The CLEAN PAPER

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

سوال کا جواب دینا اور پورا کلینیکل کیس چلانا ایک ھی بات نہیں

autonomous MIRA اے آئی ایجنٹ ھے جو *EHR sandboxed* میں تاریخ لیتا، *microbiology*/امیجنگ/*labs* ترتیب اور *interpret* کرتا، تشخیص تک پہنچتا اور علاج *orders* لکھتا ھے۔ عوامی *MIMIC-IV* کے 574 *retrospective* کیسز، آٹھ *pre-selected diagnoses*، پر اس نے 88.9% *overall* تشخیصی درستگی حاصل کی؛ 311-کیس براہِ راست موازنہ میں 87.8%، *senior* بورڈ سے تصدیق شدہ *physicians* کے 78.1% اور *mixed-seniority* گروپ کے 71.1% کے مقابلے میں۔ *guideline-concordant largely Decisions* تھے اور مصنفین نے بلند-*medication severity* غلطیاں رپورٹ نہیں کیے۔ مگر جائزہ سیمولیشن تھی، *past* ریکارڈز پر، متن-*only*، *advantage* کا بڑا حصہ واضح ٹیسٹ والی حالات میں تھا؛ *MIRA* نے ڈاکٹرز کے مقابلے میں *far* مزید *analytes blood* استعمال کیے (~51% *vs* 28%)؛ کئی علاج نتائج *chart recorded* کے خلاف اسکور ھوئے؛ اور عوامی *MIMIC-IV* تربیتی ڈیٹا *contamination* کا حقیقی خطرہ رکھتا ھے — جسے مصنفین ممکن بالائی *bound* کہتے ھیں۔ حقیقی *advance* کام-*whole* کا بہاؤ ایجنٹ ھے، نہ یہ ثبوت کہ اے آئی ڈاکٹرز سے بہتر *diagnose* کرتی ھے یا *deployment clinic* کے لیے *ready* ھے۔

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 artfiicial intelligence agents*, Dyke Ferber, Lars Hilgers, Christiane Höper and colleagues; senior author Jakob Nikolas Kather, Nature (2026) · DOI: 5-10675-026-s41586/1038.10

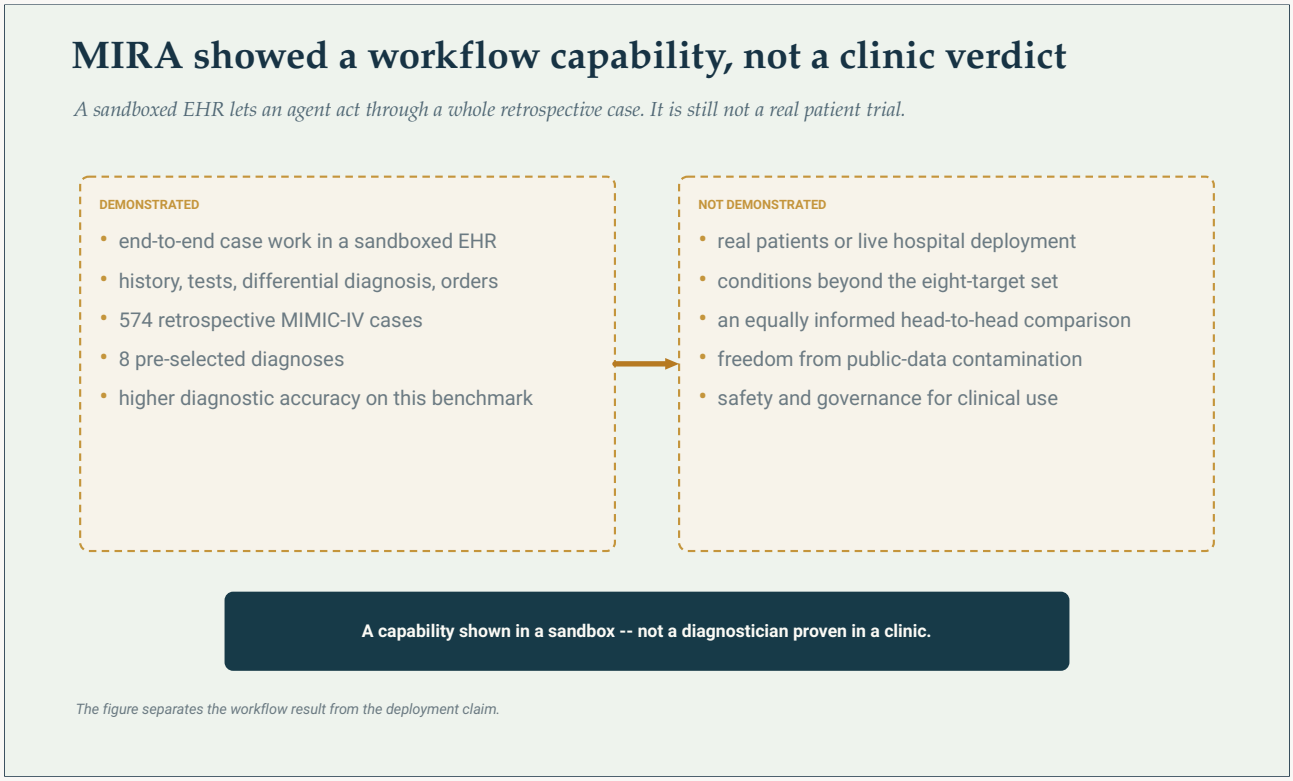
سوال کا جواب دینا اور پورا کلینیکل کیس چلانا ایک ھی بات نہیں

