



The CLEAN PAPER

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

Um agente de IA conduziu casos completos de doentes sozinho — num simulador, sobre registos passados

O MIRA é um novo tipo de IA médica: em vez de responder a uma única pergunta, conduz um caso inteiro dentro de um registo hospitalar simulado — recolhe história, pede e lê exames, chega a um diagnóstico e escreve as ordens. Em 574 casos retrospectivos de uma base de dados pública, abrangendo oito diagnósticos pré-selecionados, os autores relatam que superou médicos em precisão diagnóstica e tomou decisões em grande parte alinhadas com orientações e seguras quanto a medicamentos. Cada ressalva desta frase pesa: funcionou num ambiente isolado sobre registos passados, apenas por texto; boa parte da vantagem surgiu nas doenças com resultados de exames claros; pediu cerca do dobro das análises sanguíneas dos médicos; e vários desfechos foram pontuados contra aquilo que o processo clínico original registava. O avanço genuíno é um agente que atua ao longo de todo o fluxo — não a prova de que uma máquina agora diagnostica melhor do que um médico. Os autores dizem-no: generalização, segurança e governação ainda precisam de estudos prospetivos no mundo real.

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

Responder a uma pergunta não é o mesmo que conduzir o caso

A maioria das histórias sobre IA médica fala de um modelo que *responde*. Damos-lhe uma vinheta clínica ou uma pergunta de exame, ele devolve um diagnóstico ou um parágrafo de aconselhamento e obtém uma boa pontuação. Útil, mas estreito: um consultor inteligente que nunca toca no processo clínico.

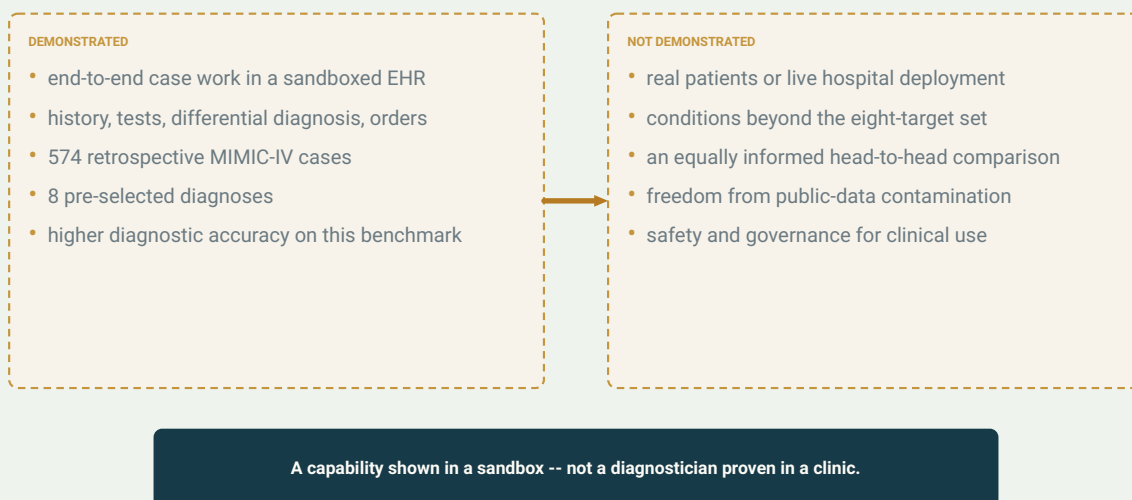
O MIRA foi construído para fazer algo diferente, e essa diferença é a verdadeira novidade. Em vez de responder a uma só pergunta, acompanha um caso inteiro: lê o registo do doente, decide que história clínica ainda precisa de obter, pede análises laboratoriais, exames de imagem e microbiologia, lê os resultados, vai estreitando o diagnóstico diferencial e, por fim, escreve as ordens que daí resultam — prescrições, internamento, referência para cirurgia. Faz isto executando ações *dentro* de um registo clínico eletrónico, como faria um clínico, em vez de produzir texto livre para uma pessoa transcrever.

