The Clean Paper

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

एक AI एजेंट ने पूरा मरीज़-मामला खुद संभाला— लेकिन सिमुलेटर में, पुराने रिकॉर्ड पर

MIRA एक अलग तरह का चिकित्सकीय AI है: एक सवाल का जवाब देने के बजाय यह नकली अस्पताल रिकॉर्ड के भीतर पूरा मामला चलता है — इतिहास लेता है, जाँचें मंगाता और पढ़ता है, निदान तक पहुँचता है और आदेश लिखता है। सार्वजनिक डेटासेट के 574 पुराने मामलों और पहले से चुने गए आठ निदानों में लेखकों के अनुसार इसका निदान-सटीकता चिकित्सकों से अधिक रही और इसके अधिकांश निर्णय दिशानिर्देशों के अनुरूप तथा दवा-सुरक्षा की दृष्टि से उचित थे। लेकिन हर सीमा महत्वपूर्ण है: यह केवल पठ-आधारित सैंडबॉक्स में पुराने रिकॉर्डों पर चली; इसकी बढ़त खासकर उन रोगों में थी जिनके जाँच-नतीजे साफ थे; इसने डॉक्टरों से लगभग दोगुनी रकम-जाँचें मंगाईं; और कई परिणामों का मूल चार्ट में दर्ज उत्तर के आधार पर अंक दिए गए। असली प्रगति यह है कि एक एजेंट पूरे कार्यप्रवाह में कार्रवाई कर सकता है — यह प्रमाण नहीं कि मशीन अब डॉक्टर से बेहतर निदान करती है। लेखकों के अनुसार समन्वयीकरण, सुरक्षा और शासन के लिए अभी वास्तविक दुनिया में भावी अध्ययन जरूरी हैं।

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

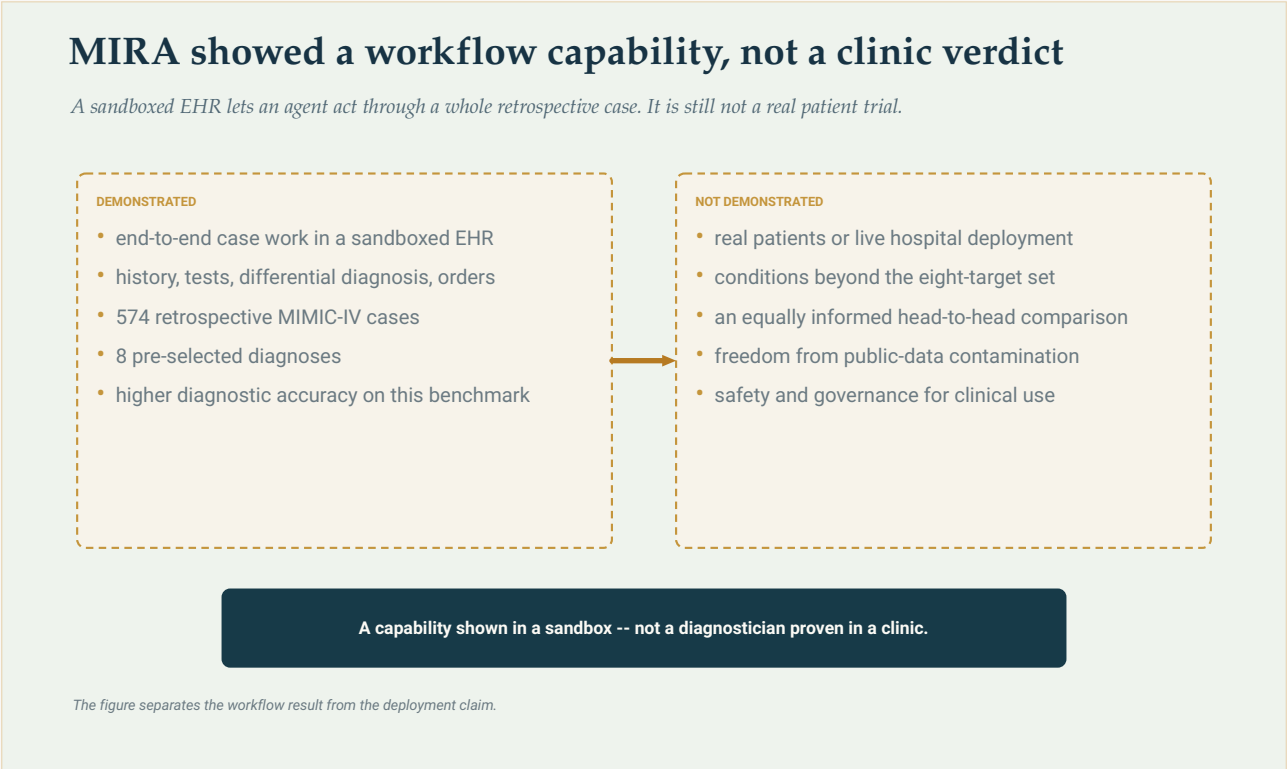
किसी सवाल का जवाब देना और पूरा मामला संभालना एक ही बात नहीं है