طبی اے آئی کی زیادہ تر کہانیوں ایسے ماڈل کے بارے میں ہوتی ہیں جو **جواب دیتا ھے**۔ آپ اسے vignette یا exam سوال دیتے ہیں، وہ تشخیص یا advice کا paragraph واپس دیتا ھے، اور اچھا اسکور کر لیتا ھے۔ مفید، مگر محدود: ایک ذہین مشیر جو chart پر خود کچھ نہیں کرتا۔

MIRA مختلف چیز کرنے کے لیے بنایا گیا ھے، اور اصل خبر یہی فرق ھے۔ ایک isolated سوال کا جواب دینے کے بجائے یہ پورا کیس چلاتا ھے: مریض ریکارڈ پڑھتا ھے، فیصلہ کرتا ھے کہ مزید کون سی تاریخ درکار ھے، لیبارٹری، امیجنگ اور microbiology ٹیسٹ ترتیب کرتا ھے، نتائج پڑھتا ھے، dffierential تشخیص محدود کرتا ھے، اور پھر اگلے مراتب لکھتا ھے — prescriptions، surgery admission، referral۔ یہ actions الیکٹرانک صحت ریکارڈ کے **اندر** لیتا ھے، معالج کی طرح، نہ کہ صرف آزاد متن generate کر کے کسی انسانی سے نقل کرواتا ھے۔

یہ حقیقی قدم ھے: advice دینے والے نظام سے کام کا بہاؤ میں actions لینے والے نظام تک۔ اور یہی وہ قدم ھے جو coverage میں جلد "اے آئی ڈاکٹرز سے بہتر" بن جاتا ھے۔ اس لیے ضروری ھے کہ کیا ٹیسٹ ھوا، اور کہاں، اسے درست رکھا جائے۔

ایک sentence میں: MIRA نے پورے کیسز شروع سے آخر تک خود چلائے اور اس بینچ مارک پر تشخیصی درستگی میں معالجین سے بہتر اسکور کیا — مگر یہ سب sandbox میں، retrospective ریکارڈز پر، پہلے سے چنی گئی آٹھ تشخیصات کے اندر، صرف متن کے ذریعے ھوا، اور اس کی بڑی برتری ان حالات میں آئی جن کے ٹیسٹ زیادہ صاف اور فیصلہ کن تھے۔ پیش رفت حقیقی ھے۔ Scoreboard کلینک نہیں۔



MIRA نے sandboxed EHR کے اندر شروع سے آخر تک کام کا بہاؤ صلاحیت دکھائی۔ اس نے حقیقی دنیا کلینیکل عملی تعیناتی، وسیع coverage، تشخیصی یا ڈاکٹرز کے ساتھ برابر معلومات پر fair براہِ راست مقابلہ ثابت نہیں کیا۔

