The Clean Paper

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

Agen AI menangani seluruh kasus pasien sendiri — dalam simulator, pada rekam medis lama

MIRA adalah jenis baru AI medis: alih-alih menjawab satu pertanyaan, ia menangani seluruh kasus di dalam rekam medis rumah sakit yang disimulasikan — mengambil riwayat, memesan dan membaca tes, mencapai diagnosis, lalu menulis perintah. Pada 574 kasus retrospektif dari basis data publik, mencakup delapan diagnosis yang telah dipilih, para penulis melaporkan MIRA mengungguli dokter dalam akurasi diagnosis dan membuat keputusan yang sebagian besar selaras dengan pedoman serta aman dari sisi obat. Setiap kualifikasi dalam kalimat itu penting: ia berjalan di sandbox pada rekam medis lama, hanya lewat teks; sebagian besar keunggulannya datang dari kondisi dengan hasil tes yang tegas; ia memesan sekitar dua kali lebih banyak tes darah daripada dokter; dan beberapa outcome dinilai terhadap apa yang tercatat di chart asli. Kemajuan yang sebenarnya adalah agen yang bertindak di seluruh alur kerja — bukan bukti bahwa mesin kini mendiagnosis lebih baik daripada dokter. Para penulis mengatakannya sendiri: generalisasi, keamanan, dan tata kelola masih membutuhkan studi prospektif di dunia nyata.

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

Menjawab pertanyaan tidak sama dengan menangani seluruh kasus

Sebagian besar kisah tentang AI medis berbicara tentang model yang *menjawab*. Kita memberinya vignette atau soal ujian, ia mengembalikan diagnosis atau paragraf saran, lalu mendapat nilai tinggi. Berguna, tetapi sempit: seperti konsultan pintar yang tidak pernah menyentuh rekam medis.

MIRA dibangun untuk melakukan sesuatu yang berbeda, dan perbedaan itulah berita sebenarnya. Alih-alih menjawab satu pertanyaan, ia menangani seluruh kasus: membaca rekam pasien, memutuskan riwayat apa yang masih dibutuhkan, memesan pemeriksaan laboratorium, pencitraan, dan mikrobiologi, membaca hasilnya, mempersempit diagnosis banding, lalu menulis perintah berikutnya — resep, rawat inap, rujukan operasi. Semua itu dilakukan dengan mengambil tindakan *di dalam* rekam medis elektronik, seperti klinisi, bukan sekadar menghasilkan teks bebas yang harus disalin manusia.

Itu memang satu langkah nyata: dari sistem yang memberi saran menjadi sistem yang bertindak di sepanjang alur kerja. Dan justru langkah seperti ini yang sering diratakan dalam pemberitaan menjadi “AI mengungguli dokter”. Karena itu kita perlu sangat tepat mengenai apa yang diuji, dan di mana.

Versi satu kalimatnya: MIRA menangani seluruh kasus secara mandiri dan, pada tolok ukur ini, mengungguli dokter dalam akurasi diagnosis — tetapi ia melakukannya di sandbox, pada rekam medis retrospektif, untuk delapan diagnosis yang sudah dipilih sebelumnya, hanya melalui teks, dan sebagian besar keunggulannya muncul pada kondisi dengan hasil tes yang paling tegas. Kemajuannya nyata. Papan skor bukanlah 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.

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 menunjukkan kemampuan alur kerja end-to-end di dalam EHR sandbox. Studi ini tidak menunjukkan penerapan klinis dunia nyata, cakupan diagnosis yang luas, atau perbandingan adil dengan dokter yang mendapat informasi setara.

