The Clean Paper

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

一個 AI 代理在模擬器中獨立處理了完整患者病例——基於歷史病歷

MIRA 是一種新型醫療 *AI*：它不是回答單一問題，而是在模擬醫院病歷中處理整個病例——採集病史、開具並解讀檢查、做出診斷、書寫醫囑。在來自公共資料庫的 *574* 個回顧性病例上，橫跨八種預選診斷，作者報告它在診斷準確率上超越了醫生，並做出了基本符合指南、用藥安全的決定。這句話中的每一個限定詞都承載分量：它在沙盒中運行，基於歷史病歷，僅限文本；其優勢大多來自檢驗結果清晰的疾病；它開具的血液檢查大約是醫生的兩倍；若干結局是以原始病歷記錄的內容為標準來打分的。真正的進步是一個橫跨整個工作流程的代理——而非證明一台機器現在診斷得比醫生更好。作者自己承認：泛化能力、安全性和治理仍需要前瞻性真實世界研究。

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

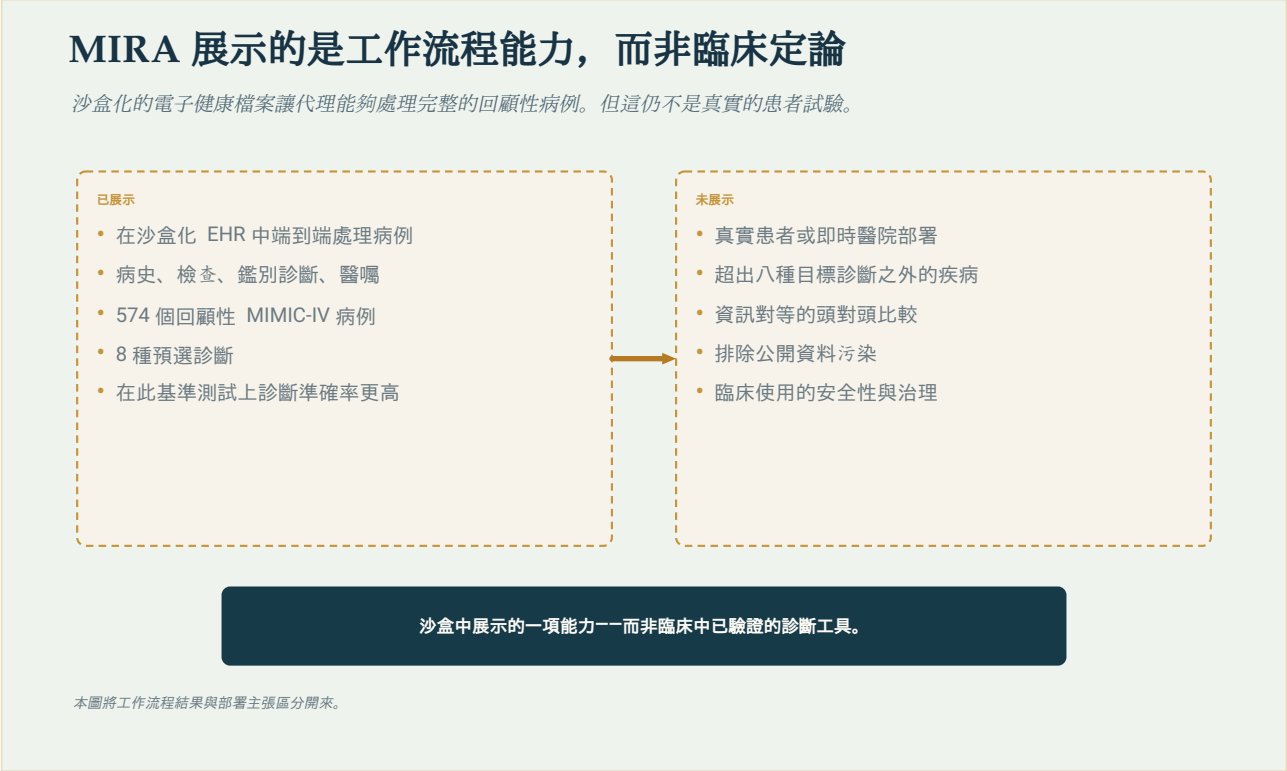
回答問題不等於處理整個病例

多數醫療 AI 的報導，談的是一個會回答問題的模型。你給它一段病例摘要或一道考題，它回傳診斷或一段建議，而且得分不錯。這很有用，但範圍很窄：就像一位從不接觸病歷的聰明會診顧問。

