The Clean Paper

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

Um agente de IA trabalhou casos completos de pacientes por conta própria: em um simulador, com prontuários antigos

MIRA é um novo tipo de IA médica: em vez de responder a uma única pergunta, trabalha um caso inteiro dentro de um prontuário hospitalar simulado: colhe a história, pede e lê exames, chega a um diagnóstico e escreve as ordens. Em 574 casos retrospectivos de uma base pública, com oito diagnósticos pré-selecionados, os autores relatam que superou médicos em acurácia diagnóstica e tomou decisões em grande parte alinhadas a diretrizes e seguras em medicação. Cada qualificador importa: foi um sandbox com prontuários passados, apenas texto; boa parte da vantagem veio de condições com exames claros; pediu cerca de duas vezes mais exames de sangue que os médicos; e vários resultados foram pontuados contra o que estava no prontuário original. O avanço real é um agente que atua ao longo de todo o fluxo de trabalho, não uma prova de que uma máquina diagnostica melhor que um médico.

Written by Lucio Vaglio · Cross-checked by Vera Rubin · Edited by Michele Renda · Figures by Laura Nesso · 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

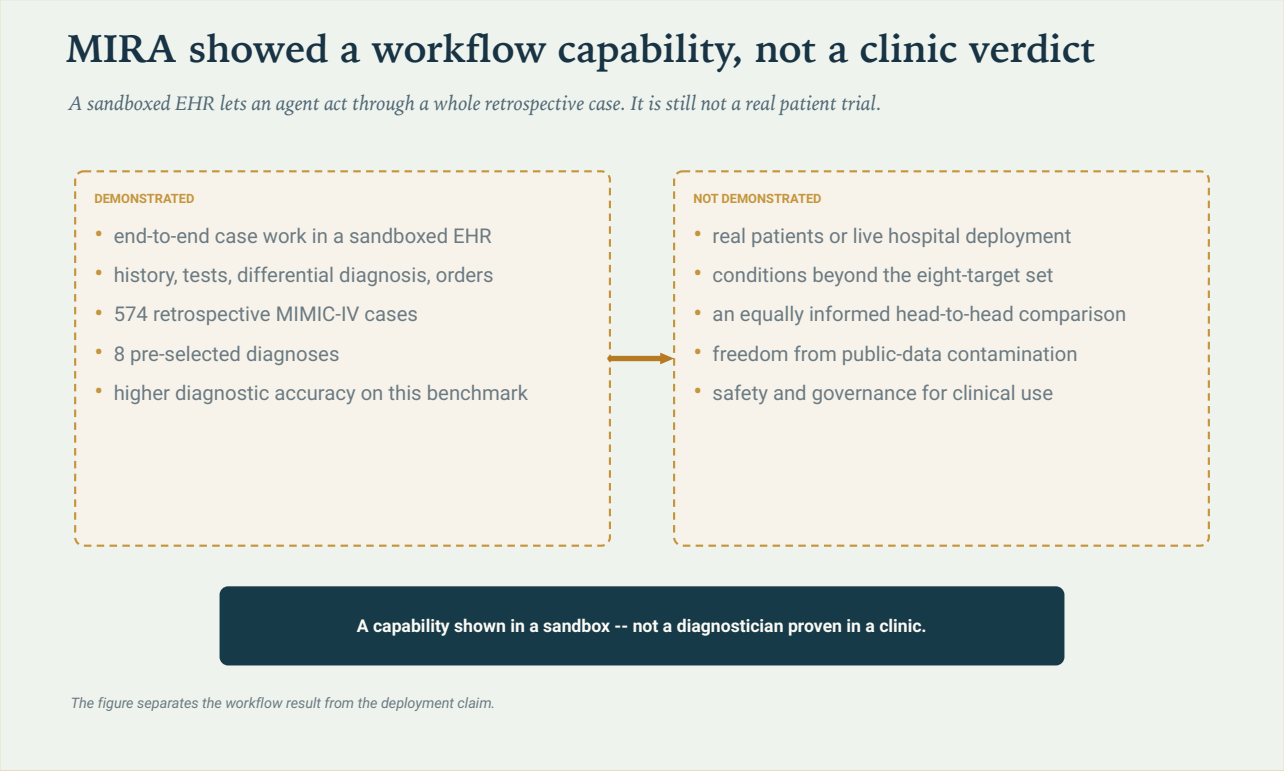
Responder a uma pergunta não é o mesmo que conduzir o caso