مصنفین نے کیا کیا

MIRA — نام کا مطلب "طبی ذہانت برائے استدلال و عمل" ہے — OpenAI کے ماڈلز پر بنایا گیا خودکار ایجنٹ ہے۔ گفتگو اور عملی اقدامات کے لیے **GPT-4o** استعمال ہوتا ہے، جبکہ الگ منصوبہ بندی کے مرحلے میں OpenAI کا **o1** استدلالی ماڈل کام کرتا ہے۔ یہ ماڈلز ایک مخصوص، طبی معیاروں کے مطابق فریم ورک کے اندر رکھے گئے ہیں جو FHIR HL7 پر مبنی ہے، یعنی اسی معیار پر جس سے حقیقی ہسپتال طبی ریکارڈ کا تبادلہ کرتے ہیں، اور اس کے پاس 80,000 سے زیادہ ممکنہ طبی اقدامات کا مجموعہ ہے۔ محفوظ تجرباتی ماحول میں MIRA مریض کی تاریخ نکال سکتا ہے، لیبارٹری جانچ، امیجنگ اور خرد حیاتیاتی ٹیسٹ طلب اور ان کی تشریح کر سکتا ہے، متبادل تشخیصات بنا سکتا ہے اور علاج کا منصوبہ تیار کر سکتا ہے — مثلاً دوائیں تجویز کرنا، جراحی طریقہ کار طے کرنا یا داخلے کی منصوبہ بندی کرنا۔ ایک تفصیل شروع ہی میں اہم ہے: MIRA کے جوابوں کو نمبر دینے والے خودکار "جج" بھی GPT-4o تھے، یعنی آزمائے گئے ماڈل ہی کے خاندان سے؛ ہم مرتبہ جائزہ لینے والے نے اس پر تشویش ظاہر کی، جس کے بعد مصنفین نے بورڈ سے تصدیق شدہ ڈاکٹروں کا جائزہ بھی شامل کیا۔

ٹیسٹ سیٹ **MIMIC-IV** سے آیا، ایک بڑا عوامی، شناخت-de شدہ EHR ـ Deaconess Israel Beth database طبی مرکز میں 2008 سے 2019 کے دوران treated تقریباً 300,000 مریض میں سے مصنفین نے **574 مریض کیسز** منتخب کیے، جو **آٹھ ہدف تشخیصات** پر مشتمل تھے — pathologies abdominal اور داخلی-طب، emergencies، مثلاً appendicitis، pancreatitis pneumonia اور urinary-tract انفیکشن۔ ہر کیس میں initial limited MIRA تصویر سے شروع کرتا اور مرحلہ وار طے کرتا کہ اگلا عمل کیا ہو — حقیقی تشخیصی عمل جیسی شکل، مگر ماضی کے ریکارڈ کے خلاف، جس کا اصل ending پہلے سے موجود ہے۔

پھر اسی کیس سیٹ پر معالجین کے ساتھ کارکردگی موازنہ کیا گیا۔

"Autonomous agent in a sandboxed EHR" کا واقعی مطلب کیا ہے؟

ہیں۔ رہے اٹھا کچھ بہت الفاظ تین یہ

ایجنٹ کا مطلب chatbot نہیں جو ایک ہدایت کا جواب دے۔ نظام لوپ چلاتا ہے: موجودہ حالت دیکھو، عمل چنو (یہ ٹیسٹ ترتیب کرو، وہ تاریخ پوچھو)، نتیجہ دیکھو، اگلا عمل چنو — تشخیص اور plan تک۔ "ذہانت" زبان ماڈل ہے؛ **agency** وہ scaffolding ہے جو ماڈل کے متن کو operations EHR permitted میں تبدیل کرتی ہے اور نتائج واپس خوراک لینا کرتی ہے۔

EHR Sandboxed کا مطلب سیمولیشن شدہ، walled-off طبی-ریکارڈ نظام ہے، hospital live نہیں۔ MIRA ٹیسٹ "ترتیب" کر سکتا ہے اور مریض کا وہی نتیجہ حاصل کر سکتا ہے جو حقیقی زندگی میں موجود تھا، کیونکہ کیس ماضی کے ہے۔ اس کا کوئی عمل حقیقی مریض یا کلینک کو نہیں چھوتا۔