MIRA 的設計目標不同，而這項差異才是真正的新聞。它不是回答一個問題，而是處理完整病例：讀取患者病歷，判斷還需要哪些病史，開立實驗室、影像與微生物學檢查，讀取結果，縮小鑑別診斷範圍，接著開立後續醫囑——處方、住院醫囑、外科轉診。它透過在電子健康紀錄內採取行動來完成這些工作，就像臨床醫師一樣，而不是產生文字讓人抄錄。

這確實是實質的一步：從提供建議的系統，走向能跨越整個工作流程採取行動的系統。但這也正是報導中很容易被簡化成「AI 勝過醫師」的那種進展。因此，有必要準確說明測試了什麼，以及在哪裡測試。

一句話版本：MIRA 能自行處理完整病例，而且在這項基準測試中，診斷準確度勝過醫師——但它是在沙盒中、使用回溯性病歷、針對預先選定的八種診斷、僅以文字溝通完成的；它相當一部分的優勢，來自檢查結果最明確的病症。這項進展是真實的。這個排行榜不等於診所。



MIRA 在沙盒化 EHR 內展示了端到端的工作流程能力。它沒有展示真實世界的臨床部署、廣泛的診斷涵蓋範圍，也沒有展示一場雙方掌握同等資訊、對醫師公平的正面比較。

作者做了什麼

MIRA——Medical Intelligence for Reasoning and Action——是一個建立在 OpenAI 模型上的自主代理：負責對話與行動的部分執行於 **GPT-4o**，另有一個獨立的規劃步驟使用 OpenAI 的 **o1** 推理模型。這些模型位於一個客製化且符合標準的框架內——以 HL7 FHIR 為基礎，也就是現實中的醫院用來傳遞病歷的資料標準——並配備超過八萬種可能的臨床行動工具。在這個沙盒中，MIRA 可以調閱患者病史，開立並解讀實驗室、影像與微生物學檢查，產生鑑別診斷，以及擬定治療計畫——開立藥物、安排手術、規劃住院。有一點值得先指出：為 MIRA 的答案評分的自動化「裁判」本身就是 GPT-4o——也就是受測模型所屬的同一模型家族；在一位同儕審查者提出這項疑慮後，作者又請一位具專科認證的醫師進行覆核。

測試集來自 **MIMIC-IV**，這是一個大型、公開且去識別化的 EHR 資料庫。作者從 2008 年至 2019 年間在 Beth Israel Deaconess Medical Center 接受治療的約 300,000 名患者中，選出涵蓋 **八種目標診斷** 的 **574 個患者病例**——包括腹部病變與內科急症，其中有闌尾炎、胰臟炎、肺炎和泌尿道感染。對每個病例，MIRA 都從有限的資訊開始，必須逐步判斷下一步該做什麼——形式上和真實診療流程相同，但它面對的是一份真實結局已經記錄在案的病歷。

接著，研究者將它在相同病例上的表現與醫師比較。

> **「沙盒化 EHR 中的自主代理」到底是什麼意思**
>
> 這三個詞在這裡承擔了很多含義，值得拆開說明。
> **代理**表示這個系統不是一個只回答提示的聊天機器人。它會執行一個循環：查看目前狀態，選擇一項行動（開這項檢查、詢問那段病史），查看結果，再選擇下一項行動——直到得到診斷與計畫。「智能」是語言模型；代理能力則來自模型周圍的支架，將模型輸出的文字轉換成獲准執行的 EHR 操作，再把結果回饋給模型。
> **沙盒化 EHR** 表示一份模擬且隔離的病歷系統副本，而不是正在運作的醫院系統。MIRA 可以「開立」檢查，並收到那位患者實際接受過的結果，因為病例是歷史病例，答案早已記錄在案。它做的任何事都不會觸及真實患者或真實診所。
> **自主**表示它能在沒有人工介入的情況下，端到端完成病例——但只是在沙盒內。這**不**表示它被放任在無人監督下管理患者。這是兩種非常不同的主張，而測試的只有前者。
> 還有一點關於沙盒的細節值得說明，因為很容易產生錯誤想像：MIRA 可以開立它想要的任何檢查，但系統只有在該檢查在現實中**確實為這名患者執行過**時，才能回傳結果。要求一項患者接受過的血液檢查，你會得到真實數值；要求一項沒有人開過的掃描，你得到的會是「N/A——無法執行」，絕不會是捏造的數字。病史也是一樣：如果病歷沒有 MIRA 詢問的內容，模擬患者就只會說自己不知道。因此，MIRA 無法憑空取得一項從未做過的決定性檢查——它只能利用真實診療流程碰巧包含的內容。
> 這個區別很重要，因為標題中的那個詞——「自主」——最容易被理解成後一種意思。

