lesart: Auf einem retrospektiven Benchmark übertraf ein Agent, der in der gesamten Akte handeln konnte, Ärzte insbesondere bei Diagnosen mit klaren Tests — teilweise indem er mehr Tests nutzte, teilweise indem er mit der aufgezeichneten Akte übereinstimmte. Das ist eine echte Demonstration einer neuen Fähigkeit, kein Urteil darüber, dass Maschinen nun grundsätzlich besser diagnostizieren als Ärzte.

Was dies *nicht* beweist

- Es zeigt **nicht**, dass MIRA bei echten Patienten funktioniert. Jeder Fall war retrospektiv und simuliert; kein Patient wurde je von MIRA behandelt.
- Es zeigt **nicht**, dass MIRA jenseits von acht Diagnosen funktioniert. Erkrankungen außerhalb dieser vorab ausgewählten Gruppe – also der unübersichtliche, undifferenzierte Großteil der Medizin – wurden nicht getestet.
- Es zeigt **nicht**, dass das System sicher eingesetzt werden kann. Die Autoren selbst schreiben, dass Generalisierbarkeit, Sicherheit und Governance weiterhin prospektive Studien in der realen Versorgung erfordern.
- Es belegt **keinen** vollständig fairen Direktvergleich mit Ärzten. MIRA griff auf deutlich mehr Bluttests zurück (etwa 51 % der verfügbaren Laborparameter gegenüber 28 %), und mehrere Behandlungsergebnisse wurden teilweise anhand der historischen Akte statt anhand einer unabhängigen optimalen Versorgung bewertet.
- Es zeigt **nicht**, dass das Ergebnis frei von Memorisation ist. Die öffentlichen MIMIC-IV-Daten könnten sich mit den Trainingsdaten des Modells überschneiden; ausschließen ließe sich das erst durch unabhängige Replikation.
- Es bedeutet **nicht** „KI ersetzt Ärzte“. Es handelt sich um Entscheidungsunterstützung, die *innerhalb* einer Akte handeln kann; nach Einschätzung der Gutachter wird ein realer Einsatz in Partnerschaft mit Klinikern erfolgen, die die Entscheidungshoheit behalten und all das beitragen, was eine Textakte nicht enthält.

Wie belastbar sind die Belege?

Man sollte die Behauptung in zwei Teile zerlegen, weil die Evidenz für beide sehr unterschiedlich ist.

Als Fähigkeitsdemonstration – dass ein Sprachmodell-Agent einen gesamten klinischen Fall in einer EHR-Sandbox bearbeiten kann, indem er Anamnese, Tests, Diagnose und Anordnungen durchgängig verkettet – ist die Arbeit tatsächlich neuartig und einigermaßen überzeugend. Genau dieser Teil ist neu und verdient Aufmerksamkeit.

Als Überlegenheitsbehauptung – dass MIRA *besser als Ärzte* sei – ist die Evidenz begrenzt und muss vorsichtig gelesen werden. Sie gilt auf einem bestimmten retrospektiven Benchmark mit acht Diagnosen, bei einer Informationsasymmetrie durch mehr Tests, einem teilweise aus der historischen Akte gewonnenen Lösungsschlüssel und einer realen Möglichkeit von Trainingsdatenkontamination. Das reicht für die Aussage: „Der Agent schnitt auf diesem Benchmark beeindruckend ab.“ Es reicht nicht für: „Der Agent ist ein besserer Diagnostiker als ein Arzt.“ Und die Autoren behaupten Letzteres auch nicht.

Die nützlichste Haltung ist weder „KI schlägt Ärzte“ noch „nur ein Spielzeug“. Sondern: Eine neue Art medizinischer KI – eine, die in der Akte *handelt*, statt nur Fragen zu beantworten – schnitt auf einem anspruchsvollen retrospektiven Benchmark gut ab und muss sich nun dort beweisen, wo sie noch nicht getestet wurde: prospektiv, an realen, nicht vorselektierten Patienten und unter klarer

Governance.

Warum das wichtig ist

Seit einigen Jahren hat sich die interessante Frage bei medizinischer KI still verändert. Früher lautete sie: „Kann ein Modell die richtige Diagnose nennen?“ — und Modelle beantworteten das auf immer saubereren Testsätzen zunehmend mit Ja. MIRA markiert den Übergang zu einer schwierigeren Frage: „Kann ein Modell *die Arbeit erledigen* — Informationen sammeln, Untersuchungen anordnen, interpretieren, entscheiden und handeln — über einen ganzen Arbeitsablauf hinweg?“ Das ist die nützlichere und ehrlichere Frage, weil beim Handeln sowohl die eigentliche Schwierigkeit als auch das Risiko liegen.

Genau deshalb ist das Framing wichtig. Der reflexartige Titel *KI schlägt Ärzte* zeigt auf den am wenigsten neuen und am schwächsten gestützten Teil des Ergebnisses. Wirklich neu ist etwas Kleineres, aber Konsequenteres: ein Agent, der sich durch die gesamte Akte bewegen kann. Wenn diese Fähigkeit standhält, verändert sie zunächst den **Arbeitsablauf**, lange bevor sie verändert, wer dafür verantwortlich ist. Der Arzt wird nicht ersetzt; Anordnen, Interpretieren, Ergebnissen hinterherlaufen — der Workflow — ist der Ort, an dem ein Werkzeug wie dieses zuerst ankommt.

