The Clean Paper

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

Un agente de IA trabajo casos completos de pacientes por su cuenta: en un simulador, con historiales pasados

MIRA es un nuevo tipo de IA medica: en vez de responder una sola pregunta, trabaja un caso entero dentro de un historial hospitalario simulado: toma la historia, pide y lee pruebas, llega a un diagnostico y redacta las ordenes. En 574 casos retrospectivos de una base publica, con ocho diagnosticos preseleccionados, los autores informan que supero a medicos en precision diagnostica y tomo decisiones en gran parte acordes con guias y seguras en medicacion. Cada calificativo importa: fue un sandbox con historiales pasados, solo texto; gran parte de la ventaja vino de condiciones con pruebas claras; pidio unas dos veces mas analisis de sangre que los medicos; y varios resultados se puntuaron contra lo registrado en la historia original. El avance real es un agente que actua a traves de todo el flujo de trabajo, no una prueba de que una maquina diagnostique mejor que un medico.

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

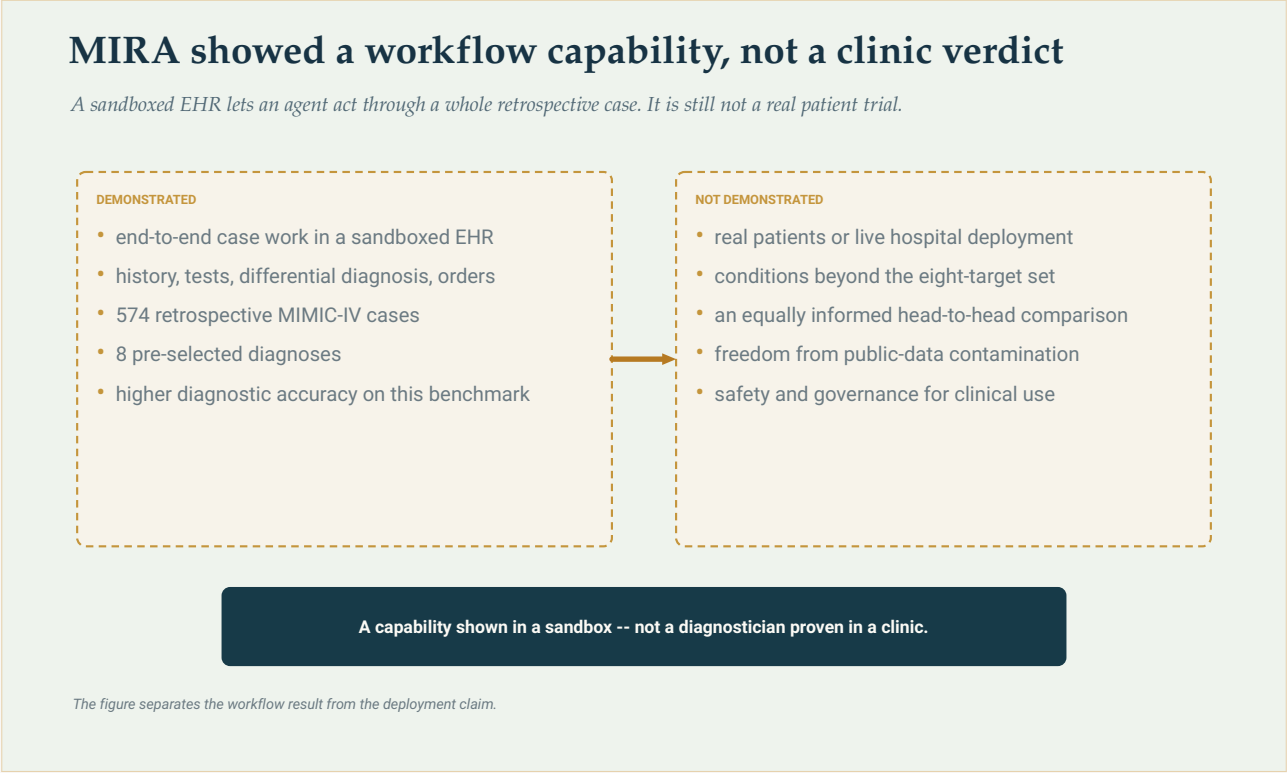
Responder una pregunta no es lo mismo que llevar el caso

La mayoría de historias de IA médica tratan de un modelo que *responde*. Le das una viñeta o una pregunta de examen, devuelve un diagnóstico o un parrafo de consejo y puntúa bien. Util, pero estrecho: un consultor listo que nunca toca la historia clínica.

