The Clean Paper

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

An AI agent worked entire patient cases on its own — in a simulator, on past records

MIRA is a new kind of medical AI: instead of answering a single question, it works an entire case inside a simulated hospital record — taking a history, ordering and reading tests, reaching a diagnosis, writing the orders. On 574 retrospective cases from a public database, across eight pre-selected diagnoses, the authors report it outperformed physicians on diagnostic accuracy and made largely guideline-concordant, medication-safe decisions. Every qualifier in that sentence carries weight: it ran in a sandbox on past records, in text only; much of its edge came on the conditions with clear-cut test results; it ordered about twice as many blood tests as the doctors; and several outcomes were scored against what the original chart recorded. The genuine advance is an agent that acts across the whole workflow — not proof that a machine now diagnoses better than a doctor. The authors say so: generalization, safety and governance still need prospective, real-world studies.

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

Answering a question is not the same as running the case

Most medical-AI stories are about a model that *answers*. You hand it a vignette or an exam question, it hands back a diagnosis or a paragraph of advice, and it scores well. Useful, but narrow: a clever consultant who never touches the chart.

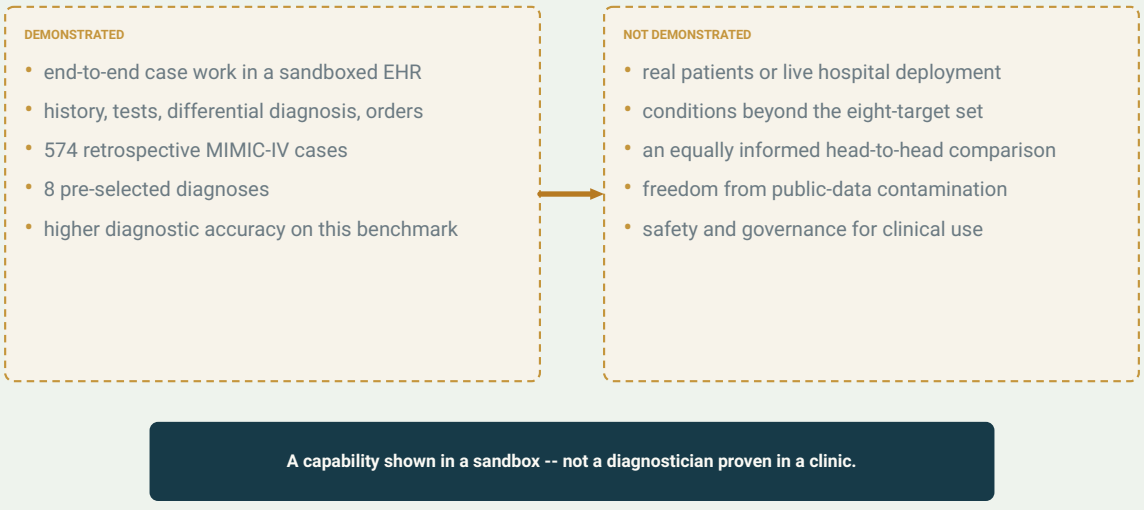
MIRA is built to do something different, and that difference is the actual news. Instead of answering one question, it works an entire case: it reads a patient's record, decides what history it still needs, orders laboratory, imaging and microbiology tests, reads the results, narrows a differential diagnosis, and then writes the orders that follow — prescriptions, an admission, a referral for surgery. It does this by taking actions *inside* an electronic health record, the way a clinician does, rather than by producing free text for a human to transcribe.

That is a real step: from a system that advises to a system that acts across the workflow. It is also exactly the kind of step that gets flattened, in the coverage, into "AI outperforms doctors." So it is worth being precise about what was tested, and where.

The one-sentence version: MIRA worked entire cases on its own and, on this benchmark, outperformed physicians on diagnostic accuracy — but it did so in a sandbox, on retrospective records, across eight pre-chosen diagnoses, communicating only in text, and it reached much of its edge on the conditions with the cleanest test results. The advance is real. The scoreboard is not the clinic.

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.



The figure separates the workflow result from the deployment claim.

MIRA demonstrated an end-to-end workflow capability inside a sandboxed EHR. It did not demonstrate real-world clinical deployment, broad diagnostic coverage, or a fair equally informed head-to-head with doctors.