É um passo real: de um sistema que aconselha para um sistema que atua ao longo do fluxo de trabalho. É também exatamente o tipo de passo que costuma ser achatado, na cobertura mediática, até se transformar em «a IA supera os médicos». Por isso vale a pena ser preciso sobre o que foi testado — e onde.

A versão numa frase: o MIRA conduziu casos inteiros autonomamente e, neste benchmark, superou médicos em precisão diagnóstica — mas fê-lo num ambiente isolado, sobre registos retrospectivos, em oito diagnósticos escolhidos previamente, comunicando apenas por texto, e obteve boa parte da vantagem nas doenças com resultados de testes mais claros. O avanço é real. O marcador não é a clínica.

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.

O MIRA demonstrou capacidade de executar um fluxo clínico de ponta a ponta dentro de um EHR isolado. Não demonstrou implementação clínica no mundo real, cobertura diagnóstica ampla nem uma comparação direta justa com médicos igualmente informados.

Original Aurora diagram — The Clean Paper CC BY 4.0

O que fizeram os autores

MIRA — Medical Intelligence for Reasoning and Action — é um agente autónomo construído sobre modelos da OpenAI: a parte que conversa e atua usa **GPT-4o**, enquanto uma etapa separada de planeamento utiliza o modelo de raciocínio **o1** da OpenAI. Estes componentes funcionam dentro de uma estrutura própria compatível com normas — construída sobre HL7 FHIR, o padrão de dados usado por hospitais reais para trocar registos — com uma caixa de ferramentas de mais de oitenta mil ações clínicas possíveis. Nesse ambiente isolado, o MIRA pode consultar a história de um doente, pedir e interpretar análises, imagiologia e microbiologia, gerar um diagnóstico diferencial e formular planos terapêuticos — prescrever medicamentos, agendar procedimentos cirúrgicos, planear internamentos. Um detalhe merece ser assinalado logo de início: os «juízes» automáticos que pontuaram as respostas do MIRA eram também GPT-4o — a mesma família de modelos que estava a ser testada — e os autores complementaram essa avaliação com revisão por um médico certificado, depois de um revisor por pares ter levantado a questão.

O conjunto de teste veio do **MIMIC-IV**, uma grande base de dados pública e desidentificada de registos clínicos eletrónicos. A partir de cerca de 300 000 doentes tratados no Beth Israel Deaconess Medical Center entre 2008 e 2019, os autores selecionaram **574 casos** abrangendo **oito diagnósticos-alvo** — patologias abdominais e emergências de medicina interna, entre elas apendicite, pancreatite,

pneumonia e infecção urinária. Em cada caso, o MIRA começava com uma imagem limitada e tinha de decidir, passo a passo, o que fazer a seguir — a mesma forma de uma investigação clínica real, mas executada sobre um registo cujo desfecho verdadeiro já estava arquivado.

O seu desempenho foi depois comparado com o de médicos nos mesmos casos.

O que significa realmente «agente autónomo num EHR isolado»

Há três palavras a carregar muito peso aqui, por isso vale a pena destrinchá-las.

Agente significa que o sistema não é um chatbot a responder a um prompt. Executa um ciclo: observa o estado atual, escolhe uma ação (pedir este exame, perguntar aquela parte da história), vê o resultado, escolhe a ação seguinte — até chegar a um diagnóstico e a um plano. A «inteligência» é um modelo de linguagem; a *agência* é a estrutura que o envolve, transforma o texto do modelo em operações permitidas no EHR e devolve os resultados ao modelo.

EHR isolado significa uma cópia simulada e separada de um sistema de registos médicos, não um sistema hospitalar em funcionamento. O MIRA pode «pedir» um exame e receber o resultado que aquele doente realmente teve, porque o caso é histórico e a resposta já está registada. Nada do que faz afeta um doente ou uma clínica reais.