他們發現了什麼

在這項基準測試中，**MIRA 的診斷準確度勝過醫師**——但標題藏起了三個不同的數字，而且很容易混在一起，所以值得分開看。

在單獨評估中，MIRA 涵蓋全部 **574 個病例**，有 **88.9%** 的時間指出了正確診斷——評分依據是 MIMIC-IV 中實際為每名患者記錄的出院診斷。至於與人類的正面比較，醫師並沒有處理全部 574 個病例；他們處理的是一個共用的 **311 個病例子集**，使用 MIRA 擁有的相同病歷與工具。在這 311 個相同病例上，MIRA 的得分是 **87.8%**，而且研究者將它與兩組分別招募的醫師比較。第一組是**資深組**：四位具專科認證、擁有 7 至 11 年經驗的醫師，得分 **78.1%**。第二組是**資歷混合組**：主要由初級住院醫師加上兩位專科醫師組成，更接近真實急診科的配置，得分 **71.1%**。相同病例、相同工具——結果依序是 AI 第一、資深醫師第二、以初級醫師為主的團隊第三。論文自身謹慎的措辭是：在全部八種疾病中，MIRA「始終不遜於醫師，且經常超越」醫師。

它後續的決策也經得起檢驗：大致符合指引、用藥安全，且住院處置適當。作者報告，在五類安全性（藥物交互作用、腎功能劑量、過敏、QT 風險與鴉片類藥物處方）中，沒有高嚴重度用藥錯誤；468 個病例中有 467 個病例的處方正確——同時也指出，系統「未達到 100% 的可靠性」。若照字面理解，這是個引人注目的結果：一個代理處理完整病例，而且在診斷上勝過醫師。

但勝出的方式和勝利本身一樣重要。閱讀論文的獨立專科人士指出，優勢實際上從何而來。在 Science Media Centre 的專家小組 中，謝偉醫師（雪菲爾大學）指出，標題數字——AI 在診斷準確度上勝過醫師——**主要由檢查結果明確的病症所推動**，例如闌尾炎和胰臟炎；在這些情況下，一項決定性的掃描或實驗室數值就能解決問題。至於肺炎和泌尿道感染——也是人們實際前往急診最常見的兩個原因——*AI* 與醫師的表現都是最差，兩者之間的差距也最小。

雙方的操作方式也存在不對稱，而這點就寫在論文自己的數字中。MIRA 更依賴實驗室檢查：它使用了常規照護中可取得的約 **51%** 實驗室分析物，相較之下，具專科認證的醫師約使用 **28%**——每個病例的中位數多出七項血液參數。作者謹慎地將此描述為低於資料集自身的常規照護基準，而不是「什麼都開」的策略，並報告沒有系統性增加成本較高的橫斷面影像檢查。不過，更多資訊本身就可能帶來更高的診斷準確度——因此，這並不是一場資訊同等充分的選手之間完全對等的競賽，謝醫師也直接指出了這項差距。一位可以自由開立每項便宜血液檢查、而不必權衡成本、不適或延誤的臨床醫師，在紙面上也會顯得更敏銳。

為什麼「勝過醫師」不等於「更好的醫師」

基準測試的勝利很容易被過度解讀。在「MIRA 得分更高」與「MIRA 更好」之間，隔著三件事。

第一，**它是相對於什麼來評分的**。 Julie Jacko 教授（愛丁堡大學）指出，MIRA 的幾項關鍵結果，是相對於底層資料集中所記錄的內容定義的——這表示系統得到獎勵，是因為**重現已記錄的臨床行為**，不一定是因為展示了最佳照護。歷史病歷成了答案卷。這是建構基準測試的合理方式，但它衡量的是與既有處置的一致性，而不是某種絕對意義上的正確性。公平地說，這個問題對治療一致性指標的影響最大——MIRA 的醫囑是否符合病歷？——而標題中的診斷準確度則是以出院診斷為評分依據，比單純重複文件記載更接近真實結果。