What the authors did

MIRA — Medical Intelligence for Reasoning and Action — is an autonomous agent built on OpenAI's models: the part that converses and acts runs on **GPT-4o**, and a separate planning step uses OpenAI's **o1** reasoning model. These sit inside a custom, standards-compliant framework — built on HL7 FHIR, the data standard real hospitals use to move records around — with a toolbox of more than eighty thousand possible clinical actions. Inside that sandbox MIRA can pull a patient's history, order and interpret labs, imaging and microbiology, generate a differential diagnosis, and formulate treatment plans — prescribing medications, scheduling surgical procedures, planning admissions. One detail worth flagging up front: the automated "judges" that scored MIRA's answers were themselves GPT-4o — the same model family being tested — which the authors backed with review from a board-certified physician after a peer reviewer raised the concern.

The test set came from **MIMIC-IV**, a large, publicly available de-identified EHR database. From roughly 300,000 patients treated at Beth Israel Deaconess Medical Center between 2008 and 2019, the authors selected **574 patient cases** spanning **eight target diagnoses** — abdominal pathologies and internal-medicine emergencies, among them appendicitis, pancreatitis, pneumonia and urinary-tract infection. For each case, MIRA started from a limited picture and had to decide, step by step, what to do next — the same shape as a real workup, but run against a record whose real ending is already on file.

Its performance was then compared with physicians on the same cases.

What "autonomous agent in a sandboxed EHR" actually means

Three words are doing a lot of work here, so it is worth unpacking them.

Agent means the system is not a chatbot answering a prompt. It runs a loop: look at the current state, choose an action (order this test, ask for that history), see the result, choose the next action — until it reaches a diagnosis and a plan. The "intelligence" is a language model; the *agency* is the scaffolding around it that turns the model's text into permitted EHR operations and feeds the results back in.

Sandboxed EHR means a simulated, walled-off copy of a medical-record system, not a live hospital one. MIRA can "order" a test and receive the result that patient actually got, because the case is historical and the answer is already recorded. Nothing it does touches a real patient or a real clinic.

Autonomous means it completes the case end to end without a human in the loop — *inside the sandbox*. It does **not** mean it was set loose to manage patients unsupervised. Those are very different claims, and only the first was tested.

One more thing about the sandbox, because it is easy to picture wrongly: MIRA may *order* any test it likes, but the system can only hand back a result if that test was **actually performed** for this patient in real life. Ask for a blood panel the patient received, and you get the real values; ask for a scan nobody ordered, and you get back "N/A — could not be performed," never an invented number. History works the same way: if the record does not contain what MIRA asks about, the simulated patient simply says it does not know. So MIRA cannot summon the one

> decisive test that was never taken — it can only work with what the real workup happened to include.
> The distinction matters because the headline word — "autonomous" — is the one most likely to be read as the second thing.

What they found

On the benchmark, **MIRA outperformed physicians on diagnostic accuracy** — but the headline hides three different numbers, and they are easy to blur, so it is worth separating them.

On its own, across all **574 cases**, MIRA named the right diagnosis **88.9%** of the time — scored against the discharge diagnosis that was actually recorded for each patient in MIMIC-IV. For the head-to-head with humans, the doctors did not work all 574 cases; they worked a shared **311-case subset**, using the same records and tools MIRA had. On those same 311 cases, MIRA scored **87.8%**, and it was measured against two separately recruited groups. The first was a **senior group**: four board-certified physicians with 7 to 11 years of experience, who scored **78.1%**. The second was a **mixed-seniority group**: mostly junior residents plus two specialists, closer to how a real emergency department is actually staffed, who scored **71.1%**. Same cases, same tools — and the order came out AI first, the seasoned doctors second, the junior-heavy team third. The paper's own careful phrasing is that MIRA was "consistently equivalent to, and often exceeded" the physicians, across all eight diseases.

Its downstream decisions held up too: largely guideline-concordant, medication-safe and appropriate on admission. The authors report no high-severity medication errors across five safety categories (drug interactions, renal dosing, allergies, QT-risk and opioid prescribing) and prescriptions correct in 467 of 468 cases — while noting the system "did not achieve 100% reliability." Taken at face value, that is a striking result: an agent running the whole case, and coming out ahead of doctors on the diagnosis.

