The Clean Paper

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

একটা AI agent পুরো patient case নিজে চালিয়েছে — simulator-এ, পুরোনো record-এর ওপর

Autonomous medical-AI agent historical case-based simulator-এ পুরো clinical workflow চালিয়েছে — test order, result interpretation, iterative diagnosis ও management। 574 case overall benchmark এবং 311 case human comparison-এ strong performance দেখালেও agent বেশি test order করেছে, environment sandbox ছিল এবং MIMIC-IV contamination/retrospective limitation আছে। Result long-horizon medical reasoning capability দেখায়; real patient care-এর autonomy বা safety নয়।

Written by Lucio Vaglio · Cross-checked by Vera Rubin · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

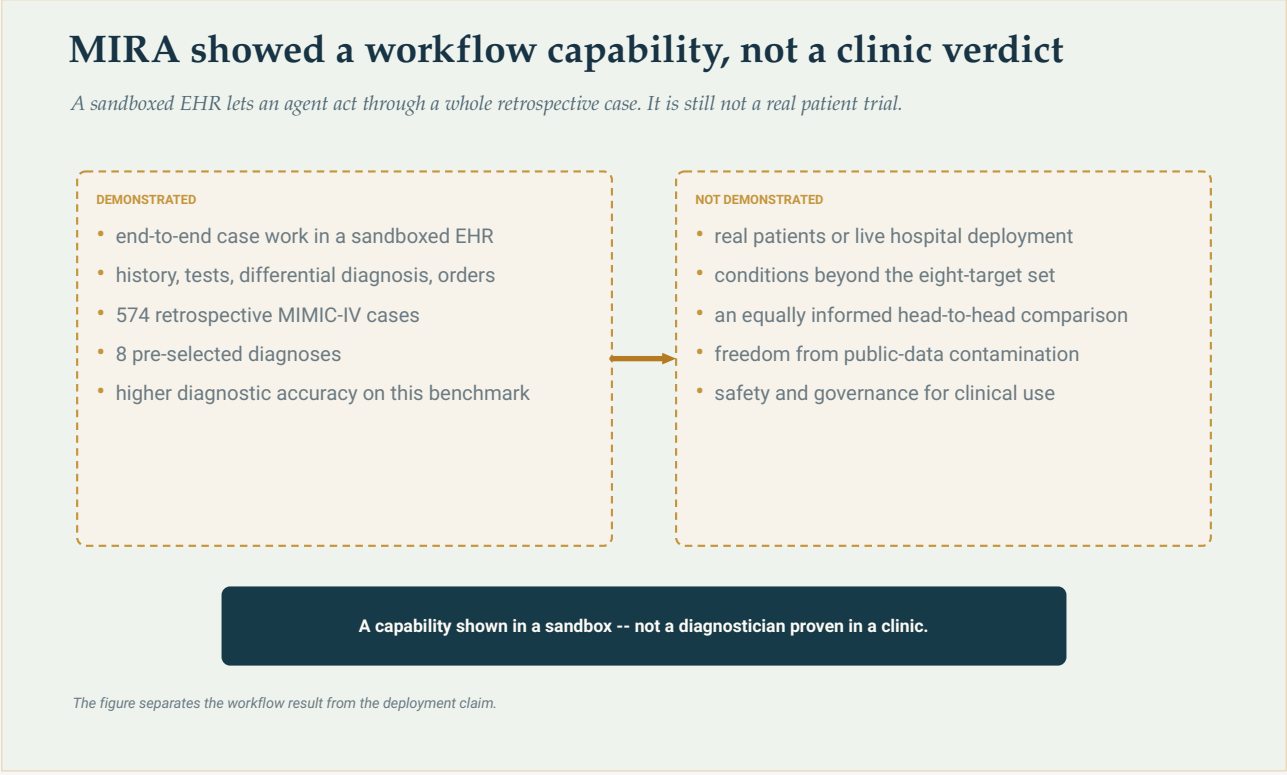
Paper: *Towards autonomous medical artificial intelligence agents*, Dyke Ferber, Lars Hilgers, Christiane Höper and colleagues; senior author Jakob Nikolas Kather, Nature (2026) · DOI: 10.1038/s41586-026-10675-5