A maioria das histórias sobre IA médica trata de um modelo que *responde*. Você lhe entrega uma vinheta ou uma questão de prova, ele devolve um diagnóstico ou um parágrafo de orientação, e pontua bem. Útil, mas estreito: um consultor inteligente que nunca toca no prontuário.

O MIRA foi construído para fazer algo diferente, e essa diferença é a notícia real. Em vez de responder a uma pergunta, ele trabalha um caso inteiro: lê o prontuário de um paciente, decide que história ainda precisa, solicita exames laboratoriais, de imagem e microbiologia, lê os resultados, estreita um diagnóstico diferencial e então escreve as ordens que decorrem disso — prescrições, uma internação, um encaminhamento para cirurgia. Faz isso tomando ações *dentro* de um prontuário eletrônico, como um clínico faz, em vez de produzir texto livre para um humano transcrever.

Isso é um passo real: de um sistema que aconselha para um sistema que age ao longo do fluxo de trabalho. É também exatamente o tipo de passo que, na cobertura, vira "IA supera médicos". Por isso vale ser preciso sobre o que foi testado, e onde.

A versão em uma frase: o MIRA trabalhou casos inteiros por conta própria e, neste benchmark, superou médicos em acurácia diagnóstica — mas fez isso em um sandbox, com prontuários retrospectivos, em oito diagnósticos pré-escolhidos, comunicando-se apenas por texto, e alcançou boa parte de sua vantagem nas condições com resultados de exame mais limpos. O avanço é real. O placar não é a clínica.



O MIRA demonstrou uma capacidade de fluxo de trabalho de ponta a ponta dentro de um prontuário eletrônico em sandbox. Não demonstrou implantação clínica no mundo real, cobertura diagnóstica ampla nem uma

comparação justa, igualmente informada, com médicos.

O que os autores fizeram

MIRA — Medical Intelligence for Reasoning and Action — é um agente autônomo construído sobre modelos da OpenAI: a parte que conversa e age roda em **GPT-4o**, e uma etapa separada de planejamento usa o modelo de raciocínio **o1** da OpenAI. Eles ficam dentro de uma estrutura personalizada e compatível com padrões — construída sobre HL7 FHIR, o padrão de dados que hospitais reais usam para mover prontuários — com uma caixa de ferramentas de mais de oitenta mil ações clínicas possíveis. Dentro desse sandbox, o MIRA pode puxar o histórico do paciente, solicitar e interpretar exames laboratoriais, de imagem e microbiologia, gerar um diagnóstico diferencial e formular planos de tratamento — prescrever medicamentos, agendar procedimentos cirúrgicos, planejar internações. Um detalhe que merece ser sinalizado logo de início: os "juízes" automatizados que pontuaram as respostas do MIRA eram eles próprios GPT-4o — a mesma família de modelos em teste —, algo que os autores apoiaram com revisão de um médico certificado depois que um parecerista levantou a preocupação.

O conjunto de teste veio do **MIMIC-IV**, uma grande base pública e desidentificada de prontuários eletrônicos. De cerca de 300.000 pacientes tratados no Beth Israel Deaconess Medical Center entre 2008 e 2019, os autores selecionaram **574 casos de pacientes** cobrindo **oito diagnósticos-alvo** — patologias abdominais e emergências de medicina interna, entre elas apendicite, pancreatite, pneumonia e infecção urinária. Para cada caso, o MIRA começava com um quadro limitado e precisava decidir, passo a passo, o que fazer em seguida — a mesma forma de uma investigação real, mas executada contra um prontuário cujo desfecho real já estava registrado.

