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 · 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

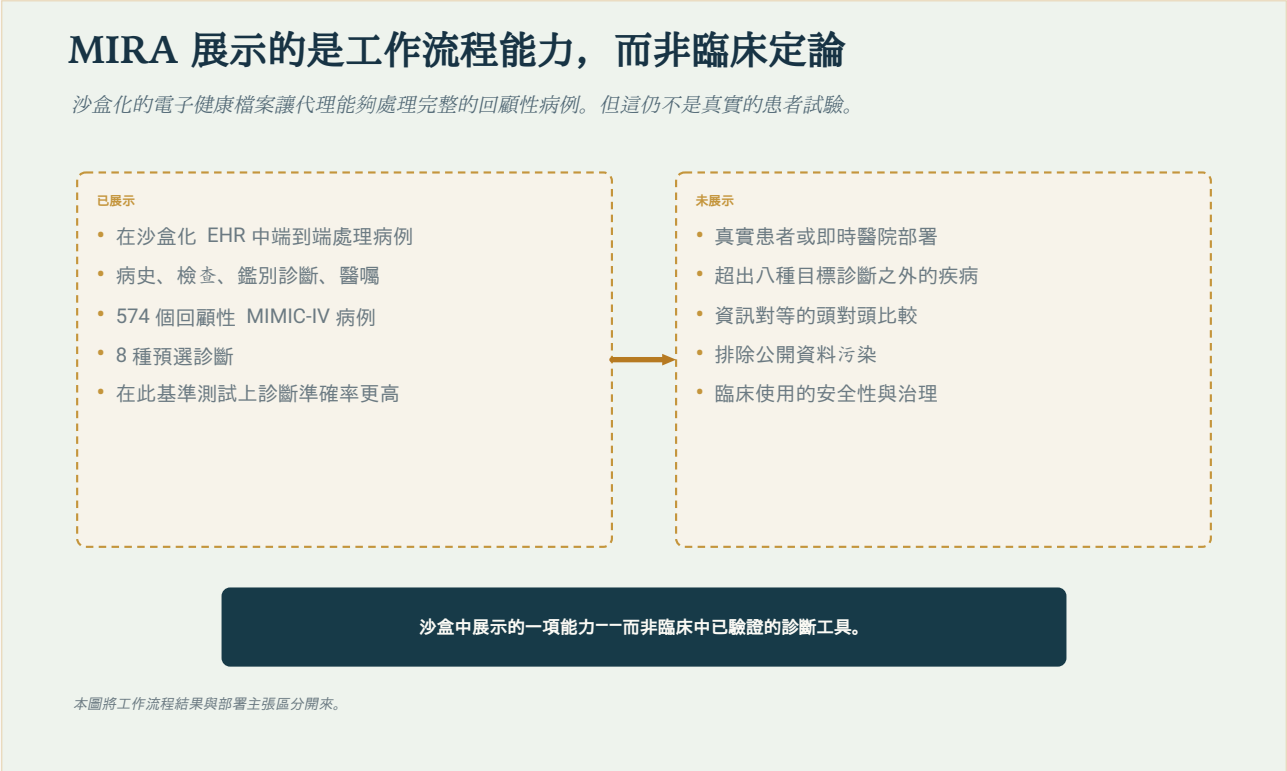
回答問題不等於處理病例

大多數醫療 AI 的故事都是關於一個回答的模型。你給它一份病例簡述或考題，它返回一個診斷或一段建議，得分很高。有用，但狹隘：一個從不碰病歷的聰明顧問。

MIRA 被構建來做不同的事，而這一區別才是真正的新聞。它不回答單一問題，而是處理整個病例：讀取患者病歷，決定還需要哪些病史，開具實驗室、影像和微生物檢查，讀取結果，縮小鑑別診斷範圍，然後書寫後續醫囑——處方、入院、轉診手術。它通過在電子健康檔案內部採取行動來實現——像臨床醫生那樣——而不是生成供人類轉錄的自由文本。

