



The CLEAN PAPER

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

MIRA یک پرونده پزشکی کامل را در sandbox اجرا کرد و در benchmark از پزشکان جلو زد — اما این هنوز کلینیک نیست

MIRA، عاملی خودمختار مبتنی بر GPT-4o و o1، در محیط شبیه‌سازی‌شده پرونده سلامت الکترونیک می‌تواند شرح حال بگیرد، آزمایش و تصویربرداری درخواست کند، نتایج را تفسیر کند، تشخیص بسازد و دستور درمان بنویسد. روی ۵۷۴ پرونده گذشته‌نگر MIMIC-IV در هشت تشخیص، دقت کلی ۸۸٫۹٪ داشت و در زیرمجموعه‌ای از ۳۱۱ پرونده به ۸۷٫۸٪ رسید، در برابر ۷۸٫۱٪ برای پزشکان دارای برد تخصصی. اما این شبیه‌سازی فقط متنی و روی پرونده‌های گذشته بود؛ MIRA حدود ۵۱٪ از موارد آزمایشگاهی موجود را به کار گرفت، در برابر ۲۸٪ برای پزشکان، و بخش بزرگی از برتری در بیماری‌هایی بود که آزمون‌های تشخیصی روشنی داشتند. این نتیجه توانایی در یک محیط بسته را نشان می‌دهد، نه پزشک خودمختار برای بیماران واقعی.

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: 5-10675-026-s41586/1038.10

پاسخ دادن به یک سؤال با اداره کردن کل پرونده یکی نیست

بیشتر داستان‌های هوش مصنوعی پزشکی درباره مدلی اند که پاسخ می‌دهد. یک vignette یا سؤال امتحانی به آن می‌دهید، تشخیص یا پاراگرافی از توصیه تحویل می‌دهد و نمره خوبی می‌گیرد. مفید است، اما محدود: مشاوره با هوش که هرگز به پرونده دست نمی‌زند.

MIRA برای کار دیگری ساخته شده و همین تفاوت خبر اصلی است. به جای پاسخ دادن به یک سؤال، کل پرونده را پیش می‌برد: سابقه بیمار را می‌خواند، تصمیم می‌گیرد چه بخش‌هایی از شرح حال هنوز لازم است، آزمایش‌های خون، تصویربرداری و میکروبیولوژی سفارش می‌دهد، نتایج را می‌خواند، تشخیص افتراقی را محدود می‌کند و سپس دستورهای بعدی را می‌نویسد — نسخه، بستری، ارجاع برای جراحی. این کارها را با اقدام درون پرونده الکترونیک سلامت انجام می‌دهد، شبیه clinician، نه با تولید متن آزادی که انسان مجبور باشد دوباره وارد سامانه کند.

این یک گام واقعی است: از سامانه‌ای که مشورت می‌دهد به سامانه‌ای که در طول workflow اقدام می‌کند. و دقیقاً همان نوع گامی است که در پوشش رسانه‌ای به «AI» از پزشکان بهتر شد» صاف می‌شود. پس باید دقیق بگویم چه چیز و کجا آزموده شد.

نسخه یک جمله‌ای: MIRA پرونده‌های کامل را خودش پیش برد و در این benchmark از پزشکان در دقت تشخیصی جلو زد — اما این کار را در sandbox، روی پرونده‌های گذشته، در هشت تشخیص از پیش انتخاب شده، فقط با متن انجام داد و بخش بزرگی از برتری‌اش در بیماری‌هایی بود که تست‌های روشن و تعیین کننده داشتند. پیشرفت واقعی است. جدول امتیاز، کلینیک نیست.

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.

DEMONSTRATED

- end-to-end case work in a sandboxed EHR
- history, tests, differential diagnosis, orders
- 574 retrospective MIMIC-IV cases
- 8 pre-selected diagnoses
- higher diagnostic accuracy on this benchmark

NOT DEMONSTRATED

- real patients or live hospital deployment
- conditions beyond the eight-target set
- an equally informed head-to-head comparison
- freedom from public-data contamination
- safety and governance for clinical use

A capability shown in a sandbox -- not a diagnostician proven in a clinic.

The figure separates the workflow result from the deployment claim.

MIRA توانایی workflow انتباه‌انها را درون EHR شده sandbox نشان داد. استقرار بالینی واقعی، پوشش تشخیصی گسترده یا مقایسه منصفانه دو طرف با اطلاعات برابر را نشان نداد.

