The Clean Paper

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

一个 AI 代理在模拟器中独立处理了完整患者病例——基于历史病历

MIRA 是一种新型医疗 *AI*：它不是回答单一问题，而是在模拟医院病历中处理整个病例——采集病史、开具并解读检查、做出诊断、书写医嘱。在来自公共数据库的 *574* 个回顾性病例上，横跨八种预选诊断，作者报告它在诊断准确率上超越了医生，并做出了基本符合指南、用药安全的决定。这句话中的每一个限定词都承载分量：它在沙盒中运行，基于历史病历，仅限文本；其优势大多来自检验结果清晰的疾病；它开具的血液检查大约是医生的两倍；若干结局是以原始病历记录的内容为标准来打分的。真正的进步是一个横跨整个工作流程的代理——而非证明一台机器现在诊断得比医生更好。作者自己承认：泛化能力、安全性和治理仍需要前瞻性真实世界研究。

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

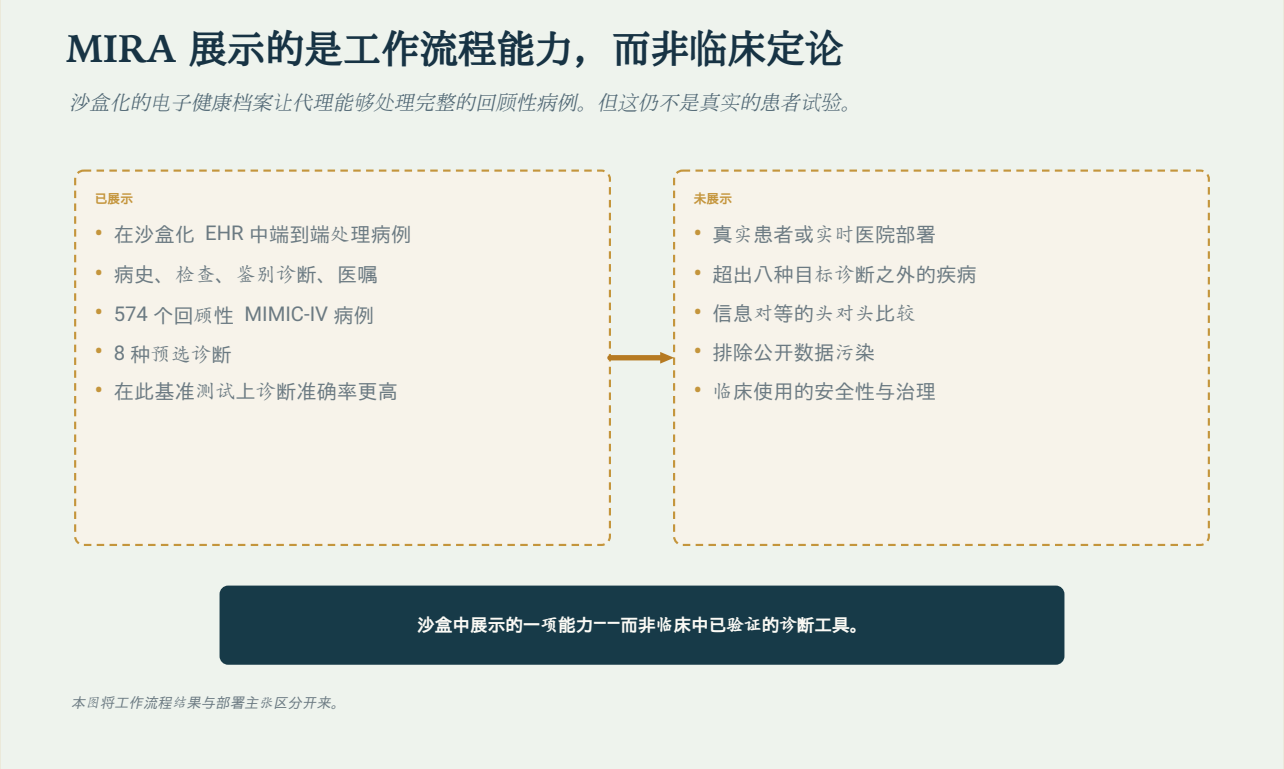
回答问题不等于处理病例

大多数医疗 AI 的故事都是关于一个回答的模型。你给它一份病例简述或考题，它返回一个诊断或一段建议，得分很高。有用，但狭隘：一个从不碰病历的聪明顾问。

MIRA 被构建来做不同的事，而这一区别才是真正的新闻。它不回答单一问题，而是处理整个病例：读取患者病历，决定还需要哪些病史，开具实验室、影像和微生物检查，读取结果，缩小鉴别诊断范围，然后书写后续医嘱——处方、入院、转诊手术。它通过在电子健康档案内部采取行动来实现——像临床医生那样——而不是生成供人类转录的自由文本。

这是真正的一步：从提供建议的系统到横跨整个工作流程的行动系统。而这正是那种在报道中被简化为「AI 超越医生」的步骤。因此值得精确说明测试了什么，以及在哪里。