Seu desempenho foi então comparado com o de médicos nos mesmos casos.

> **O que "agente autônomo em um prontuário eletrônico em sandbox" realmente significa**
>
> Três palavras fazem muito trabalho aqui, então vale destrinchá-las.
> **Agente** significa que o sistema não é um chatbot respondendo a um prompt. Ele roda um ciclo: olha o estado atual, escolhe uma ação (pedir este exame, perguntar aquela história), vê o resultado, escolhe a próxima ação — até chegar a um diagnóstico e um plano. A "inteligência" é um modelo de linguagem; a *agência* é o andaime ao redor dele que transforma o texto do modelo em operações permitidas no prontuário eletrônico e devolve os resultados.
> **Prontuário eletrônico em sandbox** significa uma cópia simulada e isolada de um sistema de prontuário, não um sistema hospitalar vivo. O MIRA pode "pedir" um exame e receber o resultado que aquele paciente realmente teve, porque o caso é histórico e a resposta já está registrada. Nada que ele faz toca um paciente real ou uma clínica real.
> **Autônomo** significa que ele completa o caso de ponta a ponta sem um humano no ciclo — *dentro do sandbox*. Não significa que foi solto para cuidar de pacientes sem supervisão. São

> afirmações muito diferentes, e só a primeira foi testada.
> Mais uma coisa sobre o sandbox, porque é fácil imaginá-lo errado: o MIRA pode *pedir* qualquer exame que quiser, mas o sistema só consegue devolver um resultado se aquele exame foi **realmente feito** para esse paciente na vida real. Peça um painel de sangue que o paciente recebeu, e você obtém os valores reais; peça uma tomografia que ninguém solicitou, e recebe "N/A — não pôde ser realizada", nunca um número inventado. A história funciona do mesmo modo: se o prontuário não contém o que o MIRA pergunta, o paciente simulado simplesmente diz que não sabe. Portanto, o MIRA não consegue invocar o exame decisivo que nunca foi feito — só pode trabalhar com o que a investigação real acabou incluindo.
> A distinção importa porque a palavra de manchete — "autônomo" — é a que mais provavelmente será lida como a segunda coisa.

O que eles encontraram

No benchmark, **o MIRA superou médicos em acurácia diagnóstica** — mas a manchete esconde três números diferentes, e é fácil misturá-los, então vale separá-los.

Sozinho, em todos os **574 casos**, o MIRA nomeou o diagnóstico correto **88,9%** das vezes — pontuado contra o diagnóstico de alta realmente registrado para cada paciente 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 prontuários e ferramentas que o MIRA tinha. Nesses mesmos 311 casos, o MIRA marcou **87,8%**, e foi medido contra dois grupos recrutados separadamente. O primeiro era um **grupo sênior**: quatro médicos certificados, com 7 a 11 anos de experiência, que marcaram **78,1%**. O segundo era um **grupo de senioridade mista**: em sua maioria residentes juniores mais dois especialistas, mais próximo de como um pronto-socorro real costuma ser preenchido, que marcou **71,1%**. Mesmos casos, mesmas ferramentas — e a ordem saiu IA primeiro, médicos experientes em segundo, equipe com muitos juniores em terceiro. A formulação cuidadosa do próprio artigo é que o MIRA foi "consistentemente equivalente a, e muitas vezes excedeu" os médicos, nas oito doenças.

Suas decisões posteriores também se sustentaram: em grande parte concordantes com diretrizes, seguras em medicação e apropriadas quanto à internação. Os autores relatam nenhum erro medicamentoso de alta gravidade em cinco categorias de segurança (interações medicamentosas, dose renal, alergias, risco de QT e prescrição de opioides) e prescrições corretas em 467 de 468 casos — observando ao mesmo tempo que o sistema "não atingiu 100% de confiabilidade". Tomado pelo valor de face, é um resultado marcante: um agente conduzindo o caso inteiro e saindo à frente de médicos no diagnóstico.