चिकित्सकीय AI की अधिकांश कहानियाँ ऐसे मॉडल के बारे में होती हैं जो *जवाब देता है*। आप उसे किसी मरीज़ का संक्षिप्त विवरण या परीक्षा का सवाल देते हैं, वह निदान या सलाह का एक पैराग्राफ़ लौटाता है और अच्छा स्कोर कर लेता है। यह उपयोगी है, लेकिन दायरा सीमित है: जैसे कोई तेज़ सलाहकार जो मरीज़ की पूरी फ़ाइल को कभी हाथ न लगाए।

MIRA को कुछ अलग करने के लिए बनाया गया है, और असली खबर यही अंतर है। एक सवाल का जवाब देने के बजाय यह पूरे मामले पर काम करता है: मरीज़ का रिकॉर्ड पढ़ता है, तय करता है कि और कौन-सी जानकारी चाहिए, प्रयोगशाला, इमेजिंग और सूक्ष्मजीव-विज्ञान जाँचें मंगाता है, परिणाम पढ़ता है, संभावित निदानों की सूची को संकरा करता है और फिर आगे के आदेश लिखता है — दवाएँ, भर्ती या सर्जरी के लिए रेफ़रल। यह किसी इंसान को कॉपी करने के लिए केवल मुक्त-पाठ सलाह नहीं बनता; यह इलेक्ट्रॉनिक स्वास्थ्य रिकॉर्ड के अंदर क्रियाएँ करता है, जैसे कोई चिकित्सक करता है।

यह सचमुच एक कदम आगे है: सलाह देने वाली प्रणाली से पूरे कार्यप्रवाह में काम करने वाली प्रणाली तक। और यही वह तरह की प्रगति है जिसे खबरों में अक्सर समेटकर “AI डॉक्टरों से बेहतर” बना दिया जाता है। इसलिए यह साफ़ रखना ज़रूरी है कि वास्तव में क्या परखा गया और किस परिस्थिति में।

एक वाक्य में: MIRA ने पूरे मामले अपने आप चलाए और इस बेंचमार्क पर निदान की सटीकता में चिकित्सकों से बेहतर स्कोर किया — लेकिन यह सैंडबॉक्स में, पुराने रिकॉर्ड पर, पहले से चुने गए आठ निदानों के भीतर और केवल पाठ के माध्यम से हुआ; उसकी बढ़त का बड़ा हिस्सा उन स्थितियों में आया जहाँ जाँच-परिणाम बहुत स्पष्ट थे। प्रगति वास्तविक है। लेकिन स्कोरबोर्ड क्लिनिक नहीं होता।



MIRA ने सैंडबॉक्स वाले EHR के भीतर शुरू से अंत तक पूरा कार्यप्रवाह चलाने की क्षमता दिखाई। उसने वास्तविक क्लिनिकल तैनाती, व्यापक निदान या समान जानकारी के साथ डॉक्टरों के विरुद्ध निष्पक्ष आमने-सामने तुलना नहीं दिखाई।

Original Aurora diagram — The Clean Paper CC BY 4.0

लेखकोंने क्या किया

MIRA — Medical Intelligence for Reasoning and Action — OpenAI के मॉडलोंपर बनाएक स्वायत्त एजेंट है। बातचीत और क्रियाओं वालाहिस्सा**GPT-4o** पर चलताहै, जबकि अलग योजना-चरण OpenAI के **o1** रीज़निंग मॉडल काउपयोग करता है। दोनोंएक विशेष रूप से बनाए गए, मानकोंके अनुरूप ढाँचे के भीतर काम करते हैं। यह HL7 FHIR पर आधारित है — वही डेटामानक जिसकाउपयोग वास्तविक अस्पताल स्वास्थ्य रिकॉर्ड के आदान-प्रदान के लिए करते हैं — और इसमें 80,000 से अधिक संभावित क्लिनिकल क्रियाओं काउपकरण-संग्रह है। इस सैंडबॉक्स में MIRA मरीज़ काइतिहास देख सकताहै, प्रयोगशाला इमेजिंग और सूक्ष्मजीव-विज्ञान जाँचें मंगाऔर समझ सकताहै, विभेदक निदान बनासकताहै और उपचार-योजनातैयार कर सकताहै — दवाएँ लिखने से लेकर सर्जरीतय करने और भर्तीकीयोजनाबनाने तक। एक विवरण शुरू में ही ध्यान देने योग्य है: MIRA के उत्तरोंकोअंक देने वाले स्वचालित “निर्णायक” स्वयं GPT-4o थे — यानीउसीमॉडल परविश्वास से जिसे परखाजारहाथा। एक सहकर्मीसमीक्षक ने इस पर आपत्ति उठाई, जिसके बाद लेखकोंने प्रमाणित चिकित्सक की समीक्षाभीजोड़ी।