Autonomous کا مطلب ہے کہ انسانی-in-the-لوپ کے بغیر کیس شروع سے آخر تک مکمل کرتا ہے — **sandbox کے اندر**۔ اس کا مطلب یہ نہیں کہ اسے حقیقی مریض کو manage unsupervised کرنے کے لیے چھوڑ دیا گیا تھا۔ صرف پہلی چیز ٹیسٹ ہوئی۔

Sandbox کے بارے میں ایک اور اہم تفصیل: MIRA کوئی بھی ٹیسٹ ترتیب کر سکتا ہے، مگر نظام نتیجہ صرف اسی ٹیسٹ کا دے سکتا ہے جو اس مریض کے حقیقی ریکارڈ میں **واقعی ہوا تھا**۔ وہ blood پینل مریض نے کروایا تھا تو حقیقی قدریں ملیں گی؛ ایسا اسکین ترتیب کیا جو کبھی ہوا ہی نہیں، تو N/A" — could نہیں be کارکردگی دکھائی" ملے گا، fabricated عدد نہیں۔ تاریخ بھی اسی طرح: ریکارڈ میں جواب نہ ہو تو سیمولیشن شدہ مریض کہتا ہے معلوم نہیں۔ یعنی MIRA وہ فیصلہ کن ٹیسٹ جادو سے summon نہیں کر سکتا جو حقیقی تشخیصی عمل میں کبھی لیا ہی نہیں گیا۔

سرخی لفظ "autonomous" سب سے زیادہ اسی لیے غلط پڑھا جا سکتا ہے۔

کیا ملا

بینچ مارک پر **MIRA تشخیصی درستگی میں معالجین سے آگے تھا** — مگر سرخی کے پیچھے تین الگ اعداد ھیں جنہیں مرکب نہیں کرنا چاھیے۔

تمام **574 کیسز** پر MIRA نے درست discharge تشخیص **9%.88** وقت شناخت کیا۔ براہِ راست انسانی براہِ راست مقابلہ میں ڈاکٹرز نے تمام 574 کیسز نہیں کیے؛ انہوں نے مشترک **311-کیس subset** پر کام کیا، وھی ریکارڈز/ٹولز استعمال کرتے ھوئے جو MIRA کے پاس تھے۔ ان 311 پر MIRA **8%.87** تھا۔ انسانی موازنہ کے دو گروپس تھے۔ پہلا **senior گروپ**: چار بورڈ سے تصدیق شدہ معالجین، 7–11 سال experience، اسکور **1%.78**۔ دوسرا **ملے جلے-seniority گروپ**: زیادہ تر junior residents، ساتھ دو specialists، اسکور **1%.71**۔ ایک ھی کیسز، ایک ھی ٹولز: اے آئی پہلا، experienced ڈاکٹرز دوسرا، بھاری-junior گروپ تیسرا۔ مقالے کی محتاط phrasing ھے کہ MIRA "consistently ... equivalent to, often and exceeded" معالجین میں تمام آٹھ diseases۔

بعد کے فیصلے بھی زیادہ تر رھنما اصولوں کے مطابق، دواؤں کے لحاظ سے محفوظ اور داخلے کے فیصلوں میں مناسب تھے۔ مصنفین نے پانچ حفاظتی زمروں — دواؤں کے باھمی اثرات، گردوں کے مطابق خوراک، الرجی، QT خطرہ اور opioid تجویز — میں شدید نوعیت کی دوائی کی غلطی رپورٹ نہیں کی، اور 468 کیسز میں 467 نسخوں کو درست قرار دیا؛ ساتھ ھی یہ بھی لکھا کہ نظام نے "100% قابلِ اعتماد کارکردگی" حاصل نہیں کی۔ پہلی نظر میں یہ متاثر کن ھے: ایجنٹ پورا کیس چلاتا ھے اور تشخیصی بینچ مارک میں ڈاکٹروں سے آگے نکلتا ھے۔