একটি প্রশ্নরে উত্তর দওেয়া আর পুরো ক্লনিকি্যাল ক্ষত্রে চালানো এক জনিসি নয়

চকিৎিসাবষিয়ক AI গল্পরে বশেরি ভাগ মডলে **উত্তর দয়ে**: সংক্ষপ্তি ক্লনিকি্যাল বর্ণনা/পরীক্ষা প্রশ্ন দলিে রো গনর্িণয় বা পরামর্শ অনুচ্ছদে দয়ে। উপযো গী, কন্িতু সংকীর্ণ—চার্টে কো নো ক্রয়িা নয়ে না।

MIRA (Medical Intelligence for Reasoning and Action)—চকিৎিসাবষিয়ক যুক্তি ও সদ্ধিান্তরে জন্য তরৈি একটি AI ব্যবস্থা অন্য কছিু পরীক্ষা করে। এটি একটি ক্লনিকি্যাল কসে শুরু থকেে শষে পর্যন্ত পরচিালনা করে: রো গী নথি পড়ে, আর কী ইতহিাস দরকার ঠকি করে, পরীক্ষাগার/ইমজেংি/মাইক্রো বায়ো লজি নর্িদশে দয়ে, ফল পড়ে, পার্থক্যমূলক রো গনর্িণয় সংকীর্ণ করে এবং প্রসেক্রপিশন, ভর্তি বা অস্ত্রো পচার রফোরলেরে মতো নর্িদশে লখিে। মুক্ত পাঠ্য মানবকে অনুলপিি করতে দওেয়ার বদলে স্যান্ডবক্সে সীমাবদ্ধ ইলকেট্রনকি স্বাস্থ্য নথরি ভতেরে চকিৎিসকরে মতো ধাপে ধাপে কাজ করে।

এটি সত্যই সক্ষমতা ধাপ: পরামর্শদাতা ব্যবস্থা থকেে কর্মপ্রবাহে সক্রয়ি অংশগ্রহণকারী ব্যবস্থা। কন্িতু "AI চকিৎিসকদরে ছাড়য়িে যায়" শরিো নামে ফল flatten হয়। এক বাক্যে বলল: MIRA পশ্চাৎদৃষ্টমিূলক স্যান্ডবক্স মানদণ্ডে পুরো ক্ষত্রে স্বয়ংক্রয়িভাবে চালয়িে রো গনর্িণয়রে নর্িভুলতায় চকিৎিসক দলরে ওপরে স্কো র করছে—কন্িতু **বাস্তব রো গী নয়, আটটি আগে থকেে নর্িবাচতি রো গনর্িণয়, শুধু পাঠ্যভত্তিকি নথি** এবং সুবধিার বড় অংশ স্পষ্ট পরীক্ষা-ফল শর্তে; আরও রক্ত পরীক্ষা ব্যবহার করছে। অগ্রগতি বাস্তব; স্কো রবো র্ড ক্লনিকি নয়।



MIRA স্যান্ডবক্সে সীমাবদ্ধ EHR-এ শুরু থকেে শষে পর্যন্ত কর্মপ্রবাহ সক্ষমতা দখোয়। এটি বাস্তব জগতরে ব্যবহারকি মো তায়নে, বস্তিৃত রো গনর্িণয়মূলক আবরণ বা সমান তথ্যপ্রাপ্ত ন্যায্য চকিৎিসক head-to-head দখোনো করনে।ি