परीक्षण-सामग्रीMIMIC-IV से आई, जोसार्वजनिक रूप से उपलब्ध और पहचान-रहित कियागयाएक बड़ाEHR डेटाबेस है। Beth Israel Deaconess Medical Center में 2008 से 2019 के बीच इलाज पाए लगभग 300,000 मरीज़ोंमें से लेखकोंने **574 मामले** चुने, जो**आठ लक्ष्य निदानों**में फैले थे — इनमें पेट से जुड़ीबीमारियाँऔर आंतरिक चिकित्साकीआपात स्थितियाँ थीं, जैसे appendicitis, pancreatitis, pneumonia और urinary-tract infection। हर मामले में MIRA सीमित जानकारीसे शुरू करताथाऔर चरण-दर-चरण तय करताथाकि आगे क्याकरनाहै। संरचनावास्तविक चिकित्सकीय जाँच जैसीथी, लेकिन वह ऐसे ऐतिहासिक रिकॉर्ड पर चल रहीथीजिसकावास्तविक अंत पहले से फ़ाइल में मौजूद था।

फिर उन्हींमामलोंपर उसके प्रदर्शन कीतुलनाचिकित्सकोंसे कीगई।

सैंडबॉक्स वाले EHR में “स्वायत्त एजेंट” कावास्तव में क्याअर्थ है

यहाँतीन शब्द बहुत काम कर रहे हैं, इसलिए उन्हें अलग-अलग समझनाउपयोगीहै।

एजेंट काअर्थ है कि प्रणालीकेवल किसीप्रॉम्प्ट काउत्तर देने वालाचैटबॉट नहींहै। यह एक चक्र चलतीहै: मौजूदा स्थिति देखो, कोई क्रियाचुनो— यह जाँच मंगाओ, वह इतिहास पूछो— परिणाम देखो, फिर अगलीक्रियाचुनो, और यह क्रम निदान तथायोजनातक चलताहै। “बुद्धिमत्ता” भाषामॉडल देताहै; *एजेंसी*उसके चारोंओर काढाँचादेताहै, जो मॉडल के पाठ कोअनुमत EHR क्रियाओं में बदलताहै और परिणाम फिर मॉडल कोलौटाताहै।

सैंडबॉक्स वालाEHR किसीसक्रिय अस्पताल प्रणालीकानाम नहींहै, बल्कि मेडिकल-रिकॉर्ड प्रणालीकीएक सिम्युलेटेड और अलग-थलग प्रतिलिपि है। MIRA कोई जाँच “मंगा” सकताहै और उस मरीज़ कोवास्तविक जीवन में संभावितपरिणाम प्राप्त कर सकताहै, क्योंकि मामलाऐतिहासिक है और उत्तर पहले से रिकॉर्ड में दर्ज है। उसकीकोई भीक्रियाकिसीवास्तविक मरीज़ याक्लिनिक कोप्रभावित नहींकरती।

स्वायत्त काअर्थ है कि वह सैंडबॉक्स के *अंदर* बिनाकिसीइंसान कोबीच में शामिल किए मामलाशुरू से अंत तक पूरा करताहै। इसकाअर्थ यह **नहीं**कि उसे वास्तविक मरीज़ोंकोबिनानिगरानीसंभालने के लिए छोड़ दियागयाथा। ये दो बहुत अलग दावे हैं, और इस अध्ययन ने केवल पहले कोपरखा।

सैंडबॉक्स के बारे में एक और बात आसानीसे गलत समझीजासकतीहै। MIRA अपनीपसंद कीकोई भीजाँच *मंगा*सकता है, लेकिन प्रणालीपरिणाम तभीलौटासकतीहै जब वह जाँच उस मरीज़ के लिए वास्तविक जीवन में **वास्तव में कीगई थी**। यदि वह ऐसारक्त-पैनल माँगताहै जोमरीज़ कोमिलाथा, तोवास्तविक मान लौटते हैं; यदि ऐसाएक्स-रे माँगताहै जो कभीकियाहीनहींगया, तोउत्तर मिलताहै “N/A — could not be performed”, कोई गढ़ीहुई संख्यानहीं। इतिहास भी इसीतरह काम करताहै: रिकॉर्ड में माँगीगई जानकारी न होतोसिम्युलेटेड मरीज़ कहताहै कि उसे नहींपता। इसलिए MIRA कोई ऐसानिर्णायक परीक्षण जिसुई ढंग से नहींबुलासकताजोअसल में कभीहुआ हीनहीं, उसे उसीसामग्री के भीतर काम करनापड़ताहै जोवास्तविक जाँच-प्रक्रियामें मौजूद थी।