MIRA está construido para hacer algo distinto, y esa diferencia es la noticia real. En vez de responder una pregunta, trabaja un caso entero: lee el historial del paciente, decide que historia le falta, pide pruebas de laboratorio, imagen y microbiología, lee los resultados, estrecha el diagnóstico diferencial y luego escribe las órdenes que siguen: prescripciones, ingreso, derivación quirúrgica. Lo hace tomando acciones *dentro* de una historia clínica electrónica, como un clínico, no produciendo texto libre para que un humano lo transcriba.

Es un paso real: de un sistema que aconseja a un sistema que actúa a través del flujo de trabajo. También es exactamente el tipo de paso que la cobertura aplana en “la IA supera a los médicos”. Por eso conviene ser precisos sobre qué se probó y dónde.

Versión en una frase: MIRA trabajó casos completos por su cuenta y, en este benchmark, superó a médicos en precisión diagnóstica; pero lo hizo en un sandbox, con historiales retrospectivos, en ocho diagnósticos preseleccionados, comunicándose solo por texto, y obtuvo buena parte de su ventaja en condiciones con pruebas claras. El avance es real. El marcador no es la clínica.



MIRA demostró una capacidad de flujo de trabajo extremo a extremo dentro de una historia clínica electrónica en sandbox. No demostró despliegue clínico real, cobertura diagnóstica amplia ni una comparación justa e igualmente informada con médicos.

Qué hicieron los autores

MIRA - Medical Intelligence for Reasoning and Action - es un agente autónomo construido sobre modelos de OpenAI: la parte que conversa y actúa usa **GPT-4o**, y un paso de planificación separado usa el modelo de razonamiento **o1**. Están dentro de un marco custom conforme a estándares, basado en HL7 FHIR, el estándar de datos que usan hospitales reales para mover historiales, con una caja de herramientas de más de ochenta mil acciones clínicas posibles. En ese sandbox, MIRA puede extraer la historia, pedir e interpretar análisis, imagen y microbiología, generar un diagnóstico diferencial y formular planes de tratamiento: prescribir medicamentos, programar cirugías, planificar ingresos. Detalle a marcar: los "jueces" automáticos que puntuaron a MIRA eran también GPT-4o, la misma familia de modelos probada, respaldados por revisión de un médico certificado tras la preocupación de un revisor.

El conjunto de prueba vino de **MIMIC-IV**, una gran base pública de historias clínicas de-identificadas. De unos 300.000 pacientes tratados en Beth Israel Deaconess Medical Center entre 2008 y 2019, los autores eligieron **574 casos** con **ocho diagnósticos objetivo**: patologías abdominales y urgencias de medicina interna, como apendicitis, pancreatitis, neumonía e infección urinaria. En cada caso, MIRA empezaba con una imagen limitada y debía decidir paso a paso qué hacer, la forma de un estudio real, pero sobre un historial cuyo final ya estaba registrado.

Luego compararon su rendimiento con médicos en los mismos casos.

> **Qué significa realmente "agente autónomo en una historia clínica electrónica en sandbox"**
>
> Tres palabras hacen mucho trabajo.
> **Agente** significa que el sistema no es un chatbot que responde a un prompt. Ejecuta un bucle: mira el estado actual, elige una acción (pedir esta prueba, preguntar aquella historia), ve el resultado, elige la siguiente, hasta llegar a diagnóstico y plan. La inteligencia es un modelo de lenguaje; la *agencia* es el andamiaje que convierte texto en operaciones EHR permitidas y devuelve resultados.
> **EHR en sandbox** significa una copia simulada y aislada de un sistema de historia clínica, no un hospital vivo. MIRA puede "pedir" una prueba y recibir el resultado que ese paciente realmente tuvo, porque el caso es historico y la respuesta ya esta registrada. Nada toca a un paciente real.
> **Autonomo** significa que completa el caso de punta a punta sin humano en el bucle, *dentro del sandbox*. No significa que se soltara para manejar pacientes sin supervision. Son afirmaciones muy distintas, y solo la primera fue probada.
> Un punto más: MIRA puede pedir cualquier prueba, pero el sistema solo devuelve resultado si esa prueba se hizo **realmente** al paciente. Si pide un escáner que nadie pidió, recibe "N/A", no un número inventado. La historia funciona igual: si el registro no contiene lo que MIRA pregunta, el paciente simulado dice que no lo sabe. No puede invocar la prueba decisiva que nunca se tomó.
> La distinción importa porque la palabra de titular, "autónomo", es la que más fácilmente se lee como la segunda cosa.