Autónomo significa que completa o caso de ponta a ponta sem uma pessoa no circuito — *dentro do ambiente isolado*. **Não** significa que tenha sido libertado para gerir doentes sem supervisão. São afirmações muito diferentes, e apenas a primeira foi testada.

Mais um ponto sobre o ambiente isolado, porque é fácil imaginá-lo de forma errada: o MIRA pode *pedir* qualquer teste que queira, mas o sistema só consegue devolver um resultado se esse teste tiver sido **realmente realizado** naquele doente. Se pedir um painel sanguíneo que o doente fez, recebe os valores reais; se pedir uma TAC que ninguém pediu, recebe «N/A — não foi possível realizar», nunca um número inventado. A história clínica funciona da mesma maneira: se o registo não contiver aquilo que o MIRA pergunta, o doente simulado responde simplesmente que não sabe. Portanto, o MIRA não pode fazer aparecer o único teste decisivo que nunca foi feito — só pode trabalhar com aquilo que a investigação clínica real acabou por incluir.

A distinção importa porque a palavra da manchete — «autónomo» — é precisamente a que mais facilmente será interpretada como a segunda coisa.

O que encontraram

No benchmark, **o MIRA superou os médicos em precisão diagnóstica** — mas a manchete esconde três números diferentes, fáceis de misturar, por isso convém separá-los.

Sozinho, nos **574 casos**, o MIRA indicou o diagnóstico correto em **88,9%** das vezes — tendo como referência o diagnóstico de alta efetivamente registado para cada doente no MIMIC-IV. Na comparação direta com humanos, os médicos não trabalharam todos os 574 casos; trabalharam um subconjunto comum de **311 casos**, usando os mesmos registos e ferramentas disponíveis para o MIRA. Nesses mesmos 311 casos, o MIRA obteve **87,8%**, e foi comparado com dois grupos recrutados sep-

aradamente. O primeiro era um **grupo sénior**: quatro médicos certificados, com 7 a 11 anos de experiência, que obtiveram **78,1%**. O segundo era um **grupo de senioridade mista**: sobretudo médicos internos mais jovens, além de dois especialistas, mais próximo da composição real de um serviço de urgência, que obteve **71,1%**. Mesmos casos, mesmas ferramentas — e a ordem foi IA em primeiro, médicos experientes em segundo e equipa com predominância de juniores em terceiro. A formulação cuidadosa do próprio artigo é que o MIRA foi «consistentemente equivalente e muitas vezes superior» aos médicos nas oito doenças.

As decisões a jusante também se mantiveram: em grande parte alinhadas com as orientações, seguras no uso de medicamentos e adequadas quanto a internamento. Os autores não relatam erros medicamentosos de elevada gravidade em cinco categorias de segurança — interações medicamentosas, ajuste renal de dose, alergias, risco de QT e prescrição de opioides — e registam prescrições corretas em 467 de 468 casos, ao mesmo tempo que observam que o sistema «não atingiu 100% de fiabilidade». Tomado pelo seu valor nominal, é um resultado impressionante: um agente a conduzir o caso inteiro e a terminar à frente dos médicos no diagnóstico.

Mas a *forma* da vitória importa tanto como a vitória. Especialistas independentes que leram o artigo chamaram a atenção para a origem da vantagem. No painel de especialistas do Science Media Centre, o Dr. Wei Xing (Universidade de Sheffield) observou que o número da manchete — a IA superar médicos em precisão diagnóstica — foi **impulsionado sobretudo pelas doenças com resultados de exames claros**, como apendicite e pancreatite, em que uma imagem ou valor laboratorial decisivo resolve a questão. Na pneumonia e nas infeções urinárias — dois dos motivos mais comuns que levam pessoas reais às urgências — *tanto* a IA como os médicos tiveram o pior desempenho, e a diferença entre ambos foi menor.