यह अंतर इसलिए महत्वपूर्ण है क्योंकि सुर्खियोंमें इस्तेमाल होने वालाशब्द — “स्वायत्त” — सबसे आसानीसे

> उसके दूसरे, कहींअधिक व्यावहारिक अर्थ में पढ़ लियाजाताहै।

उन्हें क्यामिला

बेंचमार्क पर **MIRA नैदानिक सटीकतामें चिकित्सकोंसे आगे रहा**— लेकिन इस सुर्खीके पीछे तीन अलग संख्याएँ हैं, जिन्हें अलग रखनाज़रूरीहै।

सभी**574 मामलों**में अकेले MIRA ने **88.9%** बार सहीनिदान बताया। स्कोर MIMIC-IV में दर्ज वास्तविक डिस्चार्ज निदान के विरुद्ध कियागया। इंसानोंके साथ सीधे मुकाबले में डॉक्टरोंने सभी574 मामले नहींकिए; उन्होंने एक साझा**311-मामलोंके उपसमूह** पर काम कियाऔर उनके पास वहीरिकॉर्ड तथाउपकरण थे जोMIRA कोउपलब्ध थे। इन्हीं311 मामलोंमें MIRA कास्कोर **87.8%** था। उसकीतुलनाअलग-अलग भर्तीकिए गए दोसमूहोंसे कीगई। पहला**वरिष्ठ समूह** था 7 से 11 वर्ष के अनुभव वाले चार प्रमाणित चिकित्सक, जिनकास्कोर **78.1%** रहा। दूसरा**मिश्रित अनुभव वालासमूह** था मुख्यतःजूनियर रेज़िडेंट और दोविशेषज्ञ, जोवास्तविक आपात विभाग कीस्टाफ़िंग के अधिक करीब था;उसकास्कोर **71.1%** था। मामले और उपकरण समान थे — नतीजे में AI पहले, अनुभवीडॉक्टर दूसरे और जूनियर-प्रधान टीम तीसरे स्थान पर रही। शोधपत्र की अपनीसावधान भाषाहै कि आठोंबीमारियोंमें MIRA चिकित्सकोंके “लगातार बराबर रहाऔर अक्सर उनसे आगे निकला।”

आगे के निर्णय भीकाफ़ीमजबूत रहे: अधिकांश दिशानिर्देशोंके अनुरूप, दवाओं के लिहाज़ से सुरक्षित और भर्तीके निर्णयों में उपयुक्त। लेखक पाँच सुरक्षाश्रेणियों— दवा-अंतर्क्रिया, गुर्दे के अनुसार खुराक, एलर्जी, QT जोखिम और opioid prescribing — में किसीउच्च-गंभीरतावालीदवा-त्रुटि कीरिपोर्ट नहींकरते और 468 में से 467 prescriptions कोसहीबताते हैं। साथ हीवे साफ़ लिखते हैं कि प्रणालीने “100% reliability” हासिल नहींकी। सीधे पढ़ें तोयह प्रभावशालीपरिणाम है: पूरा मामलाचलाने वालाएजेंट निदान के स्कोर में डॉक्टरोंसे आगे निकला।

लेकिन जीत कीप्रकृति *भी*उतनीही*महत्वपूर्ण* है जितनीजीत। शोधपत्र पढ़ने वाले स्वतंत्र विशेषज्ञोंने बताया कि बढ़त मुख्य रूप से कहाँसे आई। Science Media Centre के विशेषज्ञ पैनल पर University of Sheffield के Dr Wei Xing ने कहा कि AI कीनैदानिक बढ़त **मुख्यतःउन स्थितियोंसे आई जहाँजाँच-परिणाम साफ़ थे**, जैसे appendicitis और pancreatitis, जिनमें निर्णायक स्कैन यालैब मान सवाल कोकाफ़ीहद तक सुलझादेते हैं। Pneumonia और urinary-tract infection — आपात विभाग में आने के दोबहुत समान्य कारण — में *AI और डॉक्टर दोनों*सबसे खराब रहे और उनके बीच अंतर भीसबसे छोटाथा।