Mas a *forma* da vitória importa tanto quanto a vitória. Especialistas independentes que leram o artigo apontaram de onde a vantagem realmente veio. No painel de especialistas do Science Media Centre, o Dr. Wei Xing (Universidade de Sheffield) observou que o número de manchete — a IA batendo médicos em acurácia diagnóstica — foi **impulsionado principalmente pelas condições com resultados de exame claros**, como apendicite e pancreatite, em que uma tomografia ou um valor laboratorial

decisivo resolve a questão. Para pneumonia e infecção urinária — duas das razões mais comuns pelas quais pessoas realmente chegam a um pronto-socorro — *tanto* a IA quanto os médicos foram pior, e a diferença entre eles foi menor.

Também havia uma assimetria em como os dois lados jogaram, e ela está nos próprios números do artigo. O MIRA se apoiou mais no laboratório: usou cerca de **51%** dos analitos laboratoriais disponíveis no cuidado de rotina, contra aproximadamente **28%** para os médicos certificados — uma mediana de sete parâmetros sanguíneos a mais por caso. Os autores têm o cuidado de enquadrar isso como *abaixo* da linha de base de cuidado rotineiro da própria base, não como uma estratégia de pedir tudo, e relatam *nenhum* aumento sistemático em exames de imagem seccionais de maior custo. Ainda assim, mais informação pode, por si só, produzir maior acurácia diagnóstica — portanto não é exatamente uma disputa igual entre jogadores igualmente informados, uma lacuna que Xing apontou diretamente. Um clínico livre para pedir todo exame de sangue barato sem pesar custo, desconforto ou atraso também pareceria mais afiado no papel.

Por que "bateu médicos" não é "médico melhor"

Uma vitória em benchmark é fácil de ler demais. Três coisas ficam entre "MIRA pontuou mais alto" e "MIRA é melhor".

Primeiro, **contra o que foi pontuado.** A professora Julie Jacko (Universidade de Edimburgo) observou que vários dos principais resultados do MIRA são definidos em relação ao que foi documentado na base subjacente — ou seja, o sistema é recompensado por **reproduzir o comportamento clínico registrado**, não necessariamente por demonstrar cuidado ótimo. O prontuário histórico vira o gabarito. É uma forma razoável de construir um benchmark, mas mede concordância com o que foi feito, não correção em algum sentido absoluto. Em justiça ao artigo, isso pesa mais nos indicadores de *alinhamento de tratamento* — as ordens do MIRA combinaram com o prontuário? —, enquanto a acurácia diagnóstica de manchete é pontuada contra o diagnóstico de alta, que fica mais perto de um desfecho real do que de um eco da documentação.

Segundo, **a assimetria de informação** já mencionada: muito mais exames de sangue (cerca de 51% dos analitos disponíveis, contra 28%) significa mais evidência. Parte da diferença de acurácia provavelmente está sendo *comprada*, não raciocinada.

Terceiro, **onde ele rodou.** Isso foi um sandbox, com casos retrospectivos, em oito diagnósticos pré-selecionados, apenas em texto. A avaliação clínica real, como colocou o Dr. Dominic Oliver (Universidade de Oxford), depende não só do que pacientes dizem, mas de como dizem — junto com exame físico, comportamento observado e linguagem corporal, nada disso visto por um agente só de texto lendo um prontuário já fechado. E um paciente cujo problema não se encaixa em um dos oito diagnósticos escolhidos está simplesmente fora do que este estudo pode dizer.

Há também uma preocupação mais silenciosa levantada pelos revisores: **contaminação de dados.** O MIMIC-IV é público e muito discutido, então 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) apontou isso diretamente — se algumas respostas estavam nos dados

de treinamento, parte do desempenho é memória, não raciocínio clínico, e só replicação independente pode separar os dois. Notavelmente, os autores não minimizam o ponto: escrevem que seus resultados "poderiam ser interpretados cautelosamente como um possível limite superior" e "podem superestimar a generalização para outros casos públicos" — um artigo colocando um teto na própria manchete.

