



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
· 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 medica tratan de un modelo que *responde*. Le das una viñeta o una pregunta de examen, devuelve un diagnostico o un parrafo de consejo y puntua bien. Util, pero estrecho: un consultor listo que nunca toca la historia clinica.

MIRA esta 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 microbiologia, lee los resultados, estrecha el diagnostico diferencial y luego escribe las ordenes que siguen: prescripciones, ingreso, derivacion quirurgica. Lo hace tomando acciones *dentro* de una historia clinica electronica, como un clinico, no produciendo texto libre para que un humano lo transcriba.

Es un paso real: de un sistema que aconseja a un sistema que actua a traves del flujo de trabajo. Tambien es exactamente el tipo de paso que la cobertura aplanan en “la IA supera a los medicos”. Por eso conviene ser precisos sobre que se probó y donde.

Version de una frase: MIRA trabajo casos completos por su cuenta y, en este benchmark, supero a medicos en precision diagnostica; pero lo hizo en un sandbox, con historiales retrospectivos, en ocho diagnosticos preseleccionados, comunicandose solo por texto, y obtuvo buena parte de su ventaja en condiciones con pruebas claras. El avance es real. El marcador no es la clinica.

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 demostro una capacidad de flujo de trabajo extremo a extremo dentro de una historia clinica electronica en sandbox. No demostro despliegue clinico real, cobertura diagnostica amplia ni una comparacion justa e igualmente informada con medicos.

Que hicieron los autores

MIRA - Medical Intelligence for Reasoning and Action - es un agente autonomo construido sobre modelos de OpenAI: la parte que conversa y actua usa **GPT-4o**, y un paso de planificacion separado usa el modelo de razonamiento **o1**. Estan dentro de un marco custom conforme a estandares, basado en HL7 FHIR, el estandar de datos que usan hospitales reales para mover historiales, con una caja de herramientas de mas de ochenta mil acciones clinicas posibles. En ese sandbox, MIRA puede extraer la historia, pedir e interpretar analisis, imagen y microbiologia, generar un diagnostico diferencial y formular planes de tratamiento: prescribir medicamentos, programar cirugias, planificar ingresos. Detalle a marcar: los “jueces” automaticos que puntuaron a MIRA eran tambien GPT-4o, la misma familia de modelos probada, respaldados por revision de un medico certificado tras la preocupacion de un revisor.

El conjunto de prueba vino de **MIMIC-IV**, una gran base publica de historias clinicas 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 diagnosticos objetivo**: patologias abdominales y urgencias de medicina interna, como apendicitis, pancreatitis, neumonia e infeccion urinaria. En cada caso, MIRA empezaba con una imagen limitada y debia decidir paso a paso que hacer, la forma de un estudio real, pero sobre un historial cuyo final ya estaba registrado.

Luego compararon su rendimiento con medicos en los mismos casos.

Que significa realmente “agente autonomo en una historia clinica electronica 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 accion (pedir esta prueba, preguntar aquella historia), ve el resultado, elige la siguiente, hasta llegar a diagnostico 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 clinica, 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 mas: MIRA puede pedir cualquier prueba, pero el sistema solo devuelve resultado si esa prueba se hizo **realmente** al paciente. Si pide un escaner que nadie pidio, recibe “N/A”, no un numero 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 tomo.

La distincion importa porque la palabra de titular, “autonomo”, es la que mas facilmente se lee como la segunda cosa.

Que encontraron

En el benchmark, **MIRA supero a los medicos en precision diagnostica**, pero el titular mezcla tres numeros. Por su cuenta, en los **574 casos**, MIRA nombro el diagnostico correcto **88,9 %** de las veces, frente al diagnostico de alta registrado. Para la comparacion humana, los medicos 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 comparo con dos grupos. El primero, **senior**, eran cuatro medicos certificados con 7 a 11 anos de experiencia, que marcaron **78,1 %**. El segundo, de **senioridad mixta**, tenia sobre todo residentes junior mas dos especialistas, mas parecido a una urgencia real, y marco **71,1 %**. Mismos casos, mismas herramientas: IA primero, medicos experimentados despues, equipo joven tercero.

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

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 medicos - estaba **impulsada sobre todo por condiciones con pruebas claras**, como apendicitis y pancreatitis, donde un escaner o valor de laboratorio decide. Para neumonia e infeccion urinaria, dos motivos muy comunes de urgencias, tanto IA como medicos rindieron peor y la brecha fue menor.