গবষেকরো কী করছেনে

MIRA-র আলাপ ও ক্রয়িা অংশে **GPT-4o**, আর আলাদা পরকিল্পনা ধাপে OpenAI **o1** reasoning model ব্যবহার করা হয়ছে। এগুলো HL7 FHIR-সম্মত বশিষেভাবে তরৈকিাঠামো র মধ্যে কাজ করে, যখোনে **80,000-এর বশেি** ক্লনিক্যিাল ক্রয়িা রয়ছে। স্যান্ডবক্সে এজন্টে রো গীর ইতহিাস সংগ্রহ করত, পরীক্ষা নর্দিশে ও ব্যাখ্যা করত, পার্থক্যমূলক রো গনর্ণিয় তরৈকিরত, চকিৎিসা পরকিল্পনা ও প্রসেক্রপিশন দতি, এমনকি অস্ত্রো পচাররে সময়সূচবিা ভর্তপিরকিল্পনাও করতে পার। স্বয়ংক্রয়ি স্কো রংি্যওে GPT-4o পরবিাররে মডলে ছলি—একই মডলে পরবিার ব্যবহাররে পক্ষপাতরে আশঙ্কা সহকর্মী-পর্যালো চনায় ওঠার পর লখেকরো বো র্ড-সনদপ্রাপ্ত চকিৎিসকদরে পর্যালো চনা দয়িে তা যাচাই করনে।

পরীক্ষা-সটে **অনুকরণ করা-IV** জনসাধারণরে পরচিয়মুক্ত EHR ডটোবসে। Beth Israel Deaconess চকিৎিসাবষিয়ক কন্দ্রেে 2008–2019-এর প্রায় **300,000 রো গী** থকেে **574 ক্ষত্রে**, **8 লক্ষ্য রো গনর্ণিয়**—অ্যাপন্ডেসিাইটসি, প্যানক্রয়িাটাইটসি, নডিমো নয়িা, UTI-সহ abdominal/অভ্যন্তরীণ-চকিৎিসাবজ্ঞিান emergency—বছেে নওেয়া হয়। MIRA সীমতি প্রারম্ভকি চত্রি থকেে ধাপ-by-ধাপ পরবর্তী ক্রয়িা সদ্ধিান্ত নয়ে; ঐতহিাসকি নথতিে বাস্তব ফলাফল আগে থকেইে আছ।

চকিৎিসকদরে একই ক্ষত্রে/সরঞ্জাম পরবিশেে তুলনা করা হয়, কন্তিু সব 574-তে head-to-head নয়।

> **“স্বায়ত্তশাসতি এজন্টে in a স্যান্ডবক্সে সীমাবদ্ধ EHR” মানে কী**
>
> **এজন্টে:** এটি একবাররে prompt-এর উত্তর নয়। ব্যবস্থা বর্তমান অবস্থা দখেে একটি পদক্ষপে বছেে নয়ে, ফল পায়, তারপর সইে ফলরে ভত্তিতিে পরবর্তী পদক্ষপে ঠকি করে—এভাবে রো গনর্ণিয় ও পরকিল্পনা পর্যন্ত এগো য়। ভাষা মডলেটি যুক্তি ও ভাষাগত সক্ষমতা দয়ে; চারপাশরে সফটওয়্যার কাঠামো সইে সদ্ধিান্তকে অনুমো দতি EHR কাজরে মধ্যে সীমাবদ্ধ রাখ।
> **স্যান্ডবক্সে সীমাবদ্ধ EHR:** এটি চলমান হাসপাতাল ব্যবস্থা নয়, বরং ঐতহিাসকি নথি দয়িে তরৈি বন্ধ simulation। MIRA কো নো পরীক্ষা “নর্দিশে” দলিে ফল কবেল তখনই ফরে, যদি বাস্তব রো গীর ইতহিাসে সইে পরীক্ষা সত্যহিি করা হয়ে থাক। না হলে উত্তর আসে “N/A—could not be performed”; কো নো মান বানয়িে দওেয়া হয় না। নথতিে তথ্য না থাকলে simulated patient-ও সটেি “জান না”।
> **স্বায়ত্তশাসতি:** মানব-in-লুপ ছাড়া স্যান্ডবক্স ক্ষত্রে শুরু থকেে শষে পর্যন্ত সম্পূর্ণ। বাস্তব রো গী unsupervised manage করা নয়। শরিো নাম শব্দটরি এই দুই অর্থ আলাদা রাখা জরুর।