Original Aurora diagram — The Clean Paper CC BY 4.0

Apa yang dilakukan para penulis

MIRA — Medical Intelligence for Reasoning and Action — adalah agen otonom yang dibangun di atas model OpenAI: bagian yang berkomunikasi dan bertindak berjalan pada **GPT-4o**, sedangkan tahap perencanaan terpisah menggunakan model penalaran **o1** milik OpenAI. Keduanya berada dalam kerangka khusus yang mematuhi standar — dibangun di atas HL7 FHIR, standar data yang dipakai rumah sakit nyata untuk memindahkan rekam medis — dengan kotak alat berisi lebih dari delapan puluh ribu kemungkinan tindakan klinis. Di dalam sandbox itu MIRA dapat mengambil riwayat pasien, memesan dan menafsirkan pemeriksaan laboratorium, pencitraan, serta mikrobiologi, membuat diagnosis banding, dan menyusun rencana terapi — meresepkan obat, menjadwalkan prosedur bedah, merencanakan rawat inap. Satu detail patut disebut sejak awal: "juri" otomatis yang menilai jawaban MIRA juga menggunakan GPT-4o — keluarga model yang sama dengan yang sedang diuji — lalu penilaian itu dilengkapi tinjauan dokter bersertifikat setelah seorang reviewer mengangkat kekhawatiran tersebut.

Set pengujian berasal dari **MIMIC-IV**, basis data EHR besar, publik, dan telah dideidentifikasi. Dari sekitar 300.000 pasien yang dirawat di Beth Israel Deaconess Medical Center antara 2008 dan 2019, para penulis memilih **574 kasus pasien** yang mencakup **delapan diagnosis target** — patologi abdomen dan kegawatdaruratan penyakit dalam, termasuk apendisitis, pankreatitis, pneumonia, dan infeksi saluran kemih. Untuk setiap kasus, MIRA memulai dari gambaran terbatas lalu harus memutuskan, langkah demi langkah, apa yang dilakukan berikutnya — bentuknya menyerupai pemeriksaan klinis nyata, tetapi dijalankan terhadap rekam medis yang akhir sebenarnya sudah tercatat.

Kinerjanya kemudian dibandingkan dengan dokter pada kasus yang sama.

Apa sebenarnya arti "agen otonom dalam EHR sandbox"

Tiga kata di sini menanggung banyak makna, jadi perlu dibongkar satu per satu.
Agen berarti sistem bukan chatbot yang sekadar menjawab prompt. Ia menjalankan sebuah loop: lihat keadaan saat ini, pilih tindakan (pesan tes ini, tanyakan riwayat itu), lihat hasilnya, pilih tindakan berikutnya — hingga mencapai diagnosis dan rencana. "Kecerdasan"-nya adalah language model; *agency*-nya berasal dari kerangka di sekeliling model yang mengubah teks model menjadi operasi EHR yang diizinkan dan mengembalikan hasilnya sebagai masukan.
EHR sandbox berarti salinan simulasi yang terisolasi dari sistem rekam medis, bukan sistem rumah sakit yang aktif. MIRA dapat "memesan" tes dan menerima hasil yang benar-benar diperoleh pasien tersebut, karena kasusnya historis dan jawabannya sudah tercatat. Tidak ada tindakannya yang menyentuh pasien nyata atau klinik nyata.
Otonom berarti ia menyelesaikan kasus dari awal sampai akhir tanpa manusia di dalam loop

> — *di dalam sandbox*. Ini **tidak** berarti ia dilepas untuk menangani pasien tanpa pengawasan. Keduanya adalah klaim yang sangat berbeda, dan hanya yang pertama yang diuji.
> Satu hal lagi mengenai sandbox, karena mudah dibayangkan keliru: MIRA boleh *memesan* tes apa pun yang diinginkannya, tetapi sistem hanya dapat memberikan hasil jika tes tersebut **benar-benar dilakukan** pada pasien itu di dunia nyata. Jika ia meminta panel darah yang memang diterima pasien, ia mendapat nilai sebenarnya; jika ia meminta scan yang tidak pernah dilakukan, ia mendapat "N/A — could not be performed", tidak pernah angka yang dibuat-buat. Riwayat bekerja dengan cara yang sama: jika rekam medis tidak memuat jawaban atas pertanyaan MIRA, pasien simulasi sekadar mengatakan tidak tahu. Jadi MIRA tidak dapat memunculkan tes penentu yang tidak pernah dilakukan — ia hanya dapat bekerja dengan apa yang kebetulan tersedia dalam pemeriksaan nyata.
> Perbedaan ini penting karena kata pada judul — "otonom" — paling mudah dibaca sebagai hal kedua tadi.

