The Clean Paper

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

Satu ejen AI mengendalikan keseluruhan kes pesakit sendiri — dalam simulator, menggunakan rekod lama

MIRA ialah jenis baharu AI perubatan: bukan sekadar menjawab satu soalan, ia mengendalikan keseluruhan kes dalam rekod hospital simulasi — mengambil sejarah, memesan dan membaca ujian, mencapai diagnosis, lalu menulis arahan klinikal. Pada 574 kes retrospektif daripada pangkalan data awam, merentas lapan diagnosis yang dipilih terlebih dahulu, para penulis melaporkan bahawa ia mengatasi doktor dari segi ketepatan diagnosis dan membuat keputusan yang sebahagian besarnya selaras dengan garis panduan serta selamat dari segi ubat. Setiap kelayakan dalam ayat itu penting: ia berjalan dalam sandbox pada rekod lama, hanya melalui teks; sebahagian besar kelebihannya muncul pada keadaan dengan keputusan ujian yang jelas; ia menggunakan kira-kira dua kali lebih banyak ujian darah berbanding doktor; dan beberapa hasil dinilai berdasarkan apa yang direkodkan dalam carta asal. Kemajuan sebenar ialah ejen yang bertindak merentas seluruh aliran kerja — bukan bukti bahawa mesin kini mendiagnosis lebih baik daripada doktor. Para penulis sendiri menyatakannya: generalisasi, keselamatan dan tadbir urus masih memerlukan kajian prospektif dunia sebenar.

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 satu soalan tidak sama dengan mengendalikan keseluruhan kes

Kebanyakan cerita tentang AI perubatan berkisar pada model yang *menjawab*. Anda memberinya vignet atau soalan peperiksaan, ia memberi kembali diagnosis atau satu perenggan nasihat, lalu mendapat skor tinggi. Berguna, tetapi sempit: seperti perunding pintar yang tidak pernah menyentuh carta pesakit.

MIRA dibina untuk melakukan sesuatu yang berbeza, dan perbezaan itulah berita sebenarnya. Daripada menjawab satu soalan, ia mengendalikan keseluruhan kes: membaca rekod pesakit, menentukan sejarah tambahan yang masih diperlukan, memesan ujian makmal, pengimejan dan mikrobiologi, membaca hasilnya, mengecilkan diagnosis pembezaan, kemudian menulis arahan yang menyusul — preskripsi, kemasukan ke hospital, rujukan untuk pembedahan. Ia berbuat demikian dengan mengambil tindakan *di dalam* rekod kesihatan elektronik, seperti yang dilakukan oleh klinisian, bukannya sekadar menghasilkan teks bebas untuk disalin oleh manusia.

Itu ialah satu langkah sebenar: daripada sistem yang memberi nasihat kepada sistem yang bertindak merentas aliran kerja. Ia juga tepat jenis langkah yang mudah diratakan dalam liputan menjadi "AI mengatasi doktor." Jadi, penting untuk tepat tentang apa yang sebenarnya diuji, dan di mana.

Versi satu ayatnya: MIRA mengendalikan keseluruhan kes sendiri dan, pada penanda aras ini, mengatasi doktor dari segi ketepatan diagnosis — tetapi ia melakukannya dalam sandbox, menggunakan rekod retrospektif, merentas lapan diagnosis yang dipilih terlebih dahulu, berkomunikasi hanya melalui teks, dan sebahagian besar kelebihannya datang daripada keadaan dengan keputusan ujian paling jelas. Kemajuannya nyata. Papan skor itu bukan klinik sebenar.

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 keupayaan aliran kerja hujung-ke-hujung dalam EHR sandbox. Ia tidak menunjukkan penggunaan klinikal dunia sebenar, liputan diagnosis yang luas, atau perbandingan langsung yang adil dengan doktor yang mempunyai maklumat setara.