Qué encontraron

En el benchmark, **MIRA superó a los médicos en precisión diagnóstica**, pero el titular mezcla tres números. Por su cuenta, en los **574 casos**, MIRA nombró el diagnóstico correcto **88,9 %** de las veces, frente al diagnóstico de alta registrado. Para la comparación humana, los médicos no trabajaron los 574 casos; trabajaron un subconjunto compartido de **311 casos**, con los mismos historiales y herramientas que MIRA. En esos 311, MIRA marco **87,8 %** y se comparó con dos grupos. El primero, **senior**, eran cuatro médicos certificados con 7 a 11 años de experiencia, que marcaron **78,1 %**. El segundo, de **senioridad mixta**, tenía sobre todo residentes junior más dos especialistas, más parecido a una urgencia real, y marcó **71,1 %**. Mismos casos, mismas herramientas: IA primero, médicos experimentados después, equipo joven tercero.

Sus decisiones aguas abajo también aguantaron: en gran parte acordes con guías, seguras en medicacion y apropiadas para ingreso. Los autores no informan errores medicamentosos de alta severidad en cinco categorías y prescripciones correctas en 467 de 468 casos, aunque aclaran que el sistema "no alcanzó 100 % de fiabilidad". Tomado literalmente, impresiona: un agente lleva todo el caso y sale por delante en diagnóstico.

Pero la *forma* de la victoria importa tanto como la victoria. Especialistas independientes señalaron de donde venia la ventaja. En el panel del Science Media Centre, el Dr Wei Xing noto que la cifra headline - la IA batiendo a médicos - estaba **impulsada sobre todo por condiciones con pruebas claras**, como apendicitis y pancreatitis, donde un escáner o valor de laboratorio decide. Para neumonía e infección urinaria, dos motivos muy comunes de urgencias, tanto IA como médicos rindieron peor y la brecha fue menor.

También había una asimetría de información. MIRA usó más laboratorio: cerca del **51 %** de los analitos disponibles en atención rutinaria, frente a alrededor del **28 %** para los médicos certificados, una mediana de siete parámetros sanguíneos más por caso. Los autores lo enmarcan como *por debajo* de la línea base rutinaria del dataset, no como pedirlo todo, y no reportan aumento sistemático de imagen transversal costosa. Aun así, más información por sí sola puede producir más precisión; no es un duelo exactamente igual de informado.

Por qué "batir médicos" no es "mejor médico"

Un triunfo de benchmark se sobrelee con facilidad. Tres cosas separan "MIRA puntuo más" de "MIRA es mejor".

Primero, **contra qué se puntuó**. Julie Jacko señaló que varios resultados de MIRA se definen respecto a lo documentado en la base subyacente: se recompensa al sistema por **reproducir la conducta clínica registrada**, no necesariamente por demostrar atención óptima. La historia histórica se vuelve la clave de respuestas. Es razonable para un benchmark, pero mide acuerdo con lo hecho, no corrección absoluta. En justicia, esto golpea más a las métricas de tratamiento; la precisión diagnóstica principal se puntúa contra el diagnóstico de alta.

Segundo, **la asimetría de información**: muchos más análisis de sangre son más evidencia. Parte de la brecha probablemente se *compra*, no se razona.

Tercero, **donde corrió**. Fue un sandbox, con casos retrospectivos, ocho diagnósticos preseleccionados, solo texto. La evaluación clínica real depende también de cómo habla el paciente, del examen físico, conducta observada y lenguaje corporal, nada de lo cual ve un agente de texto leyendo una historia acabada. Y un paciente fuera de los ocho diagnósticos simplemente está fuera de lo que el estudio puede decir.

También hay una preocupación más silenciosa: **contaminación de datos**. MIMIC-IV es pública y muy discutida; un modelo entrenado con internet abierto podría haber visto artículos, debates de casos o los datos. Si algunas respuestas estaban en entrenamiento, parte del rendimiento sería memoria, no razonamiento clínico. Los autores no lo esquivan: escriben que sus resultados podrían interpretarse como una posible cota superior y sobreestimar la generalización a otros casos públicos.

Nada de esto hace a MIRA menos interesante. Hace que la lectura correcta sea calibrada: en un benchmark retrospectivo, un agente que actúa en todo el historial superó a médicos en diagnósticos con pruebas limpias, en parte por pedir más pruebas y en parte por coincidir con la historia registrada. Demostración real de capacidad, no veredicto de que las máquinas diagnostican mejor que los médicos.

Lo que esto no prueba

- No muestra que MIRA funcione en pacientes reales. Todos los casos fueron retrospectivos y simulados.
- No muestra que funcione fuera de ocho diagnósticos. La mayoría desordenada e indiferenciada de la medicina no fue probada.
- No muestra que sea seguro de desplegar. Generalizacion, seguridad y gobernanza necesitan estudios prospectivos reales.
- No establece una comparación justa con médicos. Uso muchos más análisis de sangre y varios resultados se puntuaron contra la historia registrada.
- No muestra que el resultado este libre de memorización. Los datos públicos de MIMIC-IV pueden solaparse con entrenamiento.
- No significa “la IA reemplaza médicos”. Es apoyo a decisiones que puede *actuar* en un historial; el consenso de revisores es uso en colaboración con clínicos.