Apa yang mereka temukan

Pada tolok ukur, **MIRA mengungguli dokter dalam akurasi diagnosis** — tetapi judul itu menyembunyikan tiga angka berbeda yang mudah tercampur, jadi sebaiknya dipisahkan.

Secara mandiri, pada seluruh **574 kasus**, MIRA menyebut diagnosis yang benar **88,9%** dari waktu — dinilai terhadap diagnosis saat pasien pulang yang benar-benar tercatat untuk tiap pasien dalam MIMIC-IV. Dalam perbandingan langsung dengan manusia, dokter tidak menangani seluruh 574 kasus; mereka menangani subset bersama **311 kasus**, menggunakan rekam dan alat yang sama dengan MIRA. Pada 311 kasus yang sama, MIRA mendapat **87,8%**, dan dibandingkan dengan dua kelompok dokter yang direkrut terpisah. Kelompok pertama adalah **kelompok senior**: empat dokter bersertifikat dengan pengalaman 7 hingga 11 tahun, yang mendapat **78,1%**. Kelompok kedua adalah **kelompok dengan senioritas campuran**: terutama residen junior ditambah dua spesialis, lebih mirip komposisi staf unit gawat darurat nyata, yang mendapat **71,1%**. Kasus sama, alat sama — urutannya AI pertama, dokter berpengalaman kedua, tim yang didominasi junior ketiga. Formulasi hati-hati dalam makalah adalah bahwa MIRA "secara konsisten setara, dan sering melampaui" para dokter pada kedelapan penyakit.

Keputusan setelah diagnosis juga cukup baik: sebagian besar selaras dengan pedoman, aman dari sisi pengobatan, dan tepat dalam keputusan rawat inap. Para penulis melaporkan tidak ada kesalahan obat berkeparahan tinggi pada lima kategori keamanan — interaksi obat, penyesuaian dosis ginjal, alergi, risiko QT, dan peresepan opioid — serta resep benar pada 467 dari 468 kasus, sambil tetap mencatat bahwa sistem "tidak mencapai keandalan 100%". Jika dibaca apa adanya, ini hasil mencolok: agen menjalankan seluruh kasus dan berakhir di atas dokter untuk diagnosis.

Namun *bentuk* kemenangan itu sama pentingnya dengan kemenangan itu sendiri. Para spesialis independen yang membaca makalah menunjukkan dari mana sebagian besar keunggulan muncul. Dalam panel ahli Science Media Centre, Dr Wei Xing (University of Sheffield) mencatat bahwa angka utama — AI mengalahkan dokter pada akurasi diagnosis — **terutama didorong oleh kondisi dengan**

hasil tes yang jelas, seperti apendisitis dan pankreatitis, ketika scan atau nilai laboratorium yang menentukan dapat menyelesaikan pertanyaan. Untuk pneumonia dan infeksi saluran kemih — dua alasan yang sangat umum bagi orang datang ke unit gawat darurat — *baik* AI maupun dokter berkinerja paling buruk, dan jarak di antara mereka paling kecil.

Ada juga ketidakseimbangan dalam cara kedua pihak "bermain", dan angkanya ada di makalah sendiri. MIRA jauh lebih banyak memakai laboratorium: ia menggunakan sekitar **51%** analit laboratorium yang tersedia dalam perawatan rutin, dibanding sekitar **28%** untuk dokter bersertifikat — median tujuh parameter darah tambahan per kasus. Para penulis berhati-hati menjelaskan bahwa angka ini masih *di bawah* baseline perawatan rutin dalam kumpulan data dan bukan strategi "pesan semuanya", serta tidak menemukan peningkatan sistematis pada pencitraan potong lintang yang lebih mahal. Tetap saja, lebih banyak informasi dapat dengan sendirinya meningkatkan akurasi diagnosis — jadi ini bukan pertandingan yang sepenuhnya sebanding antara pemain dengan informasi setara, sebuah kesenjangan yang juga disorot Xing. Seorang klinisi yang bebas memesan semua tes darah murah tanpa mempertimbangkan biaya, ketidaknyamanan, atau keterlambatan juga akan tampak lebih tajam di atas kertas.

Mengapa "mengalahkan dokter" bukan berarti "dokter yang lebih baik"

Kemenangan tolok ukur mudah dibaca terlalu jauh. Ada tiga hal di antara "skor MIRA lebih tinggi" dan "MIRA lebih baik".