第二，就是前面提到的**資訊不對稱**：血液檢查多得多（可用分析物約 51%，相較之下為 28%）就代表更多證據。準確度差距的一部分，可能是用檢查買來的，而不是推理出來的。

第三，**它在哪裡運作**。這是一個沙盒，使用回溯性病例，涵蓋八種預先選定的診斷，而且只使用文字。正如 Dominic Oliver 醫師（牛津大學）所說，真實的臨床評估不只取決於患者說了什麼，也取決於他們怎麼說——還包括身體檢查、觀察到的行為與肢體語言；這些都是只讀取一份已完成病歷的純

文字代理永遠看不到的。而患者的問題若不屬於八種選定診斷之一，也就完全超出這項研究能夠說明的範圍。

審查者還提出一項較不顯眼的疑慮：**資料污染**。MIMIC-IV 是公開資料，且有大量相關文章，因此，以開放網際網路資料訓練的語言模型可能早已看過相關論文、病例討論，甚至資料本身。Midhun Parakkal Unni 醫師（雪菲爾大學）直接指出：如果部分答案已存在於訓練資料中，那麼部分表現來自記憶，而不是臨床推理；只有獨立複製研究才能區分兩者。值得注意的是，作者並未含糊帶過這一點：他們寫道，結果「可以謹慎解讀為可能的上限」，並且「可能高估對其他公開病例的泛化能力」——這是一篇替自己的標題數字設下天花板的論文。

這些都不會讓 MIRA 變得不那麼有趣。它們只代表正確的解讀應該有所節制：在回溯性基準測試中，一個能跨越整份病歷採取行動的代理，在檢查結果明確的診斷上勝過醫師——一部分是因為開了更多檢查，一部分是因為符合已記錄的病歷。這是真實的能力展示，而不是機器現在診斷得比醫師好的判決。

這項研究**沒有**證明什麼

- 它**沒有**顯示 MIRA 能在真實患者身上運作。每個病例都是回溯性且經過模擬；沒有任何患者由它實際管理。
- 它**沒有**顯示 MIRA 能處理八種診斷以外的情況。預先選定集合之外的病症——醫療中混亂、尚未分化的多數情況——並未接受測試。
- 它**沒有**顯示 MIRA 適合安全部署。作者自己也寫道，泛化能力、安全性與治理仍需要前瞻性的真實世界研究。
- 它**沒有**建立與醫師公平的正面比較。它使用了多得多的血液檢查（可用分析物約 51%，相較之下為 28%），而且幾項治療結果的評分，部分是依據已記錄的病歷，而非依據真正的最佳照護。
- 它**沒有**顯示這項結果排除了記憶作用。公開的 MIMIC-IV 資料可能與模型訓練資料重疊，這需要獨立複製研究加以排除。
- 它**不**代表「AI 取代醫師」。它是一種能跨越整份病歷採取行動的決策支援工具；審查者的共識是，實際使用時將與臨床醫師合作，由臨床醫師保有決策權，並補足文字病歷無法提供的一切。

證據有多強？

把這項主張拆成兩半，因為兩半的證據很不一樣。

作為能力展示——也就是語言模型代理能在 EHR 沙盒中處理完整臨床病例，端到端串起病史、檢查、診斷與醫囑——這項工作確實新穎，也相當令人信服。這是其中真正嶄新的部分，也是值得關注的部分。

作為優越性主張——也就是 MIRA 比醫師更好——證據有明確邊界，應該謹慎解讀。它的結果成立於一項特定的回溯性基準測試、八種診斷，以及資訊不對稱（檢查更多）和取自歷史病歷的答案卷之上；訓練資料污染也確實可能存在。這足以說「代理在這項基準測試中表現出色」。但不足以說「代理是比醫師更好的診斷者」，而且作者並未主張後者。

最有用的立場既不是「AI 勝過醫師」，也不是「這只是玩具」。更準確的說法是：一種新型態的醫療 AI——不是回答零散問題，而是跨越整份病歷採取行動——在一項艱難的回溯性基準測試中表現良好；接下來，它必須在尚未被測試的地方證明自己：面對真實、未分化的患者，以前瞻方式、在治理框架下運作。