مگر win کی **شکل** خود win جتنی اھم ھے۔ آزاد specialists نے نوٹ کیا کہ برتری کہاں سے آ رھی تھی۔ سائنس میڈیا مرکز ماھر پینل پر Sheffield) of (University Xing Wei Dr نے کہا کہ سرخی فائدہ زیادہ تر ایسی حالات سے ڈرائیو ھوئی جن میں **واضح ٹیسٹ نتائج تھے**، جیسے appendicitis اور pancreatitis، جہاں فیصلہ کن اسکین یا لیب قدر سوال settle کر دیتی ھے۔ Pneumonia اور infections urinary-tract — departments emergency میں بہت عام — میں اے آئی اور ڈاکٹرز دونوں نے worst کارکردگی دی اور خلا سب سے چھوٹا تھا۔

دونوں sides نے برابر معلومات استعمال بھی نہیں کی۔ مقالے کے اپنے اعداد کے مطابق MIRA معمول نگہداشت میں دستیاب لیبارٹری analytes کے تقریباً **51%** کو استعمال کرتا تھا، جبکہ بورڈ سے تصدیق شدہ معالجین تقریباً **28%** — median طور پر کیس کے لیے سات اضافی blood پیرامیٹرز۔ مصنفین اسے ڈیٹا سیٹ کے معمول-نگہداشت بنیادی موازنہ سے بھی کم کہتے ھیں، نہ کہ "ترتیب everything حکمتِ عملی، اور زیادہ-لاگت cross-sectional امیجنگ میں منظم اضافہ نہیں ملا۔ پھر بھی زیادہ معلومات خود تشخیصی درستگی بڑھا سکتی ھے۔ اس لیے یہ مکمل contest like-for-like نہیں۔ blood Cheap ٹیسٹ کی لاگت/discomfort/delay کم سمجھنے والا معالج بھی مقالہ پر زیادہ درست دکھ سکتا ھے۔

"ڈاکٹرز کو سے بہتر رھا کیا" کا مطلب "بہتر ڈاکٹر" نہیں

بینچ مارک میں برتری کو ضرورت سے زیادہ معنی دینا آسان ھے۔ MIRA" نے زیادہ اسکور کیا" اور MIRA" بہتر ڈاکٹر ھے" کے درمیان کم از کم تین الگ فاصلے ھیں۔

پہلا سوال یہ ھے کہ **نظام کو کس چیز کے مقابل اسکور کیا گیا؟** Edinburgh of University کی Jacko Julie Professor نے نشاندھی کی کہ MIRA کے کئی اھم نتائج بنیادی ڈیٹا سیٹ میں درج طبی رویے کے نسبت متعین کیے گئے ھیں — یعنی نظام کو لازماً بہترین نگہداشت ثابت کرنے پر نہیں، بلکہ **تاریخی ریکارڈ میں درج طبی فیصلوں کو دوبارہ پیدا کرنے** پر انعام مل سکتا ھے۔ یوں تاریخی چارٹ ھی جواب نامہ بن جاتا ھے۔ بینچ مارک بنانے کا یہ معقول طریقہ ھے، مگر یہ اس بات سے مطابقت ناپتا ھے کہ

حقیقت میں کیا کیا گیا تھا؛ یہ مطلق معنوں میں بہترین نگہداشت کی درستگی ثابت نہیں کرتا۔ انصافاً، یہ اعتراض سب سے زیادہ علاج سے مطابقت والے پیمانوں پر لاگو ہوتا ہے — کیا MIRA کے احکامات چارٹ سے ملتے تھے؟ — جبکہ نمایاں تشخیصی درستگی کو ڈسچارج تشخیص کے مقابل ناپا گیا، جو محض دستاویز کی بازگشت کے مقابل ایک زیادہ حقیقی نتیجہ ہے۔

دوسرا، **معلومات عدم توازن**: دستیاب blood ٹیسٹ کا 51% بمقابلہ 28% شواھد کی مختلف مقدار ہے۔ درستگی خلا کا کچھ حصہ بہتر استدلال کے بجائے زیادہ شواھد خریدنے سے آ سکتا ہے۔

تیسرا، **کہاں چلایا گیا**۔ یہ retrospective sandbox، pre-selected کیسز، آٹھ تشخیصات، اور متن-صرف ترتیب تھی۔ Dominic Dr (University Oliver of Oxford) نے یاد دلایا کہ حقیقی کلینیکل جائزہ صرف مریض کیا کہتا ہے پر نہیں؛ کیسے کہتا ہے، طبیعی exam، مشاھدہ شدہ رویّہ اور جسم زبان بھی اھم ھیں — جو finished متن ریکارڈ پڑھنے والے ایجنٹ کے پاس نہیں۔ اور اگر مسئلہ ان آٹھ تشخیصات میں مطابقت نہ ھو تو مطالعہ اس کے بارے میں کچھ نہیں کہتی۔