Pertama, **apa yang menjadi kunci penilaiannya.** Profesor Julie Jacko (University of Edinburgh) mencatat bahwa beberapa outcome utama MIRA didefinisikan relatif terhadap apa yang terdokumentasi dalam kumpulan data — artinya sistem diberi nilai karena **mereproduksi perilaku klinis yang tercatat**, bukan selalu karena menunjukkan perawatan optimal. Rekam historis menjadi kunci jawaban. Itu cara yang masuk akal untuk membangun tolok ukur, tetapi mengukur kesesuaian dengan apa yang dilakukan, bukan kebenaran absolut. Secara adil, masalah ini paling kuat mengenai metrik *kesesuaian terapi* — apakah perintah MIRA cocok dengan chart? — sedangkan akurasi diagnosis utama dinilai terhadap diagnosis saat pulang, yang lebih dekat ke outcome nyata daripada sekadar gema dokumentasi.

Kedua, **ketidakseimbangan informasi** yang sudah disebut: jauh lebih banyak tes darah — sekitar 51% analit yang tersedia versus 28% — berarti lebih banyak bukti. Sebagian selisih akurasi kemungkinan sedang *dibeli*, bukan semata-mata dihasilkan oleh penalaran.

Ketiga, **di mana sistem dijalankan.** Ini sandbox, pada kasus retrospektif, dengan delapan diagnosis yang telah dipilih, dan hanya lewat teks. Penilaian klinis nyata, seperti dikatakan Dr Dominic Oliver (University of Oxford), bergantung bukan hanya pada apa yang pasien katakan tetapi juga bagaimana mereka mengatakannya — ditambah pemeriksaan fisik, perilaku yang diamati, dan bahasa tubuh, yang sama sekali tidak dilihat agen teks yang membaca rekam medis jadi. Dan pasien dengan masalah di luar delapan diagnosis yang dipilih berada di luar apa yang dapat dijelaskan studi ini.

Ada pula kekhawatiran lebih tenang yang diangkat reviewer: **kontaminasi data.** MIMIC-IV bersifat publik dan banyak dibahas, sehingga language model yang dilatih pada internet terbuka mungkin sudah melihat makalah, diskusi kasus, atau bahkan datanya. Dr Midhun Parakkal Unni (University of Sheffield) mengangkat hal ini secara langsung — jika sebagian jawaban sudah ada dalam data pelatihan, sebagian kinerja mungkin berasal dari ingatan, bukan penalaran klinis, dan hanya replikasi independen yang dapat memisahkan keduanya. Menariknya, para penulis tidak menepis masalah itu: mereka menulis bahwa hasil mereka "dapat secara hati-hati ditafsirkan sebagai kemungkinan batas atas" dan "mungkin melebihestimasi generalisasi ke kasus publik lain" — sebuah makalah yang memasang plafon pada judulnya sendiri.

Semua itu tidak membuat MIRA kurang menarik. Justru itu menentukan pembacaan yang terkalibrasi: pada tolok ukur retrospektif, agen yang mampu bertindak di seluruh rekam medis mengungguli dokter pada diagnosis dengan tes yang bersih — sebagian dengan memesan lebih banyak tes, sebagian dengan menyamai chart historis. Ini demonstrasi kemampuan yang nyata, bukan vonis bahwa mesin kini mendiagnosis lebih baik daripada dokter.