為什麼重要

過去幾年，醫療 AI 中真正有趣的問題悄悄改變了。過去的問題是「模型能不能診斷正確？」——而模型在越來越乾淨的測試集上，不斷回答可以。MIRA 將問題推向更難的一層：「模型能不能完成工作——在整個工作流程中蒐集資訊、開立檢查、解讀結果、作出決定、採取行動？」這是更有用、也更誠實的問題，因為困難與風險都存在於行動之中。

這就是如何界定這項成果很重要的原因。標題式的直覺反應——*AI* 勝過醫師——指向結果中最不新穎、也最缺乏支持的部分。真正的新事物更小，卻更具後果：一個能貫穿整份病歷採取行動的代理。如果這項能力經得起考驗，它首先改變的是**工作流程**，遠早於改變由誰掌控工作流程。醫師並沒有被取代；開立檢查、解讀結果、追蹤結果——也就是工作流程——才是這類工具最先落地的地方。

它也把舉證責任重新導向正確的方向。在回溯性病例排行榜上的勝利，是展開前瞻性試驗的理由，而不是試驗的替代品。作者也如此表示。對 MIRA 最誠實的解讀，是邀請人們進行下一項研究——接受這個領域早已熟悉的標準檢驗：在患者身上衡量，而不是在基準測試上衡量。

重點摘要

MIRA 是一個自主 AI 代理，能操作沙盒化的電子健康紀錄：它可以詢問病史、開立並解讀實驗室、影像與微生物學檢查，得到診斷，並寫下治療計畫。研究者使用公開 MIMIC-IV 資料庫中的 574 個回溯性病例進行測試，涵蓋八種預先選定的診斷；MIRA 的診斷準確度勝過醫師（整體 88.9%；正面比較為 87.8% 對 78.1%），而且做出的決策大致符合指引並具用藥安全性。但這項評估是在過去病歷上的模擬，而且僅使用文字；它相當一部分的優勢來自檢查結果明確的病症；MIRA 使用的血液檢查遠多於醫師（可用分析物約 51%，相較之下為 28%）；幾項治療結果是依據原始病歷的記載評分；公開資料集也帶來真實的訓練資料污染風險——作者自己稱這可能是其數字的上限。真正的進展，是出現了一個能跨越整個工作流程採取行動、而不是只回答孤立問題的代理。作者明確表示，泛化能力、安全性與治理仍需要前瞻性的真實世界研究——這才是正確的解讀：一項令人印象深刻的能力展示，而不是 AI 診斷得比醫師好的證明，也不是已準備好進入真實診所的系統。

不講套話的檢驗

論文展示了什麼：一個語言模型代理（MIRA）能在沙盒化 EHR 內端到端處理完整臨床病例——病史、檢查、診斷、醫囑——而且在涵蓋八種診斷的 574 個病例回溯性基準測試中，診斷準確度勝過醫師（整體 88.9%；與具專科認證的醫師比較為 87.8% 對 78.1%），並且沒有高嚴重度用藥錯誤。

合理但尚未證明的事：這項能力能為真實、未分化的患者帶來好處；準確度優勢反映的是更好的推理，而不是開了更多檢查、符合已記錄的病歷，或記住了公開資料。

它沒有展示什麼： MIRA 能在八種診斷以外運作；它適合安全部署；它在公平且資訊同等充分的比較中勝過醫師；它能取代臨床醫師；結果排除了訓練資料污染。

主要限制：回溯性、模擬、純文字的評估；八種預先選定的診斷；資訊不對稱（血液檢查多得多——可用分析物約 51%，相較之下為 28%，而且是低成本血液檢查，不是影像檢查）；治療結果部分依據歷史病歷評分；以及一項語言模型可能在訓練中見過的公開基準測試（MIMIC-IV）——作者指出，這可能是表現的上限。

一般讀者應該有多大信心？對於 MIRA 是一項真實且新穎的能力展示——一個能跨越病歷採取行動、而不只是聊天機器人的代理——應有高度信心。至於它是比醫師更好的診斷者，信心應該很低：這項主張受限於一項回溯性基準測試，並受到開立檢查、依病歷評分以及可能的資料污染所混淆。作者自己的結論是正確的——在它對患者照護具有任何意義之前，仍需要前瞻性的真實世界研究。

來源

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