¿Qué solidez tienen las pruebas?

Hay que dividir la afirmación.

Como demostración de capacidad, que un agente de modelo de lenguaje pueda llevar un caso clínico entero dentro de un EHR en sandbox, encadenando historia, pruebas, diagnóstico y órdenes, el

trabajo es realmente nuevo y razonablemente convincente. Esa es la parte que merece atención.

Como afirmación de superioridad, que MIRA sea *mejor que los médicos*, la evidencia esta acotada. Vale en un benchmark retrospectivo específico, con ocho diagnósticos, asimetría de información, clave de respuesta del historial historico y posibilidad de contaminación. Basta para decir que el agente rindió de forma impresionante en este benchmark. No basta para decir que sea mejor diagnosticador que un médico.

La postura útil no es "la IA bate médicos" ni "solo un juguete". Es: una nueva clase de IA médica, una que *actúa* a través del historial en vez de responder preguntas, lo hizo bien en un benchmark retrospectivo difícil y ahora debe probarse donde aún no se probó: pacientes reales, indiferenciados, prospectivamente y bajo gobernanza.

Por qué importa

Durante unos años, la pregunta interesante en IA médica cambio. Antes era "puede un modelo acertar el diagnóstico?", y los modelos respondían que si en conjuntos cada vez más limpios. MIRA marca el paso a una pregunta más dura: "puede un modelo *hacer el trabajo*: recoger, pedir, interpretar, decidir, actuar, a través de un flujo completo?" Esa es una pregunta más útil y honesta, porque actuar es donde viven tanto la dificultad como el riesgo.

Por eso importa el marco. El reflejo de titular, *la IA supera a los médicos*, apunta a la parte menos novedosa y menos sostenida. Lo nuevo es más pequeño y más importante: un agente que se mueve por todo el historial. Si esa capacidad se sostiene, cambia el **flujo de trabajo** mucho antes de cambiar quién manda. El médico no queda reemplazado; pedir, interpretar y perseguir resultados es donde primero aterriza una herramienta así.

Y coloca la carga de la prueba en la dirección correcta. Una victoria de ranking en casos retrospectivos es razón para hacer un ensayo prospectivo, no sustituto de el. La lectura honesta de MIRA es una invitacion a ese siguiente estudio, medido en pacientes, no en benchmarks.

En pocas palabras

MIRA es un agente de IA autónomo que opera una historia clínica electrónica en sandbox: puede tomar historia, pedir e interpretar laboratorio, imagen y microbiología, llegar a un diagnóstico y escribir planes de tratamiento. Probado en 574 casos retrospectivos de MIMIC-IV, en ocho diagnósticos preseleccionados, superó a médicos en precisión diagnóstica y tomó decisiones en gran parte acordes con guías. Pero la evaluación fue simulada, sobre historiales pasados, solo texto; buena parte de la ventaja vino de condiciones con pruebas claras; MIRA usó muchos más análisis de sangre; varias métricas se puntuaron contra lo registrado; y la base pública plantea riesgo de contaminación, que los autores llaman posible cota superior. El avance real es un agente que actúa a través de todo el flujo. No prueba que la IA diagnostique mejor que médicos ni que esté lista para una clínica real.

Sin rodeos

Lo que muestra el artículo: Un agente de modelo de lenguaje (MIRA) puede llevar un caso clínico completo dentro de un EHR en sandbox - historia, pruebas, diagnóstico, órdenes - y, en un benchmark retrospectivo de 574 casos y ocho diagnósticos, superar a médicos en precisión diagnóstica sin errores medicamentosos de alta severidad reportados.

Lo plausible pero no probado: Qué esa capacidad beneficie a pacientes reales e indiferenciados; que la ventaja refleje mejor razonamiento y no más pruebas, acuerdo con el historial o datos memorizados.

Lo que no muestra: Funcionamiento fuera de ocho diagnósticos; seguridad de despliegue; comparación justa con información igual; reemplazo de clínicos; ausencia de contaminación.

Limitaciones principales: Evaluacion retrospectiva, simulada, solo texto; ocho diagnósticos; asimetría de información; resultados de tratamiento parcialmente contra historia histórica; benchmark público que el modelo pudo ver.

¿Cuánta confianza debería tener un lector general? Alta en que MIRA es una demostración real y nueva de capacidad: un agente que *actúa* en el historial, no solo un chatbot. Baja en que sea *mejor diagnosticador que un médico*. La conclusión correcta de los autores es la adecuada: hacen falta estudios prospectivos reales antes de que esto signifique algo para la atención al paciente.

Fuentes

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