Original Aurora diagram — The Clean Paper CC BY 4.0

Apa yang dilakukan oleh para penulis

MIRA — Medical Intelligence for Reasoning and Action — ialah ejen autonomi yang dibina berasaskan model OpenAI: bahagian yang berbual dan bertindak menggunakan **GPT-4o**, manakala langkah perancangan berasingan menggunakan model penaakulan **o1** OpenAI. Semuanya berada dalam rangka kerja tersuai yang mematuhi piawaian — dibina berasaskan HL7 FHIR, piawaian data yang digunakan hospital sebenar untuk memindahkan rekod — dengan set alat lebih daripada lapan puluh ribu tindakan klinikal yang mungkin. Dalam sandbox itu, MIRA boleh mendapatkan sejarah pesakit, memesan dan mentafsir makmal, pengimejan dan mikrobiologi, menghasilkan diagnosis pembezaan, serta merumuskan pelan rawatan — menetapkan ubat, menjadualkan prosedur pembedahan dan merancang kemasukan ke hospital. Satu butiran yang wajar ditandai sejak awal: "hakim" automatik yang menilai jawapan MIRA sendiri menggunakan GPT-4o — keluarga model yang sama dengan yang sedang diuji — lalu para penulis menambah semakan oleh doktor bertauliah lembaga selepas seorang penilai rakan sebaya membangkitkan kebimbangan itu.

Set ujian datang daripada **MIMIC-IV**, pangkalan data EHR besar yang dinyahpengenalan dan tersedia secara awam. Daripada kira-kira 300,000 pesakit yang dirawat di Beth Israel Deaconess Medical Center antara 2008 dan 2019, para penulis memilih **574 kes pesakit** yang merangkumi **lapan diagnosis sasaran** — patologi abdomen dan kecemasan perubatan dalaman, termasuk apendisitis, pankreatitis, pneumonia dan jangkitan saluran kencing. Bagi setiap kes, MIRA bermula dengan gambaran terhad dan perlu memutuskan, langkah demi langkah, apa yang perlu dilakukan seterusnya — bentuk yang sama seperti siasatan klinikal sebenar, tetapi dijalankan terhadap rekod yang kesudahan sebenarnya sudah tersimpan.

Prestasinya kemudian dibandingkan dengan doktor pada kes yang sama.

Apa sebenarnya maksud "ejen autonomi dalam EHR sandbox"

Tiga perkataan di sini memikul banyak makna, jadi elok dihuraikan.

Ejen bermaksud sistem itu bukan chatbot yang sekadar menjawab arahan. Ia menjalankan satu gelung: melihat keadaan semasa, memilih tindakan (pesan ujian ini, tanya sejarah itu), melihat hasilnya, memilih tindakan seterusnya — sehingga mencapai diagnosis dan pelan. "Kecerdasan" datang daripada model bahasa; *keagenan* datang daripada struktur di sekelilingnya yang menukar teks model menjadi operasi EHR yang dibenarkan lalu memasukkan hasil operasi itu kembali kepada model.

EHR sandbox bermaksud salinan simulasi dan terasing bagi sistem rekod perubatan, bukan sistem hospital yang sedang beroperasi. MIRA boleh "memesan" ujian dan menerima keputusan yang benar-benar diperoleh pesakit tersebut kerana kes itu ialah kes sejarah dan

> jawapannya sudah direkodkan. Tiada apa yang dilakukan MIRA menyentuh pesakit sebenar atau klinik sebenar.
> **Autonomi** bermaksud ia melengkapkan kes dari awal hingga akhir tanpa manusia dalam gelung — *di dalam sandbox*. Ia **tidak** bermaksud sistem itu dilepaskan untuk mengurus pesakit tanpa pengawasan. Dua dakwaan itu sangat berbeza, dan hanya yang pertama diuji.
> Satu lagi perkara tentang sandbox, kerana mudah dibayangkan secara salah: MIRA boleh *memesan* apa sahaja ujian yang dikehendakinya, tetapi sistem hanya boleh memberikan keputusan jika ujian itu **benar-benar dilakukan** terhadap pesakit berkenaan dalam dunia sebenar. Minta panel darah yang memang diterima pesakit, dan sistem memberikan nilai sebenar; minta imbasan yang tidak pernah dipesan, dan jawapannya ialah "N/A — could not be performed," bukan nombor rekaan. Sejarah pesakit berfungsi dengan cara yang sama: jika rekod tidak mengandungi perkara yang ditanya MIRA, pesakit simulasi hanya mengatakan bahawa ia tidak tahu. Jadi MIRA tidak boleh memunculkan satu ujian penentu yang tidak pernah dibuat — ia hanya boleh bekerja dengan apa yang kebetulan termasuk dalam siasatan sebenar pesakit tersebut.
> Perbezaan ini penting kerana perkataan tajuk utama — "autonomi" — ialah yang paling mudah dibaca seolah-olah bermaksud perkara kedua.

Apa yang mereka temui

Pada penanda aras, **MIRA mengatasi doktor dari segi ketepatan diagnosis** — tetapi tajuk utama menyembunyikan tiga nombor berbeza yang mudah bercampur, jadi elok dipisahkan.