Tambien habia una asimetria de informacion. MIRA uso mas laboratorio: cerca del **51 %** de los analitos disponibles en atencion rutinaria, frente a alrededor del **28 %** para los medicos certificados, una mediana de siete parametros sanguineos mas por caso. Los autores lo enmarcan como *por debajo* de la linea base rutinaria del dataset, no como pedirlo todo, y no reportan aumento sistematico de imagen transversal costosa. Aun asi, mas informacion por si sola puede producir mas precision; no es un duelo exactamente igual de informado.

Por que “batir medicos” no es “mejor medico”

Un triunfo de benchmark se sobrelee con facilidad. Tres cosas separan “MIRA puntuo mas” de “MIRA es mejor”.

Primero, **contra que se puntuo**. Julie Jacko señalo que varios resultados de MIRA se definen respecto a lo documentado en la base subyacente: se recompensa al sistema por **reproducir la conducta clinica registrada**, no necesariamente por demostrar atencion optima. La historia historica se vuelve la clave de respuestas. Es razonable para un benchmark, pero mide acuerdo con lo hecho, no correccion absoluta. En justicia, esto golpea mas a las metricas de tratamiento; la precision diagnostica principal se puntua contra el diagnostico 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 corrio**. 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. Generalización, 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 esté 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.

Que tan fuerte es la evidencia?

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 ordenes, el

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

Como afirmacion de superioridad, que MIRA sea *mejor que los medicos*, la evidencia esta acotada. Vale en un benchmark retrospectivo especifico, con ocho diagnosticos, asimetria de informacion, clave de respuesta del historial historico y posibilidad de contaminacion. Basta para decir que el agente rindio de forma impresionante en este benchmark. No basta para decir que sea mejor diagnosticador que un medico.

La postura util no es “la IA bate medicos” ni “solo un juguete”. Es: una nueva clase de IA medica, una que *actua* a traves del historial en vez de responder preguntas, lo hizo bien en un benchmark retrospectivo dificil y ahora debe probarse donde aun no se probó: pacientes reales, indiferenciados, prospectivamente y bajo gobernanza.

Por que importa

Durante unos anos, la pregunta interesante en IA medica cambio. Antes era “puede un modelo acertar el diagnostico?”, y los modelos respondian que si en conjuntos cada vez mas limpios. MIRA marca el paso a una pregunta mas dura: “puede un modelo *hacer el trabajo*: recoger, pedir, interpretar, decidir, actuar, a traves de un flujo completo?” Esa es una pregunta mas util 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 medicos*, apunta a la parte menos novedosa y menos sostenida. Lo nuevo es mas pequeno y mas importante: un agente que se mueve por todo el historial. Si esa capacidad se sostiene, cambia el **flujo de trabajo** mucho antes de cambiar quien manda. El medico no queda reemplazado; pedir, interpretar y perseguir resultados es donde aterriza primero una herramienta asi.

Y coloca la carga de la prueba en la direccion correcta. Una victoria de ranking en casos retrospectivos es razon 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.

Resumen limpio

MIRA es un agente de IA autonomo que opera una historia clinica electronica en sandbox: puede tomar historia, pedir e interpretar laboratorio, imagen y microbiologia, llegar a un diagnostico y escribir planes de tratamiento. Probado en 574 casos retrospectivos de MIMIC-IV, en ocho diagnosticos preseleccionados, supero a medicos en precision diagnostica y tomo decisiones en gran parte acordes con guías. Pero la evaluacion fue simulada, sobre historiales pasados, solo texto; buena parte de la ventaja vino de condiciones con pruebas claras; MIRA uso muchos mas analisis de sangre; varias metricas se puntuaron contra lo registrado; y la base publica plantea riesgo de contaminacion, que los autores llaman posible cota superior. El avance real es un agente que actua a traves de todo el flujo. No prueba que la IA diagnostique mejor que medicos ni que este lista para una clinica real.

No-BS check

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, ordenes - 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: Que 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: Evaluación retrospectiva, simulada, solo texto; ocho diagnósticos; asimetría de información; resultados de tratamiento parcialmente contra historia histórica; benchmark público que el modelo pudo ver.

Cuanta 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.

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-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>