তারা কী পয়েছেনে

রো গনর্ণিয়রে নর্ভিুলতায় MIRA মানদণ্ডে চকিৎিসকরে ওপরে ছলি, কন্তিু **তনিটি সংখ্যা আলাদা**।

সব **574 ক্ষত্রে**-এ MIRA ছাড়পত্রকালীন রো গনর্ণিয়রে সঙ্গে **88.9%** মলি করে। মানব head-to-head subset **311 ক্ষত্রে**। ওই একই 311-এ MIRA **87.8%**।

জ্যষ্ঠে তুলনা: 4 বো র্ড-সনদপ্রাপ্ত চকিৎিসক, **7–11 বছর অভজ্ঞিতা**, স্কো র **78.1%**। মশ্রি-seniority দল—মূলত কনষ্ঠি resident + 2 বশিষেজ্ঞ—স্কো র **71.1%**। একই ক্ষত্রেগুলো/সরঞ্জামে নর্দিশে দওেয়া: MIRA, জ্যষ্ঠে, কনষ্ঠি-ভারী। গবষেণাপত্র নজিে “ধারাবাহকিভাবে সমতুল্য to, এবং প্রায়ই exceeded” ভাষা ব্যবহার করে, সব আটটি রো গ জুড়ে।

Downstream সদ্ধিান্ত মূলত guideline-concordant; ওষুধ-নর্দিশে উচ্চ-তীব্রতা ত্রুটি প্রতবিদেন হয়নি। প্রসেক্রপিশন **467/468** ক্ষত্রেে সংশো ধন করা, যদওি লখেকরো স্পষ্ট বলনে **100% নর্ভিরযো গ্যতা নয়**।

কন্তিু সাফল্যরে আকৃতি গুরুত্বপূর্ণ। Science মাধ্যম কন্দ্রে বশিষেজ্ঞ commentary-তে Dr Wei Xing note করনে শরিো নাম সুবধিা মূলত **স্পষ্ট নর্ণিায়ক পরীক্ষা ফল**-যুক্ত শর্ত—অ্যাপন্ডেসিাইটসি/প্যানক্রয়িাটাইটসি—এ বশেি। নডিমো নয়িা/UTI-তে AI ও চকিৎিসক

উভয়ই সবচয়ে খারাপ এবং ব্যবধান ছোট।

তথ্য অসমতা: MIRA নয়িমতি-সবো উপলভ্য পরীক্ষাগার বশিল্ষেয়রে প্রায় **51%** ব্যবহার করছে; বোর্ড-সনদপ্রাপ্ত চকিৎিসক প্রায় **28%**—মধ্যমা ক্ষত্রেে সাতটি বশেিরক্ত প্যারামটিার। লখেকরো বলনে এটি ডটোসটে নয়িমতি প্রারম্ভকি মানরে নচিে এবং ব্যয়বহুল আন্তঃ-sectional ইমজেংি ধারাবাহকিভাবে বাড়ায়নি। তবু বশেি প্রমাণ নজিইে রোগনর্িণয়মূলক নর্িভুলতা বাড়াতে পারে; সমান তথ্যপ্রাপ্ত contest নয়।

"চকিৎিসককে হারয়িছে" মানে "ভালো চকিৎিসক" নয় কনে

1. স্কোর কীসরে বর্িদ্ধ। Julie Jacko note করনে কছিু চকিৎিসা-সামঞ্জস্য মট্রেকি ঐতহিাসকি চার্টে নথবিদ্ধ আচরণ পুনর্ুত্পাদন করাকে পুরস্কার কর।ে ঐতহিাসকি নথি উত্তর মূলরে অংশ। রোগনর্িণয়মূলক নর্িভুলতা অবশ্য ছাড়পত্রকালীন রোগনর্িণয়রে বর্িদ্ধ, তাই বশিুদ্ধ নথবিদ্ধকরণ echo-এর চয়েে শক্তশিালী; তবু চকিৎিসা মানরে পরম রফোরন্েস সত্য নয়।