ایک اور quiet تشویش **ڈیٹا آلودگی** ہے۔ MIMIC-IV عوامی ہے اور discussed heavily ہے، اس لیے تربیت-internet یافتہ زبان ماڈل شاید مقالات، کیس discussions، یا ڈیٹا itsefl دیکھ چکا ھو۔ Dr Unni Parakkal Midhun (University of Sheffield) نے یہ خطرہ براہِ راست نشان کیا۔ اگر جوابات تربیت ڈیٹا میں تھے تو کارکردگی کا کچھ حصہ یادداشت ھو سکتا ہے، کلینیکل استدلال نہیں۔ آزاد replication ھی فرق واضح کرے گی۔ اھم بات یہ ہے کہ مصنفین خود تشویش dismiss نہیں کرتے: وہ نتائج کو "ممکن بالائی حد" کہتے ھیں اور لکھتے ھیں کہ عوامی کیسز پر عمومی نتیجہ کو overestimate کر سکتے ھیں۔

یہ سب MIRA کو غیر دلچسپ نہیں بناتا۔ یہ صحیح تشریح calibrate کرتا ہے: retrospective بینچ مارک پر مکمل-ریکارڈ ایجنٹ نے صاف ٹیسٹ والی تشخیصات میں ڈاکٹرز سے بہتر اسکور کیا — جزوی طور پر زیادہ ٹیسٹ ترتیب کر کے، جزوی طور پر ماضی کے chart کے ساتھ مطابقت میں۔ یہ حقیقی صلاحیت مظاھرہ ہے، فیصلہ نہیں کہ machines ڈاکٹرز سے بہتر diagnose کرتی ھیں۔

یہ مطالعہ کیا ثابت نہیں کرتی

- یہ **نہیں** دکھاتی کہ MIRA حقیقی مریض پر کام کرتا ہے۔ ھر کیس retrospective اور سیمولیشن شدہ تھا؛ کسی مریض کو MIRA نے واقعی manage نہیں کیا۔
- یہ آٹھ تشخیصات سے باھر کارکردگی **نہیں** دکھاتی۔ طب کی غیر واضح، majority undffierentiated ٹیسٹ نہیں ھوئی۔
- یہ عملی تعیناتی حفاظت **ثابت نہیں** کرتی۔ مصنفین خود لکھتے ھیں کہ عمومی نتیجہ، حفاظت اور حکمرانی کے لیے prospective حقیقی دنیا مطالعات درکار ھیں۔
- یہ ڈاکٹرز کے ساتھ براہِ راست مقابلہ **نہیں**۔ MIRA نے far مزید blood ٹیسٹ استعمال کیے (~51% دستیاب analytes بمقابلہ ~28%)، اور کئی علاج نتائج ماضی کے chart کے خلاف اسکور ھوئے۔
- یہ نتیجہ حفظ سے پاک ہے، یہ **نہیں** دکھاتی۔ عوامی MIMIC-IV ماڈل تربیت کے ساتھ overlap کر سکتا ہے؛ آزاد replication چاھیے۔
- یہ **"اے آئی ڈاکٹرز کو جگہ لینا کرتی ہے"** نہیں دکھاتی۔ یہ ریکارڈ میں actions لینے والا فیصلہ-تائید ایجنٹ ہے؛ حقیقی استعمال میں معالجین اختیار اور وہ سیاق دیتے رھیں گے جو متن ریکارڈ میں نہیں۔

شواھد کتنے مضبوط ھیں؟

دعویٰ کو دو حصوں میں تقسیم کرنا بہتر ہے۔

صلاحیت مظاہرہ کے طور پر — زبان-ماڈل ایجنٹ EHR sandboxed میں پورا کلینیکل کیس تاریخ، ٹیسٹ، تشخیص اور مراتب سمیت شروع سے آخر تک چلا سکتا ہے — کام novel genuinely اور معقول حد تک convincing ہے۔ یہی نئی چیز ہے اور یہی توجہ کے قابل ہے۔