Secara sendiri, merentas semua **574 kes**, MIRA menamakan diagnosis yang betul **88.9%** daripada masa — dinilai berbanding diagnosis semasa discaj yang benar-benar direkodkan bagi setiap pesakit dalam MIMIC-IV. Bagi perbandingan langsung dengan manusia, doktor tidak mengendalikan semua 574 kes; mereka mengendalikan **subset 311 kes** yang sama, menggunakan rekod dan alat yang sama yang tersedia kepada MIRA. Pada 311 kes itu, MIRA mendapat **87.8%**, dan dibandingkan dengan dua kumpulan yang direkrut secara berasingan. Yang pertama ialah **kumpulan senior**: empat doktor bertauliah lembaga dengan pengalaman 7 hingga 11 tahun, yang mendapat **78.1%**. Yang kedua ialah **kumpulan campuran tahap senioriti**: kebanyakannya residen junior serta dua pakar, lebih hampir kepada bagaimana jabatan kecemasan sebenar biasanya dikendalikan, yang mendapat **71.1%**. Kes sama, alat sama — dan susunannya ialah AI pertama, doktor berpengalaman kedua, pasukan yang lebih banyak junior ketiga. Frasa berhati-hati dalam kertas itu sendiri ialah bahawa MIRA "secara konsisten setara dengan, dan sering mengatasi" doktor merentas kesemua lapan penyakit.

Keputusan susulannya juga bertahan: sebahagian besarnya selaras dengan garis panduan, selamat dari segi ubat, dan sesuai dalam keputusan kemasukan. Para penulis melaporkan tiada kesilapan ubat berkeparahan tinggi merentas lima kategori keselamatan (interaksi ubat, pelarasan dos renal, alahan, risiko QT dan preskripsi opioid), serta preskripsi yang betul dalam 467 daripada 468 kes — sambil menyatakan bahawa sistem itu "tidak mencapai kebolehpercayaan 100%." Jika diambil pada nilai muka, itu hasil yang mengagumkan: ejen yang mengendalikan seluruh kes dan mendapat ketepatan diagnosis lebih tinggi daripada doktor.

Namun *bentuk* kemenangan itu sama penting dengan kemenangan itu sendiri. Pakar bebas yang membaca kertas tersebut menunjukkan dari mana sebenarnya kelebihan itu datang. Dalam panel pakar Science Media Centre, Dr Wei Xing (University of Sheffield) menyatakan bahawa angka tajuk utama — AI mengatasi doktor dalam ketepatan diagnosis — **sebahagian besarnya didorong oleh keadaan yang mempunyai keputusan ujian jelas**, seperti apendisitis dan pankreatitis, apabila imbasan atau nilai makmal penentu menyelesaikan persoalan. Bagi pneumonia dan jangkitan saluran kencing — dua sebab paling lazim orang datang ke jabatan kecemasan — *kedua-dua* AI dan doktor mempunyai prestasi paling lemah, dan jurang antara mereka paling kecil.

Terdapat juga ketidakseimbangan dalam cara kedua-dua pihak "bermain", dan ia kelihatan dalam nombor kertas itu sendiri. MIRA lebih banyak bergantung pada makmal: ia menggunakan kira-kira **51%** analit makmal yang tersedia dalam penjagaan rutin, berbanding kira-kira **28%** bagi doktor bertauliah lembaga — median tujuh parameter darah tambahan bagi setiap kes. Para penulis berhati-hati menjelaskan bahawa ini masih *di bawah* garis dasar penjagaan rutin dalam set data itu sendiri, bukan strategi memesan segala-galanya, dan mereka melaporkan *tiada* peningkatan sistematik pada pengimejan keratan rentas yang lebih mahal. Namun, lebih banyak maklumat dengan sendirinya boleh menghasilkan ketepatan diagnosis lebih tinggi — jadi ini bukan pertandingan sepenuhnya setara antara pemain yang mempunyai maklumat sama, satu jurang yang ditandai Xing secara langsung. Seorang klinisian yang bebas memesan setiap ujian darah murah tanpa perlu menimbang kos, ketidakselesaan atau kelewatan juga akan kelihatan lebih tajam di atas kertas.

Mengapa "mengatasi doktor" tidak sama dengan "doktor yang lebih baik"

Kemenangan penanda aras mudah ditafsir terlalu jauh. Tiga perkara berdiri antara "MIRA mendapat skor lebih tinggi" dan "MIRA lebih baik."