Nada disso torna o MIRA menos interessante. Torna a leitura correta uma leitura calibrada: em um benchmark retrospectivo, um agente capaz de agir ao longo de todo o prontuário superou médicos nos diagnósticos com exames limpos — em parte por pedir mais exames, em parte por concordar com o prontuário registrado. É uma demonstração real de capacidade, não um veredito de que máquinas agora diagnosticam melhor que médicos.

O que isso *não* prova

- Não mostra que o MIRA funciona com pacientes reais. Todo caso foi retrospectivo e simulado; nenhum paciente jamais foi manejado por ele.
- Não mostra que funciona além de oito diagnósticos. Condições fora do conjunto pré-selecionado — a maioria bagunçada e indiferenciada da medicina — não foram testadas.
- Não mostra que é seguro para implantação. Os próprios autores escrevem que generalização, segurança e governança ainda exigem estudos prospectivos no mundo real.
- Não estabelece uma comparação justa com médicos. Ele usou muito mais exames de sangue (cerca de 51% dos analitos disponíveis contra 28%), e vários desfechos de tratamento foram pontuados em parte contra o prontuário registrado, não contra o melhor cuidado verdadeiro.
- Não mostra que o resultado está livre de memorização. Os dados públicos do MIMIC-IV podem se sobrepor ao treinamento do modelo, algo que replicação independente precisaria descartar.
- Não significa "IA substitui médicos". É apoio à decisão que pode *agir* ao longo de um prontuário; o consenso dos revisores é que o uso real será em parceria com clínicos, que mantêm autoridade e fornecem tudo que um registro textual deixa de fora.

Quão forte é a evidência?

Divida a afirmação em duas, porque a evidência é muito diferente para cada metade.

Como demonstração de capacidade — que um agente de modelo de linguagem consegue conduzir um caso clínico inteiro dentro de um sandbox de prontuário eletrônico, encadeando história, exames, diagnóstico e ordens de ponta a ponta — o trabalho é genuinamente novo e razoavelmente convincente. Esta é a parte nova, e é a parte que merece atenção.

Como afirmação de superioridade — que o MIRA é *melhor que médicos* — a evidência é delimitada e deve ser lida com cuidado. Ela vale em um benchmark retrospectivo específico, em oito diagnósticos, com uma assimetria de informação (mais exames) e um gabarito retirado do prontuário histórico, além de uma possibilidade viva de contaminação por dados de treinamento. Isso basta para dizer

"o agente teve desempenho impressionante neste benchmark". Não basta para dizer "o agente é um diagnosticador melhor que um médico", e os autores não afirmam a segunda coisa.

A postura mais útil não é nem "IA bate médicos" nem "é só um brinquedo". É: um novo tipo de IA médica — uma que *age* ao longo do prontuário em vez de responder perguntas — foi bem em um benchmark retrospectivo difícil, e agora precisa provar-se onde ainda não foi tentada: em pacientes reais, indiferenciados, prospectivamente, sob governança.

Por que importa

Há alguns anos, a pergunta interessante em IA médica mudou discretamente. Costumava ser "um modelo consegue acertar o diagnóstico?" — e modelos continuavam respondendo que sim, em conjuntos de teste cada vez mais limpos. O MIRA marca a mudança para uma pergunta mais difícil: "um modelo consegue *fazer o trabalho* — colher, pedir, interpretar, decidir, agir — ao longo de um fluxo de trabalho inteiro?" Essa é uma pergunta mais útil, e mais honesta, porque agir é onde a dificuldade e o risco vivem.

É por isso que o enquadramento importa. O reflexo de manchete — *IA supera médicos* — aponta para a parte menos nova e menos sustentada do resultado. A coisa genuinamente nova é menor e mais consequente: um agente que consegue se mover pelo prontuário inteiro. Essa capacidade, se se sustentar, muda o **fluxo de trabalho** muito antes de mudar quem está no comando dele. O médico não é substituído; pedir, interpretar, perseguir resultados — o fluxo de trabalho — é onde uma ferramenta assim aterrissa primeiro.