一句话版本：MIRA 独立处理了整个病例，在这一基准测试上诊断准确率超越了医生——但它是在沙盒中、基于回顾性病历、横跨八种预选诊断、仅通过文本交流实现的，且其优势主要体现在检验结果最清晰的疾病上。这一进步是真实的。但记分板不是诊所。



MIRA 在沙盒化的电子健康档案中展示了端到端的工作流程能力。它并未展示真实世界的临床部署、广泛的诊断覆盖范围，或与医生公平、信息对等的正面对比。

作者做了什么

MIRA——Medical Intelligence for Reasoning and Action——是一个基于 OpenAI 模型构建的自主代理：负责对话和行动的部分运行在 **GPT-4o** 上，一个独立的规划步骤使用 OpenAI 的 **o1** 推理模型。这些位于一个定制的、符合标准的框架内——基于 HL7 FHIR，即真实医院用于传输病历的数据标准——配备一个包含超过八万种可能临床行动的工具箱。在该沙盒内，MIRA 可以调取患者病史，开具并解读实验室、影像和微生物检查，生成鉴别诊断，并制定治疗计划——开具处方、安排外科手术、计划入院。有一个细节值得前置说明：对 MIRA 答案进行评分的自动「裁判」本身是 GPT-4o——与被测试的模型属于同一模型家族——作者在以一名经过委员会认证的医生进行复核后支持了这一做法，因为一位同行评审人提出了这一关切。

测试集来自 **MIMIC-IV**，一个大型的、公开可获取的去标识化电子健康档案数据库。从 2008 年至 2019 年间在贝斯以色列女执事医疗中心接受治疗的大约 30 万名患者中，作者选取了 **574 个患者病例**，横跨 **8 种目标诊断**——腹部病变和内科急症，其中包括阑尾炎、胰腺炎、肺炎和尿路感染。对于每个病例，MIRA 从有限的图景中开始，必须一步步决定下一步做什么——这与真实检查的形态相同，但运行在结果早已记录在案的病历之上。