Pertama, **apa yang menjadi rujukan penilaiannya.** Profesor Julie Jacko (University of Edinburgh) menyatakan bahawa beberapa hasil utama MIRA ditakrifkan relatif kepada apa yang didokumentasikan dalam set data asal — bermakna sistem diberi ganjaran kerana **menghasilkan semula tingkah laku klinikal yang direkodkan**, bukan semestinya kerana menunjukkan penjagaan optimum. Carta sejarah menjadi kunci jawapan. Itu cara yang munasabah untuk membina penanda aras, tetapi ia mengukur persetujuan dengan apa yang telah dilakukan, bukan ketepatan dalam erti mutlak. Untuk berlaku adil kepada kertas itu, masalah ini paling kuat pada metrik *penjajaran rawatan* — adakah arahan MIRA sepadan dengan carta? — manakala ketepatan diagnosis tajuk utama dinilai berbanding diagnosis semasa discaj, yang lebih hampir kepada hasil sebenar daripada sekadar gema dokumentasi.

Kedua, **ketidakseimbangan maklumat** yang disebut tadi: jauh lebih banyak ujian darah (kira-kira 51% analit tersedia, berbanding 28%) bermakna lebih banyak bukti. Sebahagian jurang ketepatan berkemungkinan *dibeli* dengan maklumat tambahan, bukan semata-mata dihasilkan melalui penaakulan yang lebih baik.

Ketiga, **di mana ia berjalan.** Ini ialah sandbox, pada kes retrospektif, merentas lapan diagnosis yang

dipilih terlebih dahulu, dan hanya melalui teks. Penilaian klinikal sebenar, seperti yang dinyatakan Dr Dominic Oliver (University of Oxford), bergantung bukan sahaja pada apa yang dikatakan pesakit tetapi juga bagaimana mereka mengatakannya — di samping pemeriksaan fizikal, tingkah laku yang diperhatikan dan bahasa tubuh, yang tidak pernah dilihat oleh ejen teks yang membaca rekod siap. Dan pesakit yang masalahnya tidak termasuk dalam salah satu daripada lapan diagnosis terpilih berada di luar apa yang boleh dibicarakan oleh kajian ini.

Ada juga kebimbangan lebih senyap yang dibangkitkan penilai: **pencemaran data**. MIMIC-IV ialah data awam dan banyak dibincangkan dalam tulisan, jadi model bahasa yang dilatih pada internet terbuka mungkin sudah melihat kertas, perbincangan kes, atau data itu sendiri. Dr Midhun Parakkal Unni (University of Sheffield) menandainya secara langsung — jika sebahagian jawapan berada dalam data latihan, sebahagian prestasi itu mungkin ingatan, bukannya penaakulan klinikal, dan hanya replikasi bebas boleh membezakan kedua-duanya. Yang penting, para penulis tidak menepis isu ini: mereka menulis bahawa hasil mereka "boleh ditafsir secara berhati-hati sebagai had atas yang mungkin" dan "mungkin melebih anggar generalisasi kepada kes awam lain" — sebuah kertas yang meletakkan siling pada tajuk utamanya sendiri.

Tiada satu pun daripada perkara ini menjadikan MIRA kurang menarik. Ia cuma menjadikan tafsiran yang betul lebih seimbang dan berhati-hati: pada penanda aras retrospektif, ejen yang boleh bertindak merentas keseluruhan rekod mengatasi doktor pada diagnosis dengan ujian yang jelas — sebahagiannya dengan memesan lebih banyak ujian, sebahagiannya dengan sepadan dengan carta yang direkodkan. Itu demonstrasi keupayaan sebenar, bukan keputusan bahawa mesin kini mendiagnosis lebih baik daripada doktor.