Havia também uma assimetria na forma como os dois lados jogaram, visível nos próprios números do artigo. O MIRA apoiou-se mais no laboratório: utilizou cerca de **51%** dos analitos laboratoriais disponíveis nos cuidados de rotina, contra aproximadamente **28%** para os médicos certificados — uma mediana de mais sete parâmetros sanguíneos por caso. Os autores têm o cuidado de enquadrar isto como *abaixo* do próprio nível de utilização de rotina da base de dados, e não como uma estratégia de pedir tudo, e não relatam qualquer aumento sistemático de exames de imagem seccional mais caros. Ainda assim, mais informação pode, por si só, produzir maior precisão diagnóstica — portanto, não se trata exatamente de uma comparação entre participantes igualmente informados, uma diferença que Xing assinalou diretamente. Um clínico livre de pedir todos os exames sanguíneos baratos sem ponderar custo, desconforto ou atraso também pareceria mais certo no papel.

Porque «superar médicos» não significa «ser um médico melhor»

É fácil interpretar demais uma vitória num benchmark. Há três coisas entre «o MIRA teve pontuação superior» e «o MIRA é melhor».

Primeiro, **contra o que foi pontuado**. A professora Julie Jacko (Universidade de Edimburgo) observou que vários dos principais desfechos do MIRA são definidos relativamente ao que ficou documentado

na base de dados subjacente — o que significa que o sistema é recompensado por **reproduzir o comportamento clínico registado**, e não necessariamente por demonstrar os melhores cuidados possíveis. O processo clínico histórico torna-se a chave de respostas. É uma forma razoável de construir um benchmark, mas mede concordância com aquilo que foi feito, não correção num sentido absoluto. Em justiça ao artigo, isto pesa sobretudo nas métricas de *alinhamento terapêutico* — as ordens do MIRA coincidiram com o registo? — enquanto a precisão diagnóstica da manchete é pontuada contra o diagnóstico de alta, que está mais próximo de um desfecho real do que de um simples eco documental.

Segundo, **a assimetria de informação** já referida: muito mais análises sanguíneas — cerca de 51% dos analitos disponíveis contra 28% — significam mais evidência. Parte da diferença de precisão terá provavelmente sido *comprada* com informação, e não obtida apenas por raciocínio.

Terceiro, **onde correu**. Foi um ambiente isolado, com casos retrospectivos, oito diagnósticos pré-selecionados e apenas texto. A avaliação clínica real, como observou o Dr. Dominic Oliver (Universidade de Oxford), depende não só do que os doentes dizem, mas de como o dizem — além do exame físico, comportamento observado e linguagem corporal, nada disso disponível para um agente apenas textual que lê um registo já concluído. E um doente cujo problema não esteja entre os oito diagnósticos escolhidos fica simplesmente fora daquilo sobre que este estudo pode dizer alguma coisa.

Há ainda uma preocupação mais discreta levantada pelos revisores: **contaminação dos dados**. O MIMIC-IV é público e amplamente discutido, pelo que um modelo de linguagem treinado na internet aberta pode já ter visto artigos, discussões de casos ou os próprios dados. O Dr. Midhun Parakkal Unni (Universidade de Sheffield) assinalou isto diretamente — se algumas respostas estavam nos dados de treino, parte do desempenho pode ser memória, e não raciocínio clínico, e só uma replicação independente consegue separar as duas coisas. É importante notar que os autores não afastam a questão: escrevem que os resultados «podem ser interpretados cautelosamente como um possível limite superior» e «podem sobrestimar a generalização para outros casos públicos» — um artigo a colocar um teto na sua própria manchete.

Nada disto torna o MIRA menos interessante. Torna a leitura correta mais calibrada: num benchmark retrospectivo, um agente capaz de atuar ao longo de todo o registo superou médicos sobretudo nos diagnósticos com testes claros — em parte pedindo mais testes, em parte concordando com o processo clínico registado. É uma demonstração real de capacidade, não um veredicto de que as máquinas agora diagnosticam melhor do que os médicos.