दोनोंपक्षोंके पास उपलब्ध जानकारीमें भीअसमानताथी, और यह शोधपत्र के अपने आँकड़ोंमें दिखाई देतीहै। MIRA ने लैब जाँचोंपर अधिक भरोसाकिया;उसने सामान्य देखभाल में उपलब्ध प्रयोगशालाanalytes में से लगभग **51%** काउपयोग किया, जबकि प्रमाणित चिकित्सकोंने लगभग **28%** का— प्रति मामले औसतन सात अतिरिक्त रक्त-पैरामीटर। लेखक सावधानी से बताते हैं कि यह डेटासेट में वास्तविक नियमित देखभाल के स्तर से *कम* था;इसलिए यह हर उपलब्ध जाँच मंगाने की रणनीति नहींथी;महँगीcross-sectional imaging के उपयोग में भीकोई व्यवस्थित वृद्धि नहींदिखी। फिर भीअधिक जानकारी अपने-आप नैदानिक सटीकताबढ़ासकतीहै। इसलिए यह समान जानकारीवाले खिलाड़ियोंके बीच पूरीतरह समान मुकाबलानहीं था,जैसाXing ने सीधे इंगित किया। यदि किसीचिकित्सक कोलागत, असुविधायादेरीकीचिंताकिए बिनाहर सस्तीरक्त-जाँच मंगाने कीछूट मिले, तोउसकाप्रदर्शन भीकागज़ पर बेहतर दिख सकताहै।

“डॉक्टरों से आगे” का अर्थ “बेहतर डॉक्टर” नहीं है

बेंचमार्क पर जीत को जरूरत से ज्यादा पढ़ना आसान है। “MIRA ने अधिक स्कोर किया” और “MIRA बेहतर है” के बीच कम से कम तीन महत्वपूर्ण बातें आती हैं।

पहली, **उसे किस चीज़ के विरुद्ध अंक दिए गए।** University of Edinburgh की Professor Julie Jacko ने ध्यान दिलाया कि MIRA के कई प्रमुख परिणाम मूल डेटासेट में दर्ज चीज़ों के सापेक्ष परिभाषित हैं। यानी प्रणाली को **दर्ज चिकित्सकीय व्यवहार को दोहराने** पर इनाम मिल सकता है, जरूरी नहीं कि वह सर्वोत्तम संभव देखभाल दिखा रही हो। ऐतिहासिक चार्ट उत्तर-कुंजी बन जाता है। बेंचमार्क बनाने का यह उचित तरीका है, लेकिन यह किए गए काम से मेल मात्र है, किसी पूर्ण अर्थ में सर्वोत्तम चिकित्सा नहीं। निष्पक्षता के लिए यह भी कहना चाहिए कि यह समस्या सबसे अधिक *उपचार-संगत* वाले मेट्रिक्स पर लागू होती है — क्या MIRA के आदेश चार्ट से मेल खाते थे? — जबकि मुख्य नैदानिक सटीकता डिस्चार्ज निदान के विरुद्ध मापी गई, जो केवल दस्तावेज़ की नकल से अधिक वास्तविक परिणाम के करीब है।

दूसरी, पहले बताई गई **जानकारी की असमानता** बहुत अधिक रक्त-जाँचें — उपलब्ध analytes का लगभग 51% बनाम 28% — अधिक साक्ष्य देती हैं। सटीकता के अंतर का कुछ हिस्सा बेहतर तर्क से नहीं, अधिक जानकारी हासिल करने से आया हो सकता है।

तीसरी, **यह कहाँ चला।** यह सैंडबॉक्स था, पुराने मामलों पर, पहले से चुने गए आठ निदानों के भीतर और केवल पाठ के माध्यम से। University of Oxford के Dr Dominic Oliver के शब्दों में वास्तविक चिकित्सकीय आकलन केवल इस पर निर्भर नहीं करता कि मरीज़ क्या कहते हैं, बल्कि इस पर भी कि वे कैसे कहते हैं — साथ ही शारीरिक परीक्षण, दिखाई देने वाला व्यवहार और शरीर-भाषा भी मायने रखते हैं। तैयार रिकॉर्ड पढ़ने वाला पाठ-आधारित एजेंट इनमें से कुछ नहीं देखता। और यदि किसी मरीज़ की समस्या उन आठ चुने हुए निदानों में आती ही नहीं, तो यह अध्ययन उसके बारे में कुछ नहीं बता सकता।

समीक्षकों ने एक और शांत लेकिन महत्वपूर्ण चिंता उठाई: **प्रशिक्षण-डेटा से संभावित दूषण।** MIMIC-IV सार्वजनिक है और उस पर बहुत साहित्य उपलब्ध है, इसलिए खुले इंटरनेट पर प्रशिक्षित भाषा मॉडल ने पहले से शोधपत्र, मामले की चर्चाएँ या स्वयं डेटा के हिस्से देखे हो सकते हैं। University of Sheffield के Dr Midhun Parakkal Unni ने इसे सीधे उठाया। यदि कुछ उत्तर प्रशिक्षण डेटा में मौजूद थे, तो प्रदर्शन का एक हिस्सा चिकित्सकीय तर्क के बजाय स्मृति हो सकता है, और स्वतंत्र पुनरावृत्ति ही दोनों को अलग कर सकती है। उल्लेखनीय है कि लेखक इस चिंता को टालते नहीं हैं। वे लिखते हैं कि उनके परिणामों को सावधानी से “संभावित ऊपरी सीमा” माना जा सकता है और वे दूसरे सार्वजनिक मामलों पर समानांतर को अधिक आँक सकते हैं — यानी शोधपत्र अपनी ही सुरक्षा पर सीमा लगा रहा है।