برتری کے دعوے کے طور پر— کہ MIRA معالجین سے بہتر ہے—شواہد محدود ہیں۔ یہ نتیجہ ایک مخصوص retrospective بینچ مارک، آٹھ تشخیصات، معلومات کے عدم توازن، ماضی کے چارٹ پر مبنی اسکورنگ اور ممکنہ تربیتی ڈیٹا آلودگی کے اندر قائم رہتا ہے۔ اتنا شواہد یہ کہنے کے لیے کافی ہے کہ "اس بینچ مارک پر ایجنٹ کی کارکردگی متاثر کن تھی"، مگر یہ ثابت کرنے کے لیے نہیں کہ وہ حقیقی طبی عمل میں بہتر ڈاکٹر ہے۔

مفید مؤقف نہ "اے آئی سے بہتر ہے ڈاکٹرز" ہے، نہ toy" ماڈل"۔ بہتر مطالعہ: طبی اے آئی کی نئی قسم — جو isolated سوالات کے جواب کے بجائے ریکارڈ کے اندر **actions** لیتی ہے — مشکل retrospective بینچ مارک پر اچھی رہی، اور اب اسے وہاں ثابت کرنا ہے جہاں ابھی ٹیسٹ نہیں ہوئی: حقیقی، undffierentiated مریض، ،prospectively حکمرانی کے تحت۔

یہ کیوں اہم ہے

کچھ سالوں میں طبی اے آئی کا دلچسپ سوال quietly بدل گیا ہے۔ پہلے سوال تھا "کیا ماڈل تشخیص ٹھیک کر سکتا ہے؟"۔ اب زیادہ مشکل سوال ہے: "کیا ماڈل **پورا کام** کر سکتا ہے — gather، ترتیب، interpret، act decide — پورے کام کا بہاؤ میں؟" یہی زیادہ مفید اور دیانت دار سوال ہے، کیونکہ dffiiculty اور خطرہ دونوں عمل میں رہتے ہیں۔

اسی لیے framing اہم ہے۔ سرخی reflex — **اے آئی outperforms ڈاکٹرز** — نتیجہ کے سب سے کم novel اور سب سے کم supported حصے کی طرف اشارہ کرتا ہے۔ نئی اور consequential چیز یہ ہے کہ ایجنٹ پورے ریکارڈ میں حرکت کر سکتا ہے۔ اگر صلاحیت برقرار رہتی ہے تو یہ پہلے **کام کا بہاؤ** بدلتی ہے، اختیار نہیں۔ ڈاکٹر فوراً جگہ لینا نہیں ہوتا، ordering، نتائج chase کرنا، تشریح اور دستاویزات جیسی کام کا بہاؤ کام پہلے بدلتی ہیں۔

اور of burden ثبوت صحیح سمت میں reset ہوتا ہے۔ Retrospective win leaderboard prospective ٹرائل چلانے کی وجہ ہے، اس کا substitute نہیں۔ مصنفین بھی یہی کہتے ہیں۔ MIRA کی دیانت دار مطالعہ اگلے مطالعے کی دعوت ہے — مریض پر ناپا گیا، معیارات پر نہیں۔