這是真正的一步：從提供建議的系統到橫跨整個工作流程的行動系統。而這正是那種在報導中被簡化為「AI 超越醫生」的步驟。因此值得精確說明測試了什麼，以及在哪裡。

一句話版本：MIRA 獨立處理了整個病例，在這一基準測試上診斷準確率超越了醫生——但它是在沙盒中、基於回顧性病歷、橫跨八種預選診斷、僅通過文本交流實現的，且其優勢主要體現在檢驗結果最清晰的疾病上。這一進步是真實的。但計分板不是診所。



MIRA 在沙盒化的電子健康檔案中展示了端到端的工作流程能力。它並未展示真實世界的臨床部署、廣泛的診斷覆蓋範圍，或與醫生公平、資訊對等的正面對比。

作者做了什麼

MIRA——Medical Intelligence for Reasoning and Action——是一個基於 OpenAI 模型構建的自主代理：負責對話和行動的部分運行在 **GPT-4o** 上，一個獨立的規劃步驟使用 OpenAI 的 **o1** 推理模型。這些位於一個客製的、符合標準的框架內——基於 HL7 FHIR，即真實醫院用於傳輸病歷的資料標準——配備一個包含超過八萬種可能臨床行動的工具箱。在該沙盒內，MIRA 可以調取患者病史，開具並解讀實驗室、影像和微生物檢查，生成鑑別診斷，並制定治療計劃——開具處方、安排外科手術、計劃入院。有一個細節值得前置說明：對 MIRA 答案進行評分的自動「裁判」本身是 GPT-4o——與被測試的模型屬於同一模型家族——作者在以一名經過委員會認證的醫生進行覆核後支持了這一做法，因為一位同儕審查人提出了這一關切。

測試集來自 **MIMIC-IV**，一個大型的、公開可取得的去識別化電子健康檔案資料庫。從 2008 年至 2019 年間在貝斯以色列女執事醫療中心接受治療的大約 30 萬名患者中，作者選取了 **574 個患者病例**，橫跨 **8 種目標診斷**——腹部病變和內科急症，其中包括闌尾炎、胰腺炎、肺炎和尿路感染。對於每個病例，MIRA 從有限的圖景中開始，必須一步步決定下一步做什麼——這與真實檢查的形態相同，但運行在結果早已記錄在案的病歷之上。

其表現隨後與醫生在同一批病例上進行了比較。

> **「沙盒化電子健康檔案中的自主代理」到底意味著什麼**
>
> 這裡三個詞承載了很多含義，因此值得拆解。
> **代理**意味著該系統不是一個回答提示詞的聊天機器人。它運行一個迴圈：查看當前狀態，選擇一個行動（開這個檢查、問那段病史），看到結果，選擇下一個行動——直到得出診斷和計劃。「智慧」是語言模型；能動性是其周圍的鷹架，將模型文本轉化為被允許的電子健康檔案操作，並將結果反饋回來。
> **沙盒化電子健康檔案**意味著一個模擬的、隔離的病歷系統副本，而非真實的醫院系統。MIRA 可以「開具」一項檢查並接收該患者實際獲得的結果，因為病例是回顧性的，答案已經記錄在案。它的任何操作都不會觸及真實患者或真實診所。
> **自主**意味著它在沙盒內端到端地完成病例，無需人類參與——在沙盒內。它**不**意味著它被放任去無監督地管理患者。這是截然不同的主張，只有前者被測試了。
> 還有一點關於沙盒，因為它很容易被想像錯誤：MIRA 可以「開具」它想要的任何檢查，但系統只能在患者在實際中確實接受了該項檢查時返回結果。要求做患者接受過的血液檢查，你得到真實數值；要求做無人開具過的掃描，你得到「N/A——無法執行」，絕不會是編造的數字。病史同理：如果病歷不包含 MIRA 所詢問的內容，模擬患者只會說不知道。因此 MIRA 無法召喚那個從未被安排的決定性檢查——它只能使用真實診療恰好包含的內容。
> 這一區別很重要，因為標題詞——「自主」——最容易被解讀為後者。