इनमें से कोई बात MIRA को कम दिलचस्प नहीं बनाती। इससे केवल सही व्याख्या अधिक संतुलित होती है: पुराने मामलों वाले एक बेंचमार्क पर पूरा रिकॉर्ड संभाल सकने वाला एजेंट साफ़ जाँच-परिणाम वाली बीमारियों में डॉक्टरों से आगे निकला — कुछ हद तक अधिक जाँचें मंगाकर और कुछ हद तक दर्ज चार्ट से मेल खाते निर्णय लेकर। यह वास्तविक क्षमता का प्रदर्शन है, इस बात का फैसला नहीं कि मशीनें अब डॉक्टरों से बेहतर निदान करती हैं।

यह क्या *साबित नहीं* करता

- यह **नहीं** दिखाता कि MIRA वास्तविक मरीज़ों पर काम करता है। हर मामला पुराना और समीयुलेटेड था, किसी वास्तविक मरीज़ का इलाज MIRA ने नहीं संभाला।
- यह **नहीं** दिखाता कि यह आठ निदानों के बाहर भी काम करता है। पहले से चुने गए समूह से बाहर की स्थितियाँ — यानी चिकित्सा की अधिकांश जटिल और अस्पष्ट वास्तविकता — परखी ही नहीं गईं।
- यह **नहीं** दिखाता कि इसे तैनात करना सुरक्षित है। लेखक स्वयं लिखते हैं कि सामानांतर, सुरक्षा और शासन के लिए अभी भविष्य-दर्शी, वास्तविक दुनिया के अध्ययन चाहिए।

- यह डॉक्टरोंके साथ पूरीतरह निष्पक्ष आमने-सामने तुलना**स्थापित नहीं**करता। MIRA ने बहुत अधिक रक्त-जाँचोंका उपयोग किया— उपलब्ध analytes कालगभग 51%, डॉक्टरोंके 28% के मुकाबले — और कुछ उपचार-संबंधीपरिणाम सर्वोत्तम देखभाल के स्वतंत्र मानक के बजाय ऐतिहासिक चार्ट से आंशिक रूप से मिलान करके आंके गए।
- यह **नहीं**दिखाता कि परिणाम स्मृति के प्रभाव से मुक्त हैं। साथ्वजनिक MIMIC-IV डेटामॉडल के प्रशिक्षण-सामग्री से ओवरलैप कर सकताहै; इसे स्वतंत्र पुनरावृत्ति से हीबहर कियाजासकताहै।
- इसकाअर्थ **"AI डॉक्टरोंकीजगह लेताहै" नहीं**है। यह ऐसानिर्णय-सहायक तंत्र है जोरिकॉर्ड के भीतर क्रियाएँ कर सकताहै; समीक्षकोंकासामान्य निष्कर्ष है कि वास्तविक उपयोग चिकित्सकोंके साथ साझेदारीमें होगा, जहाँअधिकार डॉक्टर के पास रहेगाऔर वह सारीजानकारीभीउपलब्ध होगीजोपेज-रिकॉर्ड में नहींआती।

साक्ष्य कितनामजबूत है?

दावे कोदोहिस्सोंमें बाँटनाचाहिए, क्योंकि दोनोंके लिए साक्ष्य कीताकत बहुत अलग है।

क्षमताके प्रदर्शन के रूप में — यानीयह दिखाने के लिए कि भाषामॉडल एजेंट EHR सैंडबॉक्स के भीतर पूरे चिकित्सकीय मामले कोइतिहास, जाँच, निदान और आदेशोंकीकड़ीमें शुरू से अंत तक चलासकताहै — काम सचमुच नयाऔर काफ़ी विश्वसनीय है। यहीभाग वास्तविक प्रगति है और इसीपर सबसे अधिक ध्यान देनाचाहिए।