صاف خلاصہ

MIRA ایک خودکار اے آئی ایجنٹ ہے جو محفوظ EHR ماحول میں مریض کی تاریخ لیتا، لیبارٹری جانچ، امیجنگ اور خرد حیاتیاتی ٹیسٹ طلب اور ان کی تشریح کرتا، تشخیص تک پہنچتا اور علاج کے احکامات لکھتا ہے۔ عوامی MIMIC-IV کے 574 سابقہ کیسز اور آٹھ پہلے سے منتخب تشخیصات پر اس کی مجموعی تشخیصی درستگی 88.9% تھی؛ 311 کیسز کے براہِ راست موازنے میں 87.8%، جبکہ سینئر بورڈ سے تصدیق شدہ ڈاکٹروں کی 78.1% اور مختلف سینیارٹی والے گروہ کی 71.1%۔ فیصلے زیادہ تر رہنما اصولوں کے مطابق تھے اور مصنفین نے شدید نوعیت کی دوائی کی غلطیاں رپورٹ نہیں کیں۔ مگر یہ حقیقی کلینک نہیں بلکہ سابقہ ریکارڈز پر، صرف متن کے ساتھ چلائی گئی سیمولیشن تھی؛ برتری کا بڑا حصہ ان حالتوں میں تھا جہاں واضح ٹیسٹ دستیاب تھے؛ MIRA نے ڈاکٹروں سے کہیں زیادہ خون کے تجزیے استعمال کیے (تقریباً 51% بمقابلہ 28%)؛ بعض علاجی فیصلوں کو تاریخی چارٹ ہی کے مطابق نمبر ملے؛ اور چونکہ MIMIC-IV عوامی ڈیٹا ہے، تربیتی آلودگی کا حقیقی خطرہ موجود ہے — جسے مصنفین ممکنہ بالائی حد کہتے ہیں۔ اصل پیش رفت مکمل کام کے بہاؤ والا ایجنٹ ہے، نہ یہ ثبوت کہ اے آئی ڈاکٹر سے بہتر تشخیص کرتی ہے یا کلینک

میں تعیناتی کے لیے تیار ہے۔

بلا مبالغہ جانچ

مقالے کیا دکھاتا ہے: MIRA زبان-ماڈل ایجنٹ EHR sandboxed میں entire کیس شروع سے آخر تک چلا سکتا ہے — تاریخ، ٹیسٹ، تشخیص، مراتب — اور 574-کیس retrospective بینچ مارک میں بلند تشخیصی درستگی دکھاتا ہے (88.9% مجموعی؛ 311-کیس subset میں 87.8% بمقابلہ بورڈ سے تصدیق شدہ معالجین 78.1%)، جبکہ بلند-medication severity غلطیاں رپورٹ نہیں ہوئے۔

کیا قرینِ قیاس ہے مگر ثابت نہیں: یہ صلاحیت حقیقی undffierentiated مریض کے لیے فائدہ میں translate ہوگی؛ درستگی برتری بہتر استدلال سے ہے، نہ کہ زیادہ ٹیسٹ، ماضی کے chart مطابقت یا memorized عوامی ڈیٹا سے۔

یہ کیا نہیں دکھاتا: آٹھ تشخیصات سے باہر کارکردگی؛ محفوظ عملی تعیناتی؛ informed equally ڈاکٹرز کے خلاف fair برتری؛ معالجین کی متبادل؛ تربیتی ڈیٹا آلودگی سے پاک نتائج۔

اہم حدود: retrospective سیمولیشن شدہ متن-صرف جائزہ؛ آٹھ pre-selected تشخیصات؛ معلومات عدم توازن (دستیاب analytes کا ~51% بمقابلہ 28%)؛ علاج نتائج جزوی طور پر ماضی کے chart کے خلاف؛ عوامی MIMIC-IV جسے ماڈل تربیت میں دیکھ چکا ہو سکتا ہے۔

عام قاری کو کتنا اعتماد ہونا چاہیے؟ زیادہ اعتماد کہ novel MIRA صلاحیت مظاہرہ ہے — chatbot سے آگے، ریکارڈ کے اندر actions لینے والا ایجنٹ۔ کم اعتماد کہ وہ **معالج سے بہتر تشخیص کار** ثابت ہو چکا ہے: وہ دعویٰ ایک retrospective بینچ مارک، مزید-ٹیسٹ عدم توازن، پر-chart مبنی اسکورنگ اور ممکن آلودگی سے حد میں ہے۔ مصنفین کی اپنی bottom لائن درست ہے: مریض نگہداشت کے لیے معنی نکالنے سے پہلے prospective حقیقی دنیا مطالعہ ضروری ہے۔

ماخذ

(2026) Nature agents, intelligence artfiicial medical autonomous Towards al., et Ferber
https://doi.org/10.1038/s41586-026-10675-5

abstract peer-reviewed — 42310457 PubMed
https://pubmed.ncbi.nlm.nih.gov/42310457/

(2026) AMIE and MIRA to reaction expert — Centre Media Science
https://www.sciencemediacentre.org/expert-reaction-to-presentation-of-two-new-medical-ai-r

framing doctors' 'outperforms the illustrates — (2026) News-Medical
https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-si
aspx