نویسندگان چه کردند

MIRA — Action and Reasoning for Intelligence Medical — یک عامل خودمختار بر پایه مدل‌های OpenAI است: بخشی که گفت‌وگو و اقدام می‌کند روی GPT-4o اجرا می‌شود و مرحله برنامه‌ریزی جداگانه از مدل استدلالی o1 استفاده می‌کند. این‌ها درون چارچوب سفارشی و سازگار با استاندارد قرار دارند — بر پایه HL7، FHIR استاندارد داده‌ای که بیمارستان‌های واقعی برای جابه‌جایی پرونده استفاده می‌کنند — و جعبه‌ابزاری با بیش از هشتاد هزار اقدام بالینی ممکن دارند. درون MIRA sandbox می‌تواند سابقه بیمار را بگیرد، آزمایش خون، تصویربرداری و میکروبیولوژی سفارش و تفسیر کند، تشخیص افتراقی بسازد و برنامه درمان را تنظیم کند — تجویز دارو، برنامه‌ریزی جراحی، طرح بستری. یک جزئیات مهم: «داوران» خودکار که پاسخ‌های MIRA را امتیاز دادند خودشان GPT-4o بودند — همان خانواده مدلی که آزموده می‌شد — و پس از مطرح‌شدن نگرانی از سوی peer reviewer، نویسندگان ارزیابی پزشکی board-certified را نیز اضافه کردند.

مجموعه آزمون از MIMIC-IV آمد، پایگاه بزرگ و عمومی EHR با داده‌های ناشناس شده. از حدود ۳۰۰٬۰۰۰ بیمار درمان‌شده در Beth Center Medical Deaconess Israel بین ۲۰۰۸ و ۲۰۱۹، نویسندگان ۵۷۴ پرونده بیمار را در هشت تشخیص هدف انتخاب کردند — بیماری‌های شکمی و فوریت‌های داخلی، از جمله pneumonia، appendicitis، و urinary-tract infection. در هر پرونده، MIRA از تصویر محدودی شروع می‌کرد و باید گام‌به‌گام تصمیم می‌گرفت بعد چه کند — همان شکل workup واقعی، اما روی پرونده‌ای که پایان واقعی‌اش از قبل ثبت شده بود.

سپس عملکردش با پزشکان روی همان پرونده‌ها مقایسه شد.

«عامل خودمختار در EHR» شده sandbox دقیقاً یعنی چه

دارند. بازکردن ارزش و می‌کنند حمل زیادی بار واژه سه Agent یعنی سامانه‌ی chatbot نیست که فقط به prompt پاسخ دهد. یک حلقه اجرا می‌کند: وضعیت فعلی را می‌بیند، اقدامی انتخاب می‌کند (این آزمایش را بگیر، آن بخش شرح حال را پیرس)، نتیجه را می‌بیند، اقدام بعدی را انتخاب می‌کند — تا به تشخیص و برنامه برسد. «هوش» مدل زبانی است؛ scaffolding agency اطراف آن است که متن مدل را به عملیات مجاز EHR تبدیل می‌کند و نتیجه را دوباره به مدل می‌دهد.

EHR Sandboxed یعنی نسخه شبیه‌سازی‌شده و دیوارکشی‌شده پرونده پزشکی، نه سامانه زنده بیمارستان. MIRA می‌تواند آزمایشی را «سفارش» دهد و نتیجه‌ای را بگیرد که آن بیمار واقعاً گرفته، چون پرونده تاریخی و پاسخ از قبل ثبت شده است. هیچ اقدامی به بیمار یا کلینیک واقعی نمی‌رسد.

Autonomous یعنی پرونده را از ابتدا تا انتها بدون انسان در حلقه کامل می‌کند — درون sandbox. معنایش این نیست که برای مدیریت بدون نظارت بیماران واقعی آزاد شده است. این دوا دعا بسیار متفاوت‌اند و فقط اولی آزموده شده.

یک نکته دیگر درباره MIRA sandbox می‌تواند هر آزمایشی که بخواهد سفارش دهد، اما سامانه فقط وقتی نتیجه برمی‌گرداند که آن آزمایش واقعاً برای همین بیمار در دنیای واقعی انجام شده باشد. اگر آزمایش خونی را بخواهد که بیمار گرفته، مقدارهای واقعی را می‌گیرد؛ اگر اسکن هرگز انجام نشده‌ای را بخواهد، پاسخ N/A — قابل انجام نیست — می‌آید، نه عدد ساختگی. شرح حال هم همین‌طور است: اگر پرونده پاسخ پرسش را نداشته باشد، بیمار شبیه‌سازی‌شده فقط می‌گوید نمی‌داند. پس MIRA نمی‌تواند تست تعیین‌کننده‌ای را که در واقعیت هرگز گرفته نشده احضار کند؛ فقط با چیزی کار می‌کند که workup واقعی شاملش بوده است.