他們發現了什麼

在該基準測試上，**MIRA 在診斷準確率上超越了醫生**——但標題隱藏了三個不同的數字，它們很容易被混淆，因此值得分別說明。

獨自運行時，在所有 **574 個病例**中，MIRA 在 **88.9%** 的情況下給出了正確診斷——以 MIMIC-IV 中為每位患者實際記錄的出院診斷為標準。在與人類的頭對頭比較中，醫生並未處理全部 574 個病例；他們處理了一個共享的 **311 病例子集**，使用與 MIRA 相同的病歷和工具。在這相同的 311 個病例上，MIRA 得分為 **87.8%**，並且與兩個分別招募的組進行了比較。第一個是**資深組**：四名擁有 7 至 11 年經驗的經委員會認證醫生，得分為 **78.1%**。第二個是**混合資歷組**：以初級住院醫生為主加上兩名專科醫生，更接近真實急診科的實際人員配置，得分為 **71.1%**。同樣的病例，同樣的工具——排序結果為 AI 第一，經驗豐富的醫生第二，偏初級的團隊第三。論文自身的審慎措辭是 MIRA「始終與醫生相當，且經常超越」，覆蓋全部八種疾病。

其下游決策也經得起檢驗：基本符合指南，用藥安全且入院適當。作者報告在五個安全類別（藥物交互作用、腎臟劑量、過敏、QT 風險和類鴉片處方）中無嚴重用藥錯誤，468 例處方中有 467 例正確——同時指出系統「未達到 100% 的可靠性」。僅從表面來看，這是一個驚人的結果：一個代理處理了整個病例，在診斷上領先於醫生。

但勝利的形態與勝利同等重要。閱讀論文的獨立專家指出了優勢真正來自何處。在科學媒體中心的專家小組上，邢巍博士（謝菲爾德大學）指出，標題數字——AI 在診斷準確率上擊敗醫生——**主要由檢驗結果清晰的疾病驅動**，如闌尾炎和胰腺炎，一次決定性的掃描或化驗值就能解決問題。對於肺炎和尿路感染——人們實際前往急診科的最常見原因之一——雙方的表現都最差，且差距最小。

還有一項不對稱存在於論文自家的數字中。MIRA 更倚重實驗室：它利用了常規診療中可用的大約 **51%** 的實驗室分析物，而經委員會認證的醫生大約為 **28%**——每個病例中位數多了七項血液指標。作者審慎地將這一水準框定為低於資料集自身常規診療基線，而非掃光一切的策略，並報告未系統性增加較高成本的橫斷面影像。然而，更多的資訊本身就能產生更高的診斷準確率——因此這並非資訊均等的玩家之間的完全公平競爭，邢巍直接指出了這一點。一個可以自由開具每一項廉價血液檢查而無需權衡成本、不適或延遲的臨床醫生，在紙面上看起來也會更敏銳。

為什麼「擊敗醫生」不等於「更好的醫生」

基準測試的勝利很容易被過度解讀。三件事橫亙在「MIRA 得分更高」與「MIRA 更優」之間。

第一，**它被衡量的標準**。 Julie Jacko 教授（愛丁堡大學）指出，MIRA 的若干關鍵結局是相對於底層資料集中記錄的內容來定義的——這意味著系統因**復現了被記錄的臨床行為**而獲得獎勵，不一定是展示了最佳診療。歷史病歷變成了答案鍵。這是構建基準測試的合理方式，但它衡量的是與執行內容的一致，而非某種絕對意義上的正確性。公平而言，這一批評對治療對齊指標衝擊最大——MIRA 的醫囑是否與病歷匹配？——而標題診斷準確率是以出院診斷為標準衡量的，這更接近真實結局而非文件回響。

第二，**資訊的不對稱**（前文已述）：多得多的血液檢查（約 51% 對 28% 的分析物）是更多的證據。部分準確率差距很可能是被購買的，而非推理的。

第三，**它運行的環境**。這是沙盒，回顧性病例，八種預選診斷，僅限文本。真實的臨床評估，如 Dominic Oliver 博士（牛津大學）所言，不僅取決於患者說了什麼，還取決於他們怎麼說——伴隨體格檢查、觀察到的行為和身體語言，而這些是一個閱讀已完成病歷的純文本代理永遠看不到的。並且一個疾病不在八種選定診斷之列的患者，根本不在本研究能夠討論的範圍內。