But the *shape* of the win matters as much as the win. Independent specialists reading the paper pointed to where the edge actually came from. On the Science Media Centre's expert panel, Dr Wei Xing (University of Sheffield) noted that the headline figure — the AI beating doctors on diagnostic accuracy — was **mostly driven by the conditions with clear test results**, like appendicitis and pancreatitis, where a decisive scan or lab value settles the question. For pneumonia and urinary-tract infections — two of the most common reasons people actually turn up to an emergency department — *both* the AI and the doctors did worst, and the gap between them was smallest.

There was also an asymmetry in how the two sides played, and it sits in the paper's own numbers. MIRA leaned harder on the lab: it drew on about **51%** of the laboratory analytes available in routine care, against roughly **28%** for the board-certified physicians — a median of seven more blood parameters per case. The authors are careful to frame this as *below* the dataset's own routine-care baseline, not an order-everything strategy, and report *no* systematic increase in higher-cost cross-sectional imaging. Still, more information can, by itself, produce higher diagnostic accuracy — so this is not quite a like-for-like contest between equally-informed players, a gap Xing flagged directly.

A clinician free to order every cheap blood test without weighing cost, discomfort or delay would also look sharper on paper.

Why “beat doctors” is not “better doctor”

A benchmark win is easy to over-read. Three things sit between “MIRA scored higher” and “MIRA is better.”

First, **what it was scored against.** Professor Julie Jacko (University of Edinburgh) noted that several of MIRA’s key outcomes are defined relative to what was documented in the underlying dataset — meaning the system is rewarded for **reproducing the recorded clinical behaviour**, not necessarily for demonstrating optimal care. The historical chart becomes the answer key. That is a reasonable way to build a benchmark, but it measures agreement with what was done, not correctness in some absolute sense. In fairness to the paper, this bites hardest on the *treatment-alignment* metrics — did MIRA’s orders match the chart? — whereas the headline diagnostic accuracy is scored against the discharge diagnosis, which is closer to a real outcome than a documentation echo.

Second, **the asymmetry of information** already mentioned: far more blood tests (about 51% of the available analytes, versus 28%) is more evidence. Some of the accuracy gap is likely being *bought*, not reasoned.

Third, **where it ran.** This was a sandbox, on retrospective cases, across eight pre-selected diagnoses, in text only. Real clinical assessment, as Dr Dominic Oliver (University of Oxford) put it, depends not only on what patients say but on how they say it — alongside physical examination, observed behaviour and body language, none of which a text-only agent reading a finished record ever sees. And a patient whose problem does not fall into one of the eight chosen diagnoses is simply outside what this study can speak to.

There is also a quieter worry the reviewers raised: **data contamination.** MIMIC-IV is public and heavily written about, so a language model trained on the open internet may already have seen papers, case discussions, or the data themselves. Dr Midhun Parakkal Unni (University of Sheffield) flagged this directly — if some of the answers were in the training data, part of the performance is memory, not clinical reasoning, and only independent replication can tell the two apart. Notably, the authors do not hand-wave the point: they write that their results “could be cautiously interpreted as a possible upper bound” and “may overestimate generalization to other public cases” — a paper putting a ceiling on its own headline.

None of this makes MIRA less interesting. It makes the correct reading a calibrated one: on a retrospective benchmark, an agent that could act across the whole record outperformed physicians on the diagnoses with clean tests — partly by ordering more tests, partly by agreeing with the recorded chart. That is a real capability demonstration, not a verdict that machines now diagnose better than doctors.

What this does *not* prove

- It does **not** show MIRA works on real patients. Every case was retrospective and simulated; no patient was ever managed by it.
- It does **not** show it works beyond eight diagnoses. Conditions outside the pre-selected set — the messy, undifferentiated majority of medicine — were not tested.
- It does **not** show it is safe to deploy. The authors themselves write that generalization, safety and governance still need prospective, real-world studies.
- It does **not** establish a fair head-to-head with doctors. It drew on far more blood tests (about 51% of the available analytes versus 28%), and several treatment outcomes were scored partly against the recorded chart rather than against ground-truth best care.
- It does **not** show the result is free of memorization. The public MIMIC-IV data may overlap with the model's training, which independent replication would need to rule out.
- It does **not** mean "AI replaces doctors." It is decision support that can *act* across a record; the reviewers' consensus is that real use will be in partnership with clinicians, who keep authority and supply everything a text record leaves out.

How strong is the evidence?

Split the claim in two, because the evidence is very different for each half.

As a capability demonstration — that a language-model agent can run an entire clinical case inside an EHR sandbox, chaining history, tests, diagnosis and orders end to end — the work is genuinely novel and reasonably convincing. This is the part that is new, and it is the part worth paying attention to.