Apa yang *tidak* dibuktikan oleh kajian ini

- Ia **tidak** menunjukkan MIRA berfungsi pada pesakit sebenar. Setiap kes adalah retrospektif dan disimulasikan; tiada pesakit pernah benar-benar diurus olehnya.
- Ia **tidak** menunjukkan ia berfungsi di luar lapan diagnosis. Keadaan di luar set terpilih — majoriti perubatan yang rumit dan belum dibezakan — tidak diuji.
- Ia **tidak** menunjukkan bahawa ia selamat untuk digunakan. Para penulis sendiri menyatakan bahawa generalisasi, keselamatan dan tadbir urus masih memerlukan kajian prospektif dunia sebenar.
- Ia **tidak** mewujudkan perbandingan langsung yang adil dengan doktor. MIRA menggunakan jauh lebih banyak ujian darah (kira-kira 51% analit tersedia berbanding 28%), dan beberapa hasil rawatan dinilai sebahagiannya terhadap carta yang direkodkan, bukannya terhadap penjagaan terbaik yang diketahui benar.
- Ia **tidak** menunjukkan hasil itu bebas daripada penghafalan. Data MIMIC-IV yang terbuka mungkin bertindih dengan latihan model, sesuatu yang perlu ditolak melalui replikasi bebas.
- Ia **tidak** bermaksud "AI menggantikan doktor." Ia ialah sokongan keputusan yang boleh *bertindak* merentas rekod; konsensus penilai ialah penggunaan sebenar akan berlaku bersama klinisian, yang kekal memegang kuasa dan menyediakan segala maklumat yang tidak dapat ditangkap oleh rekod teks.

Sejauh mana kukuh buktinya?

Pisahkan dakwaan kepada dua bahagian, kerana bukti bagi setiap satunya sangat berbeza.

Sebagai demonstrasi keupayaan — bahawa ejen model bahasa boleh mengendalikan keseluruhan kes klinikal dalam sandbox EHR, merangkaikan sejarah, ujian, diagnosis dan arahan dari awal hingga akhir — kerja ini benar-benar baharu dan cukup meyakinkan. Inilah bahagian yang baharu, dan inilah bahagian yang patut diberi perhatian.

Sebagai dakwaan keunggulan — bahawa MIRA *lebih baik daripada doktor* — buktinya terhad dan perlu dibaca dengan berhati-hati. Ia berlaku pada satu penanda aras retrospektif khusus, pada lapan diagnosis, dengan ketidakseimbangan maklumat (lebih banyak ujian), kunci jawapan yang berasal daripada carta sejarah, dan kemungkinan nyata pencemaran data latihan. Itu cukup untuk mengatakan "ejen ini berprestasi mengagumkan pada penanda aras ini." Ia tidak cukup untuk mengatakan "ejen ini ialah pakar diagnosis yang lebih baik daripada doktor," dan para penulis tidak mendakwa perkara kedua.

Sikap paling berguna bukan "AI mengalahkan doktor" dan bukan juga "sekadar mainan." Yang lebih tepat ialah: jenis baharu AI perubatan — yang *bertindak* merentas rekod dan bukannya sekadar menjawab soalan — berprestasi baik pada penanda aras retrospektif yang sukar, dan kini perlu membuktikan dirinya di tempat yang belum pernah diuji: pada pesakit sebenar yang belum dibezakan, secara prospektif, di bawah tadbir urus.

Mengapa ia penting

Sejak beberapa tahun kebelakangan ini, soalan menarik dalam AI perubatan berubah secara senyap. Dahulu soalnya "bolehkah model mendapatkan diagnosis yang betul?" — dan model terus menjawab ya pada set ujian yang semakin bersih. MIRA menandakan peralihan kepada soalan yang lebih sukar: "bolehkah model *melakukan kerja itu* — mengumpul, memesan, mentafsir, membuat keputusan, bertindak — merentas seluruh aliran kerja?" Itu soalan yang lebih berguna dan lebih jujur, kerana tindakan ialah tempat kesukaran dan risiko sebenar berada.

Itulah sebab cara membingkai hasil ini penting. Refleks tajuk utama — *AI mengatasi doktor* — menunjuk kepada bahagian hasil yang paling kurang baharu dan paling kurang disokong. Perkara yang benar-benar baharu lebih kecil tetapi lebih besar akibatnya: ejen yang boleh bergerak melalui keseluruhan rekod. Jika keupayaan itu bertahan, ia mengubah **aliran kerja** jauh sebelum ia mengubah siapa yang bertanggungjawab ke atas aliran kerja tersebut. Doktor tidak digantikan; proses memesan, mentafsir dan mengejar keputusan — aliran kerja — ialah tempat alat seperti ini mula-mula masuk.

Dan ia mengalihkan beban bukti ke arah yang betul. Kemenangan di papan kedudukan pada kes retrospektif ialah sebab untuk menjalankan percubaan prospektif, bukan penggantinya. Para penulis sendiri menyatakannya. Tafsiran jujur tentang MIRA ialah undangan kepada kajian seterusnya — menggunakan piawaian yang sudah dikenal bidang ini: diukur pada pesakit, bukan pada penanda aras.