Und die Arbeit verschiebt die Beweislast in die richtige Richtung. Ein Sieg auf retrospektiven Fällen ist ein Grund, eine prospektive Studie durchzuführen, kein Ersatz dafür. Genau das sagen auch die Autoren. Die ehrliche Lesart von MIRA ist eine Einladung zu dieser nächsten Studie — gemessen an dem Standard, den das Feld bereits kennt: an Patienten, nicht an Benchmarks.

Kurz zusammengefasst

MIRA ist ein autonomer KI-Agent, der in einer abgeschotteten elektronischen Patientenakte arbeitet: Er kann eine Anamnese erheben, Labor-, Bildgebungs- und mikrobiologische Untersuchungen anordnen und interpretieren, eine Diagnose stellen und Behandlungspläne schreiben. Auf 574 retrospektiven Fällen aus der öffentlichen MIMIC-IV-Datenbank mit acht vorab ausgewählten Diagnosen übertraf MIRA Ärzte bei der diagnostischen Genauigkeit (88,9 % insgesamt; im Direktvergleich 87,8 % gegenüber 78,1 %) und traf weitgehend leitlinienkonforme und medikamentös sichere Entscheidungen. Doch die Evaluation war eine Simulation auf historischen Akten und ausschließlich textbasiert; ein großer Teil des Vorsprungs kam bei Erkrankungen mit eindeutigen Testergebnissen zustande; MIRA nutzte deutlich mehr Bluttests als die Ärzte (etwa 51 % der verfügbaren Laborparameter gegenüber 28 %); mehrere Behandlungsergebnisse wurden anhand dessen bewertet, was in der ursprünglichen Akte dokumentiert war; und der öffentliche Datensatz birgt ein reales Risiko von Trainingsdatenkontamination — die Autoren selbst bezeichnen ihre Zahlen daher als mögliche Obergrenze. Der echte Fortschritt ist ein Agent, der über den gesamten Workflow hinweg handelt, statt isolierte Fragen zu beantworten. Die Autoren sagen ausdrücklich, dass Generalisierbarkeit, Sicherheit und Governance noch prospektive Realweltstudien brauchen. Das ist die richtige Lesart: eine beeindruckende Fähigkeitsdemonstration, kein Beweis dafür, dass

KI besser diagnostiziert als Ärzte, und kein System, das für eine reale Klinik bereit ist.

Nüchtern geprüft

Was die Arbeit zeigt: Ein Sprachmodell-Agent (MIRA) kann einen vollständigen klinischen Fall von Anfang bis Ende in einer EHR-Sandbox bearbeiten — Anamnese, Tests, Diagnose, Anordnungen — und übertraf auf einem retrospektiven Benchmark mit 574 Fällen und acht Diagnosen Ärzte bei der diagnostischen Genauigkeit (88,9 % insgesamt; 87,8 % gegenüber 78,1 % im Vergleich mit Fachärzten), ohne Medikationsfehler hohen Schweregrades.

Was plausibel, aber nicht bewiesen ist: Dass diese Fähigkeit realen, nicht vorselektierten Patienten zugutekommt; dass der Genauigkeitsvorsprung besseres klinisches Schlussfolgern widerspiegelt und nicht mehr angeordnete Tests, Übereinstimmung mit der historischen Akte oder memorisierte öffentliche Daten.

Was die Arbeit nicht zeigt: Dass MIRA außerhalb seiner acht Diagnosen funktioniert; dass es sicher eingesetzt werden kann; dass es Ärzte in einem fairen Vergleich bei identischem Informationsstand schlägt; dass es Kliniker ersetzt; oder dass die Ergebnisse frei von Trainingsdatenkontamination sind.

Wichtigste Einschränkungen: Retrospektive, simulierte, rein textbasierte Evaluation; acht vorab ausgewählte Diagnosen; eine Informationsasymmetrie durch deutlich mehr Bluttests — etwa 51 % der verfügbaren Laborparameter gegenüber 28 %, bei günstigen Bluttests und nicht Bildgebung; Behandlungsergebnisse teilweise anhand der historischen Akte bewertet; sowie ein öffentlicher Benchmark (MIMIC-IV), den ein Sprachmodell im Training gesehen haben könnte — was die Autoren selbst als mögliche Obergrenze der Leistung kennzeichnen.

Wie viel Vertrauen sollte ein allgemeiner Leser haben? Hohes Vertrauen darin, dass MIRA eine echte und neuartige Fähigkeitsdemonstration ist — ein Agent, der *in* einer Patientenakte handelt, nicht bloß ein Chatbot. Geringes Vertrauen in die viel stärkere Behauptung, es sei ein *besserer Diagnostiker als ein Arzt*: Diese Aussage ist auf einen retrospektiven Benchmark begrenzt und wird durch unterschiedliche Testnutzung, Bewertung anhand der Akte und mögliche Datenkontamination verzerrt. Die Schlussfolgerung der Autoren selbst ist die richtige: Bevor dies für die Patientenversorgung etwas bedeutet, braucht es prospektive Studien in der realen Welt.

Quellen

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-](https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-simulated-EH.aspx)
aspx