2. তথ্য অসমতা। 51% বনাম 28% উপলভ্য পরীক্ষাগার বশিল্ষেয—MIRA বশেি তথ্য কনিছে/ব্যবহার করছে।ে খরচ, অস্বস্তি, বলিম্ব ববিচেনা না করে চকিৎিসক-ও বশেি কম খরচরে পরীক্ষা পলেে নর্িভুলতা বাড়াতে পার।ে

3. পরবিশে। পশ্চাৎদৃষ্টমিূলক স্যান্ডবক্স, আটটি নর্িবাচতি রোগনর্িণয়, **শুধু পাঠ্যভত্তিকি।** বাস্তব মূল্যায়নে ভৌত পরীক্ষা, tone, আচরণ, দহে ভাষা, ববির্তনশীল উপসর্গ আছ।ে আটটি রোগনর্িণয়রে বাইরে অস্পষ্ট চকিৎিসাবজ্ঞিান নয়িে গবষেণা কথা বলে না।

4. সম্ভাব্য প্রশক্ষিণ-তথ্য দূষণ। যে কসেগুলো অনুকরণ করা হয়ছেে সগেুলো জনসাধারণরে জন্য উন্মুক্ত এবং বহুল আলোচতি। LLM প্রশক্ষিণরে সময় গবষেণাপত্র, কসে বা সগেুলো থকেে তরৈি উপাদান দখেে থাকতে পার।ে Dr Midhun Parakkal Unni এই উদ্বগেটি তুলছেনে। লখেকরোও সতর্কভাবে বলনে, ফলটি **সম্ভাব্য ঊর্ধ্বসীমা** হতে পারে এবং অন্য প্রকাশ্য কসেে সাধারণীকরণরে ক্ষমতাকে বশেি দখোতে পার।ে স্বাধীন ব্যক্তগিত বা ভবষ্যিতমুখী পুনরাবৃত্তি ছাড়া মুখস্থ করা আর প্রকৃত যুক্তি আলাদা করা কঠনি।

এই সতর্কতা সক্ষমতাকে মুছে দয়ে না। সংশোধন করা পাঠ: পশ্চাৎদৃষ্টমিূলক মানদণ্ডে পুরো-নথি এজন্টে চকিৎিসকরে ওপরে স্কোর করছে,ে বশিষে করে পরষ্কিার-পরীক্ষা রোগ; বশেি পরীক্ষাগার info ব্যবহার করছে এবং ঐতহিাসকি চার্টরে সঙ্গে আংশকি সামঞ্জস্য পয়েছে।ে এটি সক্ষমতা প্রদর্শন, যন্ত্র ভালো চকিৎিসক—রায় নয়।

এই গবষেণা কী প্রমাণ করে না

- **বাস্তব রোগীতে কাজ করে**—না; সব ক্ষত্রেে পশ্চাৎদৃষ্টমিূলক/সমিুলটেডে।
- **আটটি রোগনর্িণয়রে বাইরে সাধারণীকরণ করে**—না।
- **Deploy নরিাপদে করা যায়**—না; ভবষ্যিতমুখী নরিাপত্তা/শাসনব্যবস্থা দরকার।
- ন্যায্য সমান তথ্যপ্রাপ্ত চকিৎিসক head-to-head—না; MIRA বশেি রক্ত বশিল্ষেয ব্যবহার করছে এবং চকিৎিসা মট্রেকি চার্টরে বর্িদ্ধে আংশকি স্কোর করছে।
- মুখস্থকরণ-মুক্ত—না; জনসাধারণরে অনুকরণ করা-IV প্রশক্ষিণ ওভারল্যাপ exclude হয়নি।
- **AI replaces চকিৎিসক**—না। বাস্তব ব্যবহার চকিৎিসক partnership/কর্তৃত্ব এবং পাঠ্য নথরি বাইরে মানব পর্যবক্ষেণ দরকার।

প্রমাণ কতটা শক্ত?