श्रेष्ठताके दावे के रूप में — यानीयह कहनाकि MIRA चिकित्सकोंसे *बेहतर* है — साक्ष्य सीमित है और सावधानीसे पढ़ा जानाचाहिए। यह एक खास पुराने मामलोंवाले बेंचमार्क, आठ निदानों जानकारीकीअसमानता— अधिक जाँचें — ऐतिहासिक चार्ट से बनीउत्तर-कुंजीऔर प्रशिक्षण-डेटादूषण कीवास्तविक संभावनाके भीतर लागू होताहै। इतनाकहनाउचित है कि "इस बेंचमार्क पर एजेंट ने प्रभावशालीप्रदर्शन किया।" इससे यह कहनाउचित नहींहोजाताकि "यह डॉक्टर से बेहतर निदानकरताहै", और लेखक भीदूसरादावानहींकरते।

सबसे उपयोगीदृष्टिकोण न "AI ने डॉक्टरोंकोहरादिया" है, न "यह सिर्फ खिलौनाहै।" बेहतर पढ़ाई यह है: चिकित्सकीय AI काएक नयाप्रकार — जोकेवल सवालोंकाउत्तर नहींदेताबल्कि रिकॉर्ड के भीतर क्रियाएँ करताहै — एक कठिन पुराने मामलोंवाले बेंचमार्क पर अच्छाचला। अब उसे उस जगह खुद कोसाबित करनाहै जहाँअभीपरीक्षण नहींहुआ: वास्तविक, पहले से वर्गीकृत न किए गए मरीज़ोंपर, भविष्य-दर्शीअध्ययन में और उचित नगिरानीतथाशासन के तहत।

यह क्योंमायने रखताहै

पिछले कुछ वर्षोंमें चिकित्सकीय AI कादिलचस्प सवाल चुपचाप बदल गयाहै। पहले सवाल था "क्यामॉडल सहीनिदान कर सकताहै?" — और अधिक साफ़-सुथरे परीक्षण-समूहोंपर मॉडल बार-बार हाँकहते रहे। MIRA इससे कठिन सवाल की ओर बदलाव दिखाताहै: "क्यामॉडल पूराकाम कर सकताहै — जानकारीजुटासकताहै, जाँच मंगासकताहै, परिणाम समझ सकता है, फैसलाकर सकताहै और कार्रवाई कर सकताहै — पूरे कार्यप्रवाह में?" यह अधिक उपयोगीऔर अधिक ईमानदार सवाल है, क्योंकि कठिनाई और जोखिम दोनोंकार्रवाई में हीसामने आते हैं।

इसीलिए परिणाम कोकैसे पेश कियाजाताहै, यह मायने रखताहै। सुर्खीवालासहज निष्कर्ष — *AI डॉक्टरोंसे बेहतर* — अध्ययन के सबसे कम नए और सबसे कम समर्थित हिस्से कीओर ध्यान खींचताहै। वास्तव में नयाभाग छोटालेकिन अधिक महत्वपूर्ण है: ऐसाएजेंट जोपूरे रिकॉर्ड में आगे बढ़ सकताहै। यदि यह क्षमतावास्तविक दुनियामें भीटिकतीहै, तोसबसे पहले यह **कार्यप्रवाह** बदलेगी, यह नहींकि उसकाप्रभारीकौन है। डॉक्टर गायब नहींहोता; जाँच मंगाना, परिणाम पढ़ना और बाक़ीबचे कामोंकापीछाकरना— यानीकार्यप्रवाह — वह जगह है जहाँऐसाउपकरण सबसे पहले उतरताहै।

और इससे प्रमाण का बोझ सही दिशा में जाता है। पुराने मामलों के लीडरबोर्ड पर जीत भविष्य-दर्शी परीक्षण करने का कारण है, उसका विकल्प नहीं। लेखक भी यही कहते हैं। MIRA को ईमानदारी से पढ़ें तो यह अगले अध्ययन का निमंत्रण है — उसी मानक पर जिसकी चिकित्सा को पहले से समझ है: मरीजों पर माप, केवल बेंचमार्क पर नहीं।