Apa yang *tidak* dibuktikan

- Studi ini **tidak** menunjukkan MIRA bekerja pada pasien nyata. Semua kasus retrospektif dan disimulasikan; tidak ada pasien yang pernah ditangani olehnya.
- Studi ini **tidak** menunjukkan sistem bekerja di luar delapan diagnosis. Kondisi di luar set yang telah dipilih — mayoritas kedokteran yang berantakan dan belum terbedakan — tidak diuji.
- Studi ini **tidak** menunjukkan sistem aman untuk diterapkan. Para penulis sendiri menulis bahwa generalisasi, keamanan, dan tata kelola masih membutuhkan studi prospektif di dunia nyata.
- Studi ini **tidak** menghasilkan perbandingan langsung yang sepenuhnya adil dengan dokter. MIRA memakai jauh lebih banyak tes darah (sekitar 51% analit yang tersedia versus 28%), dan beberapa outcome terapi dinilai sebagian terhadap chart yang tercatat, bukan terhadap standar perawatan terbaik yang benar-benar independen.
- Studi ini **tidak** menunjukkan hasil bebas dari memorisasi. Data publik MIMIC-IV mungkin bertumpang tindih dengan data pelatihan model, yang perlu disingkirkan melalui replikasi independen.
- Studi ini **tidak** berarti "AI menggantikan dokter". Ini pendukung keputusan yang dapat *bertindak* di seluruh rekam medis; konsensus reviewer adalah penggunaan nyata akan bermitra dengan klinisi, yang tetap memegang otoritas dan menyediakan semua hal yang hilang dari rekam teks.

Seberapa kuat buktinya?

Pisahkan klaim menjadi dua, karena bukti bagi masing-masing sangat berbeda.

Sebagai demonstrasi kemampuan — bahwa agen language model dapat menjalankan seluruh kasus klinis dalam EHR sandbox, merangkai riwayat, tes, diagnosis, dan perintah dari awal sampai akhir — penelitian ini benar-benar baru dan cukup meyakinkan. Inilah bagian yang baru, dan bagian yang

layak diperhatikan.

Sebagai klaim superioritas — bahwa MIRA *lebih baik daripada dokter* — buktinya terbatas dan harus dibaca hati-hati. Klaim itu berlaku pada tolok ukur retrospektif tertentu, untuk delapan diagnosis, dengan ketidakseimbangan informasi (lebih banyak tes), kunci jawaban yang bersumber dari chart historis, serta kemungkinan kontaminasi data pelatihan yang nyata. Itu cukup untuk mengatakan "agen berkinerja mengesankan pada tolok ukur ini". Tidak cukup untuk mengatakan "agen adalah diagnostikus yang lebih baik daripada dokter", dan para penulis juga tidak membuat klaim kedua itu.

Posisi yang paling berguna bukan "AI mengalahkan dokter" dan bukan "cuma mainan". Yang lebih tepat: jenis AI medis baru — yang *bertindak* di seluruh rekam medis alih-alih sekadar menjawab pertanyaan — berkinerja baik pada tolok ukur retrospektif yang sulit, dan sekarang harus membuktikan dirinya di tempat yang belum pernah diuji: pasien nyata yang belum terbedakan, secara prospektif, di bawah tata kelola.

Mengapa ini penting

Selama beberapa tahun, pertanyaan menarik dalam AI medis diam-diam berubah. Dulu pertanyaannya "bisakah model mendapat diagnosis yang benar?" — dan model terus menjawab ya pada set pengujian yang semakin rapi. MIRA menandai pergeseran ke pertanyaan yang lebih sulit: "bisakah model *melakukan pekerjaannya* — mengumpulkan, memesan, menafsirkan, memutuskan, bertindak — sepanjang satu alur kerja?" Itu pertanyaan yang lebih berguna dan lebih jujur, karena tindakan adalah tempat kesulitan dan risiko sama-sama berada.

Itulah sebabnya pembingkaian penting. Refleks judul — *AI mengungguli dokter* — justru menunjuk bagian hasil yang paling tidak baru dan paling kurang didukung. Hal yang benar-benar baru lebih kecil tetapi lebih konsekuensial: agen yang dapat bergerak di seluruh rekam medis. Jika kemampuan itu bertahan, ia mengubah **alur kerja** jauh sebelum mengubah siapa yang memegang kendali. Dokter tidak digantikan; pemesanan tes, penafsiran hasil, mengejar hasil yang belum kembali — alur kerja itulah tempat alat seperti ini pertama kali masuk.

Dan itu mengatur ulang beban pembuktian ke arah yang benar. Menang di papan peringkat kasus retrospektif adalah alasan untuk menjalankan uji prospektif, bukan pengganti uji tersebut. Para penulis mengatakan hal yang sama. Pembacaan MIRA yang jujur adalah undangan untuk studi berikutnya — dengan standar yang sudah diketahui bidang ini: ukur pada pasien, bukan tolok ukur.