還有一個更微妙的擔憂被審查人提出：**資料污染**。 MIMIC-IV 是公開的且被大量書寫，因此一個在開放網際網路上訓練的語言模型可能已經見過論文、病例討論或資料本身。Midhun Parakkal Unni 博士（謝菲爾德大學）直接指出了這一點——如果部分答案存在於訓練資料中，則部分表現是記憶而非臨床推理，只有獨立復現才能區分兩者。值得注意的是，作者沒有迴避這一點：他們寫道，其結果「可以審慎地解釋為可能的上限」，並且「可能高估了對其他公開病例的泛化能力」——一篇給自己的標題設置上限的論文。

這篇論文沒有證明什麼

- 它**沒有**證明 AI 在真實臨床環境中比醫生更好。這是在沙盒中，基於回顧性資料，僅有八種診斷，僅限文本。
- 它**沒有**展示真實世界的患者結局——沒有死亡率、沒有併發症、沒有患者滿意度。
- 它**沒有**證明更廣泛的診斷能力。一個不在八種診斷之列的病例不在研究範圍。
- 它**沒有**證明自主 AI 在診所中是安全的。說沙盒中無傷害與說它可以被部署是兩回事。
- 它**沒有**完全排除資料污染。作者承認了這一點。

證據有多強？

對於核心主張——在沙盒中，MIRA 能夠處理整個臨床工作流程並在受控回顧性病例上達到高診斷準確率——證據作為能力演示是強有力的。這是一個真實的工程成就：在 FHIR 上運行的代理、超過八萬種行動、跨八種診斷的工作流程。那依然是令人印象深刻的。

但作為效能主張——AI 在診斷上擊敗了醫生——應該更加小心地看待。優勢被集中在了檢驗結果最清晰的疾病上，且 MIRA 比醫生多獲取了大量廉價實驗室資料。差距至少部分是由資訊而非智力驅動的。此外，該論文沒有測量真實患者結局，僅測量了病歷上的準確率。

最誠實的框架：一項令人印象深刻的能力演示，而非一項臨床驗證。

清潔摘要

MIRA 是一個在沙盒化電子健康檔案中運行並獨立處理完整臨床病例的 AI 代理。在 MIMIC-IV 基準測試（574 個病例，8 種診斷）上，它在診斷準確率上超越了醫生——獨自運行時達到 88.9%，在共享的 311 病例子集上為 87.8%，而資深醫生為 78.1%，混合資歷團隊為 71.1%。但其優勢集中在具有決定性強檢驗結果的疾病上，且 MIRA 比人類獲取了更多的實驗室資料。效能主張令人震驚，但沙盒、有限的診斷範圍、資訊不對稱和資料污染風險都限制了泛化能力。這是一個真正的工程步驟，而非 AI 已準備好無監督執業的證明。

No-BS 檢查

論文所展示的：一個 AI 代理能夠端到端地處理臨床工作流程，並在八種預選診斷上超越醫生，在沙盒中，基於回顧性病歷。**合理但未經證實的：**相同的表現會在真實診所中成立；AI 實際上在推理，而非記憶；差距在同類資訊下依然能保持。**它沒有展示的：**真實世界患者結局、廣泛的診斷覆蓋範圍、臨床安全性、或已部署系統的有效性。**主要侷限：**沙盒、回顧性資料、8 種診斷、文本僅限、資訊不對稱、可能的資料污染、MIMIC-IV 的特定性。**普通讀者應該抱有多大信心？**高度信心認為這是一個有能力的代理。中度信心認為診斷擊敗是乾淨的。低度信心認為這是「更好的醫生」。恰當的立場：令人印象深刻的工程技術，而非經過驗證的臨床工具。

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-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