Ringkasan utama

MIRA ialah ejen AI autonomi yang beroperasi dalam rekod kesihatan elektronik sandbox: ia boleh mengambil sejarah, memesan dan mentafsir makmal, pengimejan dan mikrobiologi, mencapai diagnosis, serta menulis pelan rawatan. Diuji pada 574 kes retrospektif daripada pangkalan data awam MIMIC-IV, merentas lapan diagnosis yang dipilih terlebih dahulu, ia mengatasi doktor dari segi ketepatan diagnosis (88.9% keseluruhan; 87.8% berbanding 78.1% dalam perbandingan langsung) dan membuat keputusan yang sebahagian besarnya selaras dengan garis panduan serta selamat dari segi ubat. Tetapi penilaian itu ialah simulasi pada rekod lama dan hanya melalui teks; sebahagian besar kelebihan datang pada keadaan dengan keputusan ujian yang jelas; MIRA menggunakan jauh lebih banyak ujian darah daripada doktor (kira-kira 51% analit tersedia berbanding 28%); beberapa hasil rawatan dinilai terhadap apa yang direkodkan dalam carta asal; dan set data awam menimbulkan risiko sebenar pencemaran data latihan — yang para penulis sendiri sifatkan sebagai kemungkinan had atas bagi nombor mereka. Kemajuan sebenar ialah ejen yang bertindak merentas seluruh aliran kerja, bukannya menjawab soalan terpencil. Para penulis secara jelas mengatakan bahawa generalisasi, keselamatan dan tadbir urus masih memerlukan kajian prospektif dunia sebenar — itulah tafsiran yang betul: demonstrasi keupayaan yang mengagumkan, bukan bukti bahawa AI mendiagnosis lebih baik daripada doktor, dan bukan sistem yang sudah bersedia untuk klinik sebenar.

Semakan tanpa gembar-gembur

Apa yang ditunjukkan oleh kertas ini: Ejen model bahasa (MIRA) boleh mengendalikan keseluruhan kes klinikal dari awal hingga akhir dalam EHR sandbox — sejarah, ujian, diagnosis, arahan — dan, pada penanda aras retrospektif 574 kes merentas lapan diagnosis, mengatasi doktor dari segi ketepatan diagnosis (88.9% keseluruhan; 87.8% berbanding 78.1% terhadap doktor bertauliah lembaga), tanpa kesilapan ubat berkeparahan tinggi.

Apa yang munasabah tetapi belum terbukti: Bahawa keupayaan ini membawa manfaat kepada pesakit sebenar yang belum dibezakan; bahawa kelebihan ketepatan mencerminkan penaakulan lebih baik dan bukannya lebih banyak ujian yang dipesan, persetujuan dengan carta asal, atau data awam yang telah dihafal.

Apa yang tidak ditunjukkannya: Bahawa MIRA berfungsi di luar lapan diagnosisnya; bahawa ia selamat untuk digunakan; bahawa ia mengatasi doktor dalam perbandingan yang adil dengan maklumat setara; bahawa ia menggantikan klinisian; bahawa hasilnya bebas daripada pencemaran data latihan.

Had utama: Penilaian retrospektif, simulasi, hanya teks; lapan diagnosis terpilih; ketidakseimbangan maklumat (jauh lebih banyak ujian darah — kira-kira 51% analit tersedia berbanding 28%, pada ujian darah kos rendah, bukan pengimejan); hasil rawatan sebahagiannya dinilai terhadap carta sejarah; dan penanda aras awam (MIMIC-IV) yang mungkin pernah dilihat model bahasa semasa latihan — yang para penulis tandakan sebagai kemungkinan had atas prestasi.

Berapa besar keyakinan yang patut ada pada pembaca umum? Tinggi bahawa MIRA ialah demonstrasi keupayaan yang sebenar dan baharu — ejen yang *bertindak* merentas rekod, bukan sekadar chatbot. Rendah bahawa ia ialah *pakar diagnosis yang lebih baik daripada doktor*: dakwaan itu terhad kepada satu penanda aras retrospektif dan dikelirukan oleh pesanan ujian, penilaian terhadap carta, serta kemungkinan pencemaran data. Kesimpulan para penulis sendiri ialah yang betul — ini memerlukan kajian prospektif dunia sebenar sebelum ia membawa makna untuk penjagaan pesakit.

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