Singkatnya

MIRA adalah agen AI otonom yang beroperasi pada rekam medis elektronik sandbox: ia dapat mengambil riwayat, memesan dan menafsirkan laboratorium, pencitraan, dan mikrobiologi, mencapai diagnosis, serta menulis rencana terapi. Diuji pada 574 kasus retrospektif dari basis data

publik MIMIC-IV, mencakup delapan diagnosis yang telah dipilih, MIRA mengungguli dokter pada akurasi diagnosis (88,9% secara keseluruhan; 87,8% versus 78,1% dalam perbandingan langsung) dan membuat keputusan yang sebagian besar selaras dengan pedoman dan aman dari sisi obat. Namun evaluasinya merupakan simulasi pada rekam medis lama, hanya dalam teks; sebagian besar keunggulan datang dari kondisi dengan hasil tes yang tegas; MIRA menggunakan jauh lebih banyak tes darah daripada dokter (sekitar 51% analit yang tersedia versus 28%); beberapa outcome terapi dinilai terhadap apa yang dicatat dalam chart asli; dan kumpulan data publik menimbulkan risiko nyata kontaminasi data pelatihan — yang oleh penulis sendiri disebut kemungkinan batas atas angka kinerja. Kemajuan yang sebenarnya adalah agen yang bertindak di seluruh alur kerja, bukan sekadar menjawab pertanyaan terpisah. Para penulis secara eksplisit menyatakan generalisasi, keamanan, dan tata kelola masih memerlukan studi prospektif dunia nyata — pembacaan yang benar: demonstrasi kemampuan yang mengesankan, bukan bukti bahwa AI mendiagnosis lebih baik daripada dokter, dan bukan sistem yang siap untuk klinik nyata.

Penilaian tanpa berlebihan

Apa yang ditunjukkan makalah: Agen language model (MIRA) dapat menjalankan seluruh kasus klinis dari awal sampai akhir di dalam EHR sandbox — riwayat, tes, diagnosis, perintah — dan pada tolok ukur retrospektif 574 kasus untuk delapan diagnosis, mengungguli dokter dalam akurasi diagnosis (88,9% keseluruhan; 87,8% versus 78,1% melawan dokter bersertifikat) tanpa kesalahan obat berkeparahan tinggi.

Apa yang masuk akal tetapi belum terbukti: Bahwa kemampuan ini menjadi manfaat bagi pasien nyata yang belum terbedakan; bahwa keunggulan akurasi berasal dari penalaran lebih baik, bukan lebih banyak tes yang dipesan, kesesuaian dengan chart historis, atau data publik yang diingat model.

Apa yang tidak ditunjukkan: Bahwa MIRA bekerja di luar delapan diagnosisnya; bahwa aman untuk diterapkan; bahwa ia mengalahkan dokter dalam perbandingan adil dengan informasi setara; bahwa ia menggantikan klinisi; bahwa hasil bebas kontaminasi data pelatihan.

Keterbatasan utama: Evaluasi retrospektif, simulasi, dan hanya teks; delapan diagnosis yang telah dipilih; ketidakseimbangan informasi (jauh lebih banyak tes darah — sekitar 51% analit yang tersedia versus 28%, terutama tes darah murah, bukan pencitraan); outcome terapi sebagian dinilai terhadap chart historis; serta tolok ukur publik MIMIC-IV yang mungkin pernah dilihat language model dalam pelatihan — sesuatu yang oleh penulis sendiri ditandai sebagai kemungkinan batas atas kinerja.

Seberapa besar keyakinan yang sebaiknya dimiliki pembaca umum? Tinggi bahwa MIRA merupakan demonstrasi kemampuan yang nyata dan baru — agen yang *bertindak* di seluruh rekam medis, bukan sekadar chatbot. Rendah bahwa ia adalah *diagnostikus yang lebih baik daripada dokter*: klaim itu terbatas pada satu tolok ukur retrospektif dan dikacaukan oleh pemesanan tes, penilaian terhadap chart, serta kemungkinan kontaminasi. Kesimpulan para penulis sendiri paling tepat — studi prospektif di dunia nyata diperlukan sebelum hasil ini berarti sesuatu bagi perawatan pasien.

Sumber

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