این تفاوت مهم است چون واژه تیرپسند «خودمختار» بیش از همه احتمال دارد به معنای دوم خوانده شود.

چه پیدا کردند

در MIRA benchmark از پزشکان در دقت تشخیصی جلوزد — اما تیتز سه عدد مختلف را پنهان می‌کند و راحت می‌شود آن‌ها را قاطی کرد، پس جداشان کنیم.

در کل ۵۷۴ پرونده، MIRA در ۸۸٫۹٪ موارد تشخیص درست را نام برد — در برابر diagnosis هنگام ترخیص که در MIMIC-IV برای هر بیمار ثبت شده بود. در مقایسه مستقیم با انسان، پزشکان همه ۵۷۴ مورد را کار نکردند؛ روی زیرمجموعه مشترک ۳۱۱ پرونده کار کردند، با همان پرونده‌ها و ابزارهایی که MIRA داشت. روی همین ۳۱۱ مورد، MIRA ۸۷٫۸٪ گرفت و در برابر دو گروه جداگانه سنجیده شد. گروه اول ارشد بود: چهار پزشک board-certified با ۷ تا ۱۱ سال تجربه که ۷۸٫۱٪ گرفتند. گروه دوم ترکیبی از سطوح تجربه بود: بیشترهای junior resident به اضافه دو متخصص، نزدیک تر به staffing واقعی اورژانس، که ۷۱٫۱٪ گرفتند. پرونده‌ها و ابزارها یکسان بودند — ترتیب: AI اول، پزشکان باتجربه دوم، تیم junior-heavy سوم. عبارت محتاطانه خود مقاله این است که MIRA در هر هشت بیماری «به‌طور پایدار هم‌سطح و اغلب بهتر» از پزشکان بود.

تصمیم‌های پایین‌دستی هم نسبتاً خوب بود: عمدتاً مطابق guideline، از نظر دارویی این و برای بستری مناسب. نویسندگان در پنج رده ایمنی (تداخل دارویی، تنظیم دوز کلیوی، آلرژی، خطر QT و تجویز opioid) هیچ خطای دارویی با شدت بالا گزارش نکردند و نسخه‌ها در ۴۶۷ مورد از ۴۶۸ درست بودند — در عین حال یادآوری می‌کنند سامانه «به ۱۰۰٪ قابلیت اطمینان نرسید». در نگاه اول نتیجه چشمگیر است: عاملی که کل پرونده را اجرا می‌کند و در تشخیص از پزشکان جلو می‌زند.

اما شکل این پیروزی به اندازه خود پیروزی مهم است. متخصصان مستقل که مقاله را خواندند نشان دادند برتری واقعاً از جکا آمده. در پنل متخصصان Centre Media Science، دکتر Xing Wei از Sheffield of University گفت عدد تیتز — جلوزدن AI در دقت تشخیص — بیشتر توسط بیماری‌هایی با نتیجه آزمایش روشن مانند appendicitis و pancreatitis رانده شده، جایی که اسکن یا عدد آزمایشگاهی تعیین‌کننده سؤال را می‌بندد. در pneumonia و infection urinary-tract — دو علت بسیار شایع مراجعه به اورژانس — هم AI و هم پزشکان بدترین عملکرد را داشتند و فاصله‌شان کمترین بود.

همچنین دو طرف با اطلاعات کاملاً متقارن بازی نکردند؛ این در اعداد خود مقاله دیده می‌شود. MIRA بیشتر به آزمایشگاه تکیه کرد: حدود ۵۱٪ های analyte آزمایشگاهی موجود در مراقبت عادی را استفاده کرد، در برابر حدود ۲۸٪ برای پزشکان board-certified — median هفت پارامتر خون بیشتر در هر پرونده. نویسندگان با دقت می‌گویند این هنوز کمتر از baseline خود dataset برای مراقبت واقعی است و «همه‌چیز را سفارش بده» نبوده؛ همچنین افزایش منظم تصویربرداری cross-sectional گران‌تر دیده نشد. با این حال اطلاعات بیشتر خودش می‌تواند دقت تشخیصی را بالا ببرد — پس این کاملاً رقابت مثل‌به‌مثل میان دو طرف با اطلاعات برابر نیست، نکته‌ای که Xing مستقیماً مطرح کرد. ی‌clinician که آزاد باشد هر تست خونی ارزان را بدون وزن‌دادن به هزینه، ناراحتی یا تأخیر سفارش دهد نیز روی کاغذ تیزتر به نظر می‌رسد.