O que isto *não* demonstra

- **Não** mostra que o MIRA funcione em doentes reais. Todos os casos eram retrospectivos e simulados; nenhum doente foi alguma vez gerido por ele.
- **Não** mostra que funcione para além de oito diagnósticos. As doenças fora do conjunto pré-selecionado — a maioria confusa e indiferenciada da medicina — não foram testadas.
- **Não** mostra que seja seguro implementá-lo. Os próprios autores escrevem que generalização,

segurança e governação ainda exigem estudos prospetivos no mundo real.

- **Não** estabelece uma comparação direta justa com médicos. O MIRA recorreu a muito mais análises sanguíneas (cerca de 51% dos analitos disponíveis contra 28%), e vários desfechos terapêuticos foram pontuados parcialmente contra o processo clínico registado, não contra uma verdade independente sobre os melhores cuidados.
- **Não** mostra que o resultado esteja livre de memorização. Os dados públicos do MIMIC-IV podem sobrepor-se ao treino do modelo, algo que uma replicação independente teria de excluir.
- **Não** significa «a IA substitui os médicos». É apoio à decisão capaz de *atuar* ao longo de um registo; o consenso dos revisores é que o uso real será em parceria com clínicos, que mantêm a autoridade e fornecem tudo aquilo que um registo textual deixa de fora.

Quão forte é a evidência?

É útil dividir a afirmação em duas, porque a evidência é muito diferente em cada metade.

Enquanto demonstração de capacidade — a ideia de que um agente baseado num modelo de linguagem consegue conduzir um caso clínico inteiro dentro de um EHR isolado, encadeando história, testes, diagnóstico e ordens de ponta a ponta — o trabalho é genuinamente novo e razoavelmente convincente. Esta é a parte nova, e é a que merece atenção.

Enquanto afirmação de superioridade — de que o MIRA é *melhor do que os médicos* — a evidência é limitada e deve ser lida com cuidado. Vale num benchmark retrospectivo específico, em oito diagnósticos, com uma assimetria de informação (mais exames), uma chave de respostas retirada do processo clínico histórico e uma possibilidade real de contaminação por dados de treino. É suficiente para dizer «o agente teve um desempenho impressionante neste benchmark». Não é suficiente para dizer «o agente é um melhor diagnosticador do que um médico», e os autores não fazem essa segunda afirmação.

A posição mais útil não é «a IA supera os médicos» nem «é apenas um brinquedo». É esta: um novo tipo de IA médica — que *atua* ao longo do registo em vez de apenas responder a perguntas — teve bom desempenho num benchmark retrospectivo exigente e agora tem de provar-se onde ainda não foi testado: em doentes reais e indiferenciados, prospetivamente e sob governação.

Porque importa

Nos últimos anos, a pergunta interessante sobre IA médica mudou discretamente. Antes era «um modelo consegue acertar no diagnóstico?» — e os modelos continuaram a responder que sim, em conjuntos de teste cada vez mais limpos. O MIRA assinala a passagem para uma pergunta mais difícil: «um modelo consegue *fazer o trabalho* — recolher informação, pedir exames, interpretar, decidir, agir — ao longo de todo um fluxo?» É uma pergunta mais útil e mais honesta, porque é na ação que vivem tanto a dificuldade como o risco.

É precisamente por isso que o enquadramento importa. O reflexo da manchete — *IA supera médicos* — aponta para a parte menos nova e menos sustentada do resultado. A novidade genuína é menor e mais consequente: um agente que consegue percorrer todo o registo. Essa capacidade, se se confirmar, muda o **fluxo de trabalho** muito antes de mudar quem o dirige. O médico não é substituído; pedir exames, interpretá-los, perseguir resultados — o fluxo — é onde uma ferramenta destas entra primeiro.