দুটি দাবি আলাদা।

সক্ষমতা প্রদর্শন—ভাষা-মডলে এজন্টে স্যান্ডবক্সে সীমাবদ্ধ EHR-এ ইতহিাস→পরীক্ষাগুলো→রোগনির্ণয়→নর্দিশে পুরো লুপ চালাতে পারে—নতুন ও যথষ্টে বশ্বিাসযোগ্য। এটইি সত্যইি নতুন অংশ।

শ্রষ্ঠেত্বরে দাবি—MIRA চকিৎিসকদরে চয়েে ভালো রোগনির্ণয়কারী—খুব সীমতিভাবে পড়তে হবে। এটি নর্দিষ্টি retrospective benchmark, মাত্র আটটি diagnosis, অসম তথ্যপ্রাপ্তি, ঐতহিাসকি উত্তরকে gold standard ধরা এবং সম্ভাব্য training-data contamination-এর মধ্যে মাপা ফল। তাই "এই benchmark-এ impressive performance" বলা যায়; "মানুষরে চয়েে ভালো চকিৎিসক" বলা যায় না—লখেকরোও দ্বতিীয় দাবি করনে না।

সঠকি অবস্থানটি "AI চকিৎিসকদরে হারয়িে দয়িছে" কংিবা "এটি খলেনা"—কোনোটইি নয়। নথরি ভতেরে বাস্তব ক্রয়িা নতিে পারে এমন চকিৎিসাবষিয়ক এজন্টেটি কঠনি পশ্চাৎদৃষ্টমিূলক মানদণ্ডে ভালো করছে; এখন দরকার বাস্তব, অস্পষ্ট রোগীর ক্ষত্রেে ভবষি্যতমুখী এবং যথাযথ তত্ত্বাবধান-সহ পরীক্ষা।

কনে এটি গুরুত্বপূর্ণ

চকিৎিসাবষিয়ক AI-এর আকর্ষণীয় প্রশ্ন বদলাচ্ছে। আগে "মডলে রোগনির্ণয় ঠকি বলতে পার?"—এখন আরও কঠনি প্রশ্ন "**কর্মপ্রবাহ চালাতে পার?** ইতহিাস নওেয়া, পরীক্ষা নর্দিশে দওেয়া, ফল ব্যাখ্যা করা, সদ্ধিান্ত, ক্রয়িা?" MIRA এই সরণ দখোয়। Acting-এ উপযোগতিা এবং ঝুঁকি দুটোই বাড়ে।

শরিোনাম "AI চকিৎিসকদরে ছাড়য়িে যায়" সর্বনম্নি নতুন/সর্বনম্নি সমর্থতি অংশ। সত্যইি গুরুত্বপূর্ণ অংশ: **এজন্টে পুরো নথকি র্মপ্রবাহে পদক্ষপে করতে পার**। সক্ষমতা টকিে থাকা করলে প্রথমে কে চকিৎিসক থাকবে তা নয়, কর্মপ্রবাহ—ক্রম, ফলাফল chase, ব্যাখ্যা—বদলাবে।

Leaderboard-এ ভালো ফল ভবষি্যতমুখী clinical testing-এর বকিল্প নয়। রোগী-স্তররে বাস্তব পরবিশেে prospective গবষেণাই পরবর্তী বড় প্রমাণরে ধাপ।