چرا «از پزشکان جلوزد» به معنی «پزشک بهتری است» نیست

بردن benchmark را آسان می‌شود بیش‌ازحد خواند. سه چیز میان MIRA «امتیاز بالاتر گرفت» و MIRA «بهتر است» قرار دارند.

اول، در برابر چه چیزی امتیاز داده شده. پروفیسور Jacko Julie از Edinburgh of University اشاره کرد چند پیامد کلیدی MIRA نسبت به چیزی تعریف شده‌اند که در dataset پایه ثبت شده — یعنی سامانه برای بازسازی رفتار بالینی ثبت‌شده پاداش می‌گیرد، نه لزوماً برای نشان‌دادن مراقبت بهینه. پرونده تاریخی key answer می‌شود. این روش معقولی برای benchmark است، اما توافق با آنچه انجام شد را می‌سنجد، نه درستی مطلق. برای انصاف، این مشکل بیشتر به‌های metric هم‌راستایی درمان ضربه می‌زند — آیا دستورهای MIRA با chart جور بود؟ — در حالی که دقت تشخیصی تیتز در برابر diagnosis ترخیص سنجیده می‌شود که به outcome واقعی نزدیک‌تر است.

دوم، عدم تقارن اطلاعات: آزمایش خون بسیار بیشتر (حدود ۵۱٪ های analyte موجود در برابر ۲۸٪) یعنی شواهد بیشتر. احتمالاً بخشی از فاصله دقت خریداری شده، نه صرفاً استدلال شده.

سوم، محیط اجرا. sandbox پرونده گذشته، هشت diagnosis از پیش انتخاب شده، و فقط متن. ارزیابی بالینی واقعی، همانطور که دکتر Oliver Dominic از Oxford University گفت، نه فقط به چیزی که بیمار می گوید بلکه به چطور گفتنش وابسته است — همراه با معاینه فیزیکی، رفتار مشاهده شده و زبان بدن که عامل متنی خواننده پرونده تکمیل شده هیچ کدام را نمی بیند. بیماری که مشکلش در یکی از آن هشت diagnosis نباشد اصلاً خارج از دامنه این مطالعه است.

نگرانی آرام تر دیگری که ها reviewer مطرح کردند آلودگی داده آموزش است. MIMIC-IV عمومی و موضوع مقالات فراوان است، بنابراین مدل زبانی آموزش دیده روی اینترنت باز شاید قبلاً مقاله ها، بحث های موردی یا خود داده ها را دیده باشد. دکتر Midhun Unni Parakkal از Sheffield of University مستقیم این را مطرح کرد — اگر بعضی جواب ها در داده آموزش بوده باشند، بخشی از عملکرد حافظه است نه استدلال بالینی، و فقط تکرار مستقل می تواند این دو را جدا کند. نکته مثبت این که نویسندگان موضوع را با دست رد نمی کنند: می نویسند نتیجه ها «می توانند با احتیاط به عنوان حد بالای ممکن» تفسیر شوند و «ممکن است تعمیم به پرونده های عمومی دیگر را بیش برآورد کنند» — مقاله ای که خودش سقفی روی تیتراش می گذارد.

هیچ کدام MIRA را کم جالب نمی کنند. فقط خوانش درست را calibrated می کنند: در retrospective، benchmark، عاملی که توانست در کل پرونده اقدام کند، در های diagnosis با تست های تمیز از پزشکان جلوزد — تا حدی با سفارش آزمایش بیشتر، تا حدی با توافق با chart ثبت شده. این نمایش توانایی واقعی است، نه حکم این که ماشین اکنون بهتر از پزشک تشخیص می دهد.