E o estudo redefine o ónus da prova na direção certa. Ganhar uma tabela classificativa em casos retrospectivos é uma razão para realizar o ensaio prospetivo, não um substituto dele. Os autores dizem praticamente o mesmo. A leitura honesta do MIRA é um convite a esse próximo estudo — submetido ao padrão que o campo já conhece: medido em doentes, não em benchmarks.

Resumo claro

O MIRA é um agente autónomo de IA que opera num registo clínico eletrónico isolado: pode recolher uma história, pedir e interpretar análises, imagiologia e microbiologia, chegar a um diagnóstico e escrever planos terapêuticos. Testado em 574 casos retrospectivos da base de dados pública MIMIC-IV, abrangendo oito diagnósticos pré-selecionados, superou médicos em precisão diagnóstica (88,9% no total; 87,8% contra 78,1% na comparação direta) e tomou decisões em grande parte alinhadas com orientações e seguras quanto a medicamentos. Mas a avaliação foi uma simulação sobre registos passados, apenas por texto; grande parte da vantagem ocorreu em doenças com resultados de testes claros; o MIRA utilizou muito mais análises sanguíneas do que os médicos (cerca de 51% dos analitos disponíveis contra 28%); vários desfechos terapêuticos foram pontuados contra aquilo que o processo clínico original registava; e a base de dados pública levanta um risco real de contaminação do treino — algo que os próprios autores dizem poder tornar os números um possível limite superior. O verdadeiro avanço é um agente que atua ao longo de todo o fluxo em vez de responder a perguntas isoladas. Os autores são explícitos: generalização, segurança e governação ainda exigem estudos prospetivos no mundo real — que é a leitura correta: uma demonstração impressionante de capacidade, não a prova de que a IA diagnostica melhor do que os médicos, nem um sistema pronto para uma clínica real.

Verificação sem exageros

O que o artigo mostra: Um agente baseado num modelo de linguagem (MIRA) consegue conduzir um caso clínico inteiro de ponta a ponta dentro de um EHR isolado — história, exames, diagnóstico, ordens — e, num benchmark retrospectivo de 574 casos e oito diagnósticos, superou médicos em precisão diagnóstica (88,9% no total; 87,8% contra 78,1% relativamente a médicos certificados), sem erros medicamentosos de elevada gravidade.

O que é plausível, mas não está demonstrado: Que esta capacidade se traduza em benefício para doentes reais e indiferenciados; que a vantagem de precisão reflita melhor raciocínio e não mais

exames pedidos, concordância com o processo registado ou memorização de dados públicos.

O que não mostra: Que o MIRA funcione fora dos oito diagnósticos; que seja seguro implementá-lo; que supere médicos numa comparação justa e com a mesma informação; que substitua clínicos; que os resultados estejam livres de contaminação dos dados de treino.

Principais limitações: Avaliação retrospectiva, simulada e apenas textual; oito diagnósticos pré-selecionados; assimetria de informação (muito mais análises sanguíneas — cerca de 51% dos analitos disponíveis contra 28%, em análises de baixo custo, não imagiologia); desfechos terapêuticos parcialmente pontuados contra o registo histórico; e um benchmark público (MIMIC-IV) que um modelo de linguagem pode ter visto durante o treino — algo que os autores assinalam como possível limite superior do desempenho.

Quanta confiança deve ter um leitor geral? Elevada de que o MIRA é uma demonstração real e nova de capacidade — um agente que *atua* ao longo do registo, não apenas um chatbot. Baixa de que seja um *diagnosticador melhor do que um médico*: essa afirmação está limitada a um único benchmark retrospectivo e confundida pela quantidade de exames pedidos, pela pontuação contra o processo clínico e por possível contaminação. A conclusão dos próprios autores é a correta — isto precisa de estudo prospetivo no mundo real antes de significar algo para os cuidados aos doentes.

Fontes

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>