E ela recoloca o ônus da prova na direção certa. Uma vitória em ranking com casos retrospectivos é motivo para fazer o ensaio prospectivo, não substituto dele. Os autores dizem isso. A leitura honesta do MIRA é um convite para esse próximo estudo — submetido ao padrão que a área já conhece: medido em pacientes, não em benchmarks.

Resumo limpo

MIRA é um agente autônomo de IA que opera um prontuário eletrônico em sandbox: consegue colher história, pedir e interpretar exames laboratoriais, de imagem e microbiologia, chegar a um diagnóstico e escrever planos de tratamento. Testado em 574 casos retrospectivos da base pública MIMIC-IV, em oito diagnósticos pré-selecionados, superou médicos em acurácia diagnóstica (88,9% no geral; 87,8% contra 78,1% na comparação direta) e tomou decisões em grande parte concordantes com diretrizes e seguras em medicação. Mas a avaliação foi uma simulação com prontuários passados, apenas em texto; boa parte da vantagem veio em condições com resultados de exame claros; o MIRA usou muito mais exames de sangue que os médicos (cerca de 51% dos analitos disponíveis contra 28%); vários desfechos de tratamento foram pontuados contra o que o prontuário original registrava; e a base pública levanta um risco real de contaminação por dados de treinamento — que os próprios autores chamam de possível limite superior para seus números. O avanço genuíno é um agente que

age ao longo de todo o fluxo de trabalho, em vez de responder perguntas isoladas. Os autores são explícitos que generalização, segurança e governança ainda exigem estudos prospectivos no mundo real — que é a leitura correta: uma demonstração impressionante de capacidade, não prova de que IA diagnostica melhor que médicos, e não um sistema pronto para uma clínica real.

Checagem sem rodeios

O que o artigo mostra: Um agente de modelo de linguagem (MIRA) consegue conduzir um caso clínico inteiro de ponta a ponta dentro de um prontuário eletrônico em sandbox — história, exames, diagnóstico, ordens — e, em um benchmark retrospectivo de 574 casos em oito diagnósticos, superou médicos em acurácia diagnóstica (88,9% no geral; 87,8% contra 78,1% frente a médicos certificados) sem erros medicamentosos de alta gravidade.

O que é plausível, mas não provado: Que essa capacidade se traduza em benefício para pacientes reais e indiferenciados; que a vantagem de acurácia reflita melhor raciocínio, em vez de mais exames solicitados, concordância com o prontuário registrado ou memorização de dados públicos.

O que não mostra: Que o MIRA funciona fora de seus oito diagnósticos; que é seguro para implantação; que bate médicos em uma comparação justa, igualmente informada; que substitui clínicos; que os resultados estão livres de contaminação por dados de treinamento.

Principais limitações: Avaliação retrospectiva, simulada e apenas textual; oito diagnósticos pré-selecionados; assimetria de informação (muito mais exames de sangue — cerca de 51% dos analitos disponíveis contra 28%, em exames baratos de sangue, não imagem); desfechos de tratamento pontuados em parte contra o prontuário histórico; e um benchmark público (MIMIC-IV) que um modelo de linguagem pode ter visto no treinamento — algo que os autores sinalizam como possível limite superior de desempenho.

Quanta confiança um leitor geral deve ter? Alta de que o MIRA é uma demonstração real e nova de capacidade — um agente que *age* ao longo do prontuário, não apenas um chatbot. Baixa de que ele seja um *diagnosticador melhor que um médico*: essa afirmação está delimitada a um benchmark retrospectivo e confundida por solicitação de exames, pontuação contra o prontuário e possível contaminação. A conclusão dos próprios autores é a correta — isso precisa de estudo prospectivo no mundo real antes de significar algo para o cuidado de pacientes.

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-patient-management-mira-and-amie/

News-Medical (2026) — illustrates the 'outperforms doctors' framing
https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-simulated-EHR-cases.aspx