این مقاله چه چیزی را اثبات نمی کند

- نشان نمی دهد MIRA روی بیمار واقعی کار می کند. همه پرونده ها retrospective و شبیه سازی شده بودند؛ هیچ بیمار واقعی هرگز توسط آن مدیریت نشد.
- نشان نمی دهد فراتر از هشت تشخیص کار می کند. بیماری های بیرون از مجموعه از پیش انتخاب شده — اکثریت نامرتب و تفکیک نشده پزشکی — آزموده نشدند.
- نشان نمی دهد برای استقرار ایمن است. خود نویسندگان می نویسند تعمیم، ایمنی و governance هنوز به مطالعه prospective و واقعی نیاز دارند.
- مقایسه مستقیم کاملاً منصفانه با پزشکان را ثابت نمی کند. آزمایش خون بسیار بیشتری استفاده کرد (حدود ۵۱٪ های analyte موجود در برابر ۲۸٪)، و چند پیامد درمانی تا حدی با chart ثبت شده سنجیده شدند نه بهترین مراقبت. ground-truth.
- نشان نمی دهد نتیجه از حفظ کردن داده ها آزاد است. داده عمومی MIMIC-IV شاید با training مدل هم پوشانی داشته باشد؛ تکرار مستقل باید این را رد کند.
- معنایش AI «جای پزشک را می گیرد» نیست. پشتیبان تصمیمی است که می تواند درون پرونده اقدام کند؛ اجماع ها reviewer این است که استفاده واقعی در partnership با clinician خواهد بود، با اختیار انسانی و اطلاعاتی که متن پرونده نمی تواند بدهد.

قدرت شواهد چقدر است؟

ادعا را دو نیم کنید، چون شواهد برای هر نیمه بسیار متفاوت است.

به عنوان نمایش توانایی — این که عامل مدل زبانی می تواند یک پرونده بالینی را در sandbox EHR از ابتدا تا انتها پیش ببرد و شرح حال، تست، تشخیص و دستورها را زنجیر کند — کار واقعاً تازه و نسبتاً قانع کننده است. این بخش جدید است و ارزش توجه دارد.

به عنوان ادعای برتری — این که MIRA بهتر از پزشکان است — شواهد محدودند و باید محتاط خوانده شوند. نتیجه روی یک benchmark retrospective خاص، هشت diagnosis، با عدم تقارن اطلاعات (تست بیشتر)، key answer برگرفته از chart تاریخی و احتمال آلودگی data training صدق می کند. کافی است بگوییم «عامل در این benchmark چشمگیر بود». برای گفتن «از پزشک diagnostician بهتری است» کافی نیست، و خود نویسندگان هم ادعای دوم را نمی کنند.

موضع مفید نه AI «پزشکان را شکست داد» و نه «فقط اسباب بازی است». این است: نوع تازه ای از AI پزشکی — عاملی که در پرونده اقدام می کند نه فقط جواب می دهد — روی retrospective benchmark سخت خوب عمل کرد و اکنون باید خودش را جایی ثابت کند که هنوز آزموده نشده: روی بیماران واقعی و تفکیک نشده، به صورت prospective و زیر governance.

چرا مهم است

در چند سال گذشته پرسش جالب هوش مصنوعی پزشکی آرام تغییر کرده. قبلاً «آیا مدل تشخیص درست می دهد؟» بود — و مدل ها روی test های set مرتب تر مرتباً جواب بله می دادند. MIRA گذار به پرسش سخت تر را نشان می دهد: «آیا مدل می تواند کار را انجام دهد — جمع آوری، سفارش، تفسیر، تصمیم، اقدام — در کل «workflow» این پرسش مفید تر و صادقانه تر است، چون هم دشواری و هم خطر در «اقدام کردن» زندگی می کنند.

به همین دلیل framing مهم است. واکنش تیر — AI از پزشکان بهتر شد — به کم نوآورانه ترین و کم پشتیبانی ترین بخش نتیجه اشاره می کند. چیز واقعاً تازه کوچک تر و پیامدارتر است: عاملی که می تواند در کل پرونده حرکت کند. اگر این توانایی دوام بیاورد، workflow را بسیار پیش تر از این که مشخص کند چه کسی مسئول آن است تغییر می دهد. پزشک حذف نمی شود؛ سفارش دادن، تفسیر کردن، دنبال کردن نتیجه ها — خود workflow — نخستین جایی است که چنین ابزاری وارد می شود.

و بار اثبات را در جهت درست تنظیم می کند. بردن leaderboard روی پرونده retrospective دلیل برای اجرای prospective trial است، نه جایگزین آن. خود نویسندگان همین را می گویند. خوانش صادقانه MIRA دعوت به مطالعه بعدی است — با همان استاندارد که پزشکی از قبل می شناسد: اندازه گیری روی بیمار، نه benchmark.