As a superiority claim — that MIRA is *better than physicians* — the evidence is bounded and should be read with care. It holds on a specific retrospective benchmark, on eight diagnoses, with an information asymmetry (more tests) and an answer key drawn from the historical chart, and with a live possibility of training-data contamination. That is enough to say "the agent performed impressively on this benchmark." It is not enough to say "the agent is a better diagnostician than a doctor," and the authors do not claim the second thing.

The most useful stance is neither "AI beats doctors" nor "just a toy." It is: a new kind of medical AI — one that *acts* across the record instead of answering questions — did well on a hard retrospective benchmark, and now has to prove itself where it has not yet been tried: on real, undifferentiated patients, prospectively, under governance.

Why it matters

For a few years the interesting question in medical AI has quietly changed. It used to be "can a model get the diagnosis right?" — and models kept answering yes, on cleaner and cleaner test sets. MIRA marks the shift to a harder question: "can a model *do the job* — gather, order, interpret, decide, act — across a whole workflow?" That is a more useful question, and a more honest one, because acting is where the difficulty and the risk both live.

Which is why the framing matters. The headline reflex — *AI outperforms doctors* — points at the least novel and least supported part of the result. The genuinely new thing is smaller and more consequential: an agent that can move through the entire record. That capability, if it holds up, changes the **workflow** long before it changes who is in charge of it. The doctor is not replaced; the ordering, the interpreting, the chasing of results — the workflow — is where a tool like this first lands.

And it resets the burden of proof in the right direction. A leaderboard win on retrospective cases is a reason to run the prospective trial, not a substitute for it. The authors say as much. The honest reading of MIRA is an invitation to that next study — held to the standard the field already knows: measured on patients, not on benchmarks.

Clean summary

MIRA is an autonomous AI agent that operates a sandboxed electronic health record: it can take a history, order and interpret labs, imaging and microbiology, reach a diagnosis, and write treatment plans. Tested on 574 retrospective cases from the public MIMIC-IV database, across eight preselected diagnoses, it outperformed physicians on diagnostic accuracy (88.9% overall; 87.8% versus 78.1% head-to-head) and made largely guideline-concordant, medication-safe decisions. But the evaluation was a simulation on past records, in text only; much of the edge came on conditions with clear-cut test results; MIRA drew on far more blood tests than the doctors (about 51% of the available analytes versus 28%); several treatment outcomes were scored against what the original chart recorded; and the public dataset raises a real risk of training-data contamination — which the authors themselves call a possible upper bound on their numbers. The genuine advance is an agent that acts across the whole workflow rather than answering isolated questions. The authors are explicit that generalization, safety and governance still require prospective, real-world studies — which is the correct reading: an impressive capability demonstration, not proof that AI diagnoses better than doctors, and not a system ready for a real clinic.

No-BS check

What the paper shows: A language-model agent (MIRA) can run an entire clinical case end to end inside a sandboxed EHR — history, tests, diagnosis, orders — and, on a 574-case retrospective benchmark across eight diagnoses, outperformed physicians on diagnostic accuracy (88.9% overall; 87.8%

versus 78.1% against board-certified physicians) with no high-severity medication errors.

What is plausible but not proven: That this capability translates into benefit for real, undifferentiated patients; that the accuracy edge reflects better reasoning rather than more tests ordered, agreement with the recorded chart, or memorized public data.

What it does not show: That MIRA works outside its eight diagnoses; that it is safe to deploy; that it beats doctors in a fair, equally-informed comparison; that it replaces clinicians; that the results are free of training-data contamination.

Main limitations: Retrospective, simulated, text-only evaluation; eight pre-selected diagnoses; an information asymmetry (far more blood tests — about 51% of the available analytes versus 28%, in low-cost bloods, not imaging); treatment outcomes partly scored against the historical chart; and a public benchmark (MIMIC-IV) that a language model may have seen in training — which the authors flag as a possible upper bound on performance.

How much confidence should a general reader have? High that MIRA is a real and novel capability demonstration — an agent that *acts* across the record, not just a chatbot. Low that it is a *better diagnostician than a physician*: that claim is bounded to one retrospective benchmark and confounded by test-ordering, scoring against the chart, and possible contamination. The authors' own bottom line is the right one — this needs prospective, real-world study before it means anything for patient care.

Sources

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