संक्षेप में

MIRA एक स्वायत्त AI एजेंट है जो सैंडबॉक्स वाले इलेक्ट्रॉनिक स्वास्थ्य रिकॉर्ड के भीतर काम करता है। यह इतिहास ले सकता है, प्रयोगशाला इमेजिंग और सूक्ष्मजीव-विज्ञान जाँचें मंगा और समझ सकता है, निदान तक पहुँच सकता है और उपचार-योजना लिख सकता है। सार्वजनिक MIMIC-IV डेटाबेस के 574 पुराने मामलों और पहले से चुने गए आठ निदानों पर इसे परखने पर यह नैदानिक सटीकता में चिकित्सकों से आगे रहा — कुल 88.9%; सीधे मुकाबले में 87.8% बनाम 78.1% — और इसके अधिकांश निर्णय दिशानिर्देशों के अनुरूप तथा दवाओं के लिहाज़ से सुरक्षित थे। लेकिन मूल्यांकन केवल पठ पर आधारित पुराने रिकॉर्ड की सिमुलेशन थी; बढ़त का बड़ा हिस्सा सामान्य जाँच-परिणाम वाली बीमारियों में आया; MIRA ने डॉक्टरों की तुलना में बहुत अधिक रक्त-जाँचों का उपयोग किया — उपलब्ध analytes का लगभग 51% बनाम 28%; कुछ उपचार-संबंधी नतीजों को मूल चार्ट में दर्ज चीज़ों के विरुद्ध अंकित किया गया; और सार्वजनिक डेटासेट प्रशिक्षण-डेटा दूषण का वास्तविक जोखिम पैदा करता है, जिसे लेखक स्वयं संभावित ऊपरी सीमा मानते हैं। असली प्रगति यह है कि एजेंट अलग-अलग सवालों का जवाब देने के बजाय पूरे कार्यप्रवाह में कार्रवाई कर सकता है। समन्वयीकरण, सुरक्षा और शासन के लिए अभी वास्तविक दुनिया में भविष्य-दर्शी अध्ययन चाहिए। इसलिए सही निष्कर्ष है: प्रभावशाली क्षमता का प्रदर्शन, न कि यह प्रमाण कि AI डॉक्टरों से बेहतर निदान करता है, और न ही ऐसा तंत्र जो आज वास्तविक क्लिनिक के लिए तैयार हो।

बिना बढ़ा-चढ़ाकर जाँच

शोधपत्र क्या दिखाता है: भाषा-मॉडल एजेंट MIRA सैंडबॉक्स वाले EHR के भीतर पूरा चिकित्सकीय मामला शुरू से अंत तक चला सकता है — इतिहास, जाँचें, निदान और आदेश — और आठ निदानों वाले 574 पुराने मामलों के बेंचमार्क पर नैदानिक सटीकता में चिकित्सकों से आगे रहा: कुल 88.9%, और प्रमाणित चिकित्सकों के विरुद्ध सीधे मुकाबले में 87.8% बनाम 78.1%, जबकि कोई उच्च-गंभीरता वाली दवा-त्रुटि दर्ज नहीं हुई।

क्या संभव है, पर सिद्ध नहीं कि यही क्षमता वास्तविक, पहले से वर्गीकृत न किए गए मरीजों के लिए लाभ में बदलेगी; और यह कि सटीकता की बढ़त वास्तव में बेहतर तर्क से आई, न कि अधिक जाँचें मंगाने, ऐतिहासिक चार्ट से मेल या सार्वजनिक डेटा को पहले देखने से।

यह क्या नहीं दिखाता कि MIRA अपने आठ निदानों के बाहर काम करता है; कि इसे वास्तविक क्लिनिक में तैनात करना सुरक्षित है; कि समान जानकारी वाले निष्पक्ष मुकाबले में यह डॉक्टरों से आगे रहता है; कि यह चिकित्सकों की जगह ले सकता है; या कि परिणाम प्रशिक्षण-डेटा दूषण से मुक्त हैं।

मुख्य सीमाएँ: पुराने मामलों पर आधारित सिम्युलेटेड, केवल-पठ मूल्यांकन; पहले से चुने गए आठ निदान; जानकारी में असमानता — MIRA ने उपलब्ध analytes का लगभग 51% इस्तेमाल किया, डॉक्टरों ने लगभग 28%, मुख्यतः सस्ती रक्त-जाँचों में, इमेजिंग में नहीं; कुछ उपचार-परिणाम ऐतिहासिक चार्ट से मिलान करके अंकित किए गए; और सार्वजनिक MIMIC-IV बेंचमार्क को भाषा-मॉडल ने प्रशिक्षण के दौरान देखा हो सकता है, जिसे लेखक स्वयं प्रदर्शन की संभावित ऊपरी सीमा मानते हैं।

सामान्य पाठक को किन बातों से सावधान रहना चाहिए? इस बात पर उच्च कि MIRA वास्तविक और नई क्षमता दिखाता है — ऐसा एजेंट जो केवल चैटबॉट की तरह जवाब देने के बजाय रिकॉर्ड के भीतर कार्रवाई करता है। इस दावे पर कम कि वह *चिकित्सक*

से बेहतर निदानकर्ता है: वह निष्कर्ष एक पुराने मानदंड वाले बेंचमार्क तक सीमित है और अधिक जाँचें मंगाने, चार्ट के विरुद्ध स्कोरिंग तथा संभावित डेटा-दूषण से प्रभावित हो सकता है। लेखक स्वयं सही अगला कदम बताने हैं — मरीज़ों की देखभाल के लिए इसका अर्थ निकालने से पहले भविष्य-दर्शी, वास्तविक दुनिया के अध्ययन चाहिए।

स्रोत

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