خلاصه تمیز

MIRA یک عامل AI خودمختار است که در پرونده الکترونیک sandbox کار می کند: می تواند شرح حال بگیرد، آزمایش خون، تصویربرداری و میکروبیولوژی سفارش و تفسیر کند، به diagnosis برسد و برنامه درمان بنویسد. روی ۵۷۴ پرونده retrospective از پایگاه عمومی MIMIC-IV و هشت diagnosis از پیش انتخاب شده، در دقت تشخیص از پزشکان جلو زد (۸۸٫۹٪ در کل؛ ۸۷٫۸٪ در برابر ۷۸٫۱٪ در مقایسه مستقیم) و تصمیم هایی عمدتاً مطابق guideline و از نظر دارویی ایمن گرفت. اما ارزیابی شبیه سازی روی پرونده های گذشته و فقط متنی بود؛ بخش بزرگی از برتری روی بیماری هایی با نتایج تست روشن بود؛ MIRA آزمایش خون بسیار بیشتری از پزشکان استفاده کرد (حدود ۵۱٪ های analyte موجود در برابر ۲۸٪)؛ چند outcome درمانی با چیزی که chart اصلی ثبت کرده بود امتیاز داده شدند؛ و dataset عمومی خطر واقعی contamination training-data دارد — چیزی که خود نویسندگان آن را حد بالای ممکن برای اعدادشان می نامند. پیشرفت واقعی، عاملی است که در کل workflow اقدام می کند نه ی chatbot که سؤال منفرد جواب دهد. نویسندگان صریحاًند که تعمیم، ایمنی و governance هنوز مطالعه prospective و واقعی می خواهد — خوانش درست همین است: نمایش توانایی چشمگیر، نه اثبات این که AI بهتر از پزشک تشخیص می دهد یا آماده کلینیک است.

بررسی بی‌پرده

مقاله چه نشان می‌دهد: عامل مدل زبانی MIRA می‌تواند یک پرونده بالینی را درون sandbox EHR از ابتدا تا انتها اجرا کند — شرح حال، تست، diagnosis دستور — و در retrospective benchmark شامل ۵۷۴ پرونده و هشت diagnosis از پزشکان در دقت تشخیصی جلو بزند (۸۸٫۹٪ کل؛ ۸۷٫۸٪ در برابر ۷۸٫۱٪ پزشکان (board-certified) بدون خطای دارویی با شدت بالا).

چه چیزی محتمل است اما اثبات نشده: این توانایی به منفعت برای بیماران واقعی و تفکیک نشده تبدیل شود؛ یا برتری دقت واقعاً ناشی از reasoning بهتر باشد نه تست بیشتر، توافق با chart ثبت شده یا حفظ کردن داده عمومی.

چه چیزی را نشان نمی‌دهد: MIRA خارج از هشت diagnosis کار می‌کند؛ برای استقرار ایمن است؛ در مقایسه برابر از نظر اطلاعات از پزشکان جلو می‌زند؛ clinician را جایگزین می‌کند؛ یا نتیجه از contamination training-data آزاد است.

محدودیت‌های اصلی: ارزیابی، retrospective، شبیه‌سازی شده و فقط متنی؛ هشت diagnosis از پیش انتخاب شده؛ عدم تقارن اطلاعات (آزمایش خون بیشتر — حدود ۵۱٪ ها analytela در برابر ۲۸٪، در تست‌های کم‌هزینه نه؛ imaging های outcome درمانی تا حدی امتیاز گرفته در برابر chart تاریخی؛ و benchmark عمومی MIMIC-IV که مدل زبانی ممکن است در training دیده باشد — چیزی که خود نویسندگان آن را حد بالای احتمالی عملکرد می‌نامند.

یک خواننده عمومی چقدر باید مطمئن باشد؟ اطمینان بالا که MIRA نمایش توانایی واقعی و تازه‌ای است — عاملی که درون پرونده عمل می‌کند، نه صرفاً chatbot. اطمینان پایین که «diagnostician» بهتری از پزشک» باشد: این ادعا به یک retrospective benchmark محدود است و با سفارش تست، scoring در برابر chart و contamination احتمالی مخدوش می‌شود. جمع‌بندی خود نویسندگان درست است: پیش از آن که برای مراقبت بیمار معنایی داشته باشد، به مطالعه prospective و دنیای واقعی نیاز دارد.

منابع

(2026) Nature agents, intelligence artificial 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>