সংক্ষপ্তি সারাংশ

MIRA স্যান্ডবক্সে সীমাবদ্ধ EHR-এ স্বায়ত্তশাসতি এজন্টে: ইতহিাস, পরীক্ষাগার/ইমজেংি/মাইক্রোবায়োলজি নর্দিশে দওেয়া/ব্যাখ্যা করা, রোগনির্ণয় এবং চকিৎিসা নর্দিশে দওেয়া শুরু থকেে শষে পর্যন্ত। জনসাধারণরে অনুকরণ করা-IV-এর **574 পশ্চাৎদৃষ্টমিূলক ক্ষত্রে, 8 রোগনির্ণয়**-এ সামগ্রকি রোগনির্ণয়মূলক নর্ভুলতা **88.9%**। যথোথ **311-ক্ষত্রে** চকিৎিসক তুলনায় MIRA **87.8%**, জ্যষ্ঠে বোর্ড-সনদপ্রাপ্ত দল **78.1%**, কনষ্ঠি-ভারী দল **71.1%**। ওষুধ সদ্ধিান্ত মূলত নরিাপদ/guideline-concordant, **467/468** প্রসেক্রপিশন সংশোধন করা, উচ্চ-তীব্রতা ওষুধ ত্রুটি প্রতবিদেন নইে। কন্তিু স্যান্ডবক্স/শুধু পাঠ্যভত্তিকি, নর্বাচতি রোগনির্ণয়, স্পষ্ট-পরীক্ষা শর্তে প্রান্ত বড়, MIRA **~51% বনাম চকিৎিসক ~28%** উপলভ্য পরীক্ষাগার বশ্লিষ্যে ব্যবহার করছে, কছিু ফলাফল ঐতহিাসকি চার্টরে বরিুদ্ধে স্কোর করছে, এবং জনসাধারণরে অনুকরণ করা-IV দূষণ ঝুঁকি আছে। বাস্তব অগ্রগতি হলো পুরো-কর্মপ্রবাহ actor; বাস্তব-ক্লনিকি শ্রষ্ঠেত্ব/ব্যবহারকি মোতায়নে প্রমাণ নয়।

সরাসরি যাচাই

পপোর যা দখোয়: ভাষা-মডলে এজন্টে স্যান্ডবক্সে সীমাবদ্ধ EHR-এ পুরো ক্ষত্রে চালাতে পারে এবং নর্দিষ্টি পশ্চাৎদৃষ্টমিূলক মানদণ্ডে উচ্চ রোগনর্ণিয়মূলক স্কোর করছে।

যা সম্ভব, কন্তিু প্রমাণতি নয়: বাস্তব রোগী উপকার; সুবধিা বশিুদ্ধ ববিচেনামূলক চন্তিার কারণে; ব্যক্তগিত অদখো তথ্যওে একই কার্যসম্পাদন।

যা দখোয় না: বস্তিৃত চকিৎিসাবজ্ঞিান সাধারণীকরণ, ব্যবহারকি মোতায়নে নরিাপত্তা, ন্যায্য সমান তথ্যপ্রাপ্ত চকিৎিসক শ্রষ্ঠেত্ব, চকিৎিসক প্রতস্থিাপন, দূষণ-মুক্ত ফল।

প্রধান সীমাবদ্ধতা: পশ্চাৎদৃষ্টমিূলক সমিুলশেন, আটটি রোগনর্ণিয়, শুধু পাঠ্যভত্তিকি, আরও পরীক্ষাগার পরীক্ষাগুলো, চার্ট-ভত্তিকি স্কোরংি elements, জনসাধারণরে ডটোসটে দূষণ সম্ভাবনা।

সাধারণ পাঠকরে কতটা আস্থা রাখা উচতি? স্যান্ডবক্সে শুরু থকেে শষে পর্যন্ত কাজ সম্পাদনরে সক্ষমতায়—উচ্চ। "AI চকিৎিসকদরে চয়েে ভালো রোগনর্ণিয় কর" এই বৃহত্তর দাবতিে—কম; ফলটি benchmark-এর সীমানার মধ্য।

উৎস

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