其表现随后与医生在同一批病例上进行了比较。

> **「沙盒化电子健康档案中的自主代理」到底意味着什么**
>
> 这里三个词承载了很多含义，因此值得拆解。
> **代理**意味着该系统不是一个回答提示词的聊天机器人。它运行一个循环：查看当前状态，选择一个行动（开这个检查、问那段病史），看到结果，选择下一个行动——直到得出诊断和计划。「智能」是语言模型；能动性是其周围的脚手架，将模型文本转化为被允许的电子健康档案操作，并将结果反馈回来。
> **沙盒化电子健康档案**意味着一个模拟的、隔离的病历系统副本，而非真实的医院系统。MIRA 可以「开具」一项检查并接收该患者实际获得的结果，因为病例是回顾性的，答案已经记录在案。它的任何操作都不会触及真实患者或真实诊所。
> **自主**意味着它在沙盒内端到端地完成病例，无需人类参与——在沙盒内。它**不**意味着它被放任去无监督地管理患者。这是截然不同的主张，只有前者被测试了。
> 还有一点关于沙盒，因为它很容易被想象错误：MIRA 可以「开具」它想要的任何检查，但系统只能在患者在实际中确实接受了该项检查时返回结果。要求做患者接受过的血液检查，你得到真实数值；要求做无人开具过的扫描，你得到「N/A——无法执行」，绝不会是编造的数字。病史同理：如果病历不包含 MIRA 所询问的内容，模拟患者只会说不知道。因此 MIRA 无法召唤那个从未被安排的决定性检查——它只能使用真实诊疗恰好包含的内容。
> 这一区别很重要，因为标题词——「自主」——最容易被解读为后者。

他们发现了什么

在该基准测试上，**MIRA 在诊断准确率上超越了医生**——但标题隐藏了三个不同的数字，它们很容易被混淆，因此值得分别说明。

独自运行时，在所有 **574 个病例**中，MIRA 在 **88.9%** 的情况下给出了正确诊断——以 MIMIC-IV 中为每位患者实际记录的出院诊断为标准。在与人类的头对头比较中，医生并未处理全部 574 个病例；他们处理了一个共享的 **311 病例子集**，使用与 MIRA 相同的病历和工具。在这相同的 311 个病例上，MIRA 得分为 **87.8%**，并且与两个分别招募的组进行了比较。第一个是**资深组**：四名拥有 7 至 11 年经验的经委员会认证医生，得分为 **78.1%**。第二个是**混合资历组**：以初级住院医生为主加上两名专科医生，更接近真实急诊科的实际人员配置，得分为 **71.1%**。同样的病例，同样的工具——排序结果为 AI 第一，经验丰富的医生第二，偏初级的团队第三。论文自身的审慎措辞是 MIRA「始终与医生相当，且经常超越」，覆盖全部八种疾病。

其下游决策也经得起检验：基本符合指南，用药安全且入院适当。作者报告在五个安全类别（药物相互作用、肾脏剂量、过敏、QT 风险和类阿片处方）中无严重用药错误，468 例处方中有 467 例正确——同时指出系统「未达到 100% 的可靠性」。仅从表面来看，这是一个惊人的结果：一个代理处理了整个病例，在诊断上领先于医生。

但胜利的形态与胜利同等重要。阅读论文的独立专家指出了优势真正来自何处。在科学媒体中心的专家小组上，邢巍博士（谢菲尔德大学）指出，标题数字——AI 在诊断准确率上击败医生——**主要由检验结果清晰的疾病驱动**，如阑尾炎和胰腺炎，一次决定性的扫描或化验值就能解决问题。对于肺炎和尿路感染——人们实际前往急诊科的最常见原因之一——双方的表现都最差，且差距最小。

还有一项不对称存在于论文自家的数字中。MIRA 更倚重实验室：它利用了常规诊疗中可用的大约 **51%** 的实验室分析物，而经委员会认证的医生大约为 **28%**——每个病例中位数多了七项血液指标。作者审慎地将这一水平框定为低于数据集自身常规诊疗基线，而非扫光一切的策略，并报告未系统性增加较高成本的横断面影像。然而，更多的信息本身就能产生更高的诊断准确率——因此这并非信息均等的玩家之间的完全公平竞争，邢巍直接指出了这一点。一个可以自由开具每一项廉价血液检查而无需权衡成本、不适或延迟的临床医生，在纸面上看起来也会更敏锐。

为什么「击败医生」不等于「更好的医生」

基准测试的胜利很容易被过度解读。三件事横亘在「MIRA 得分更高」与「MIRA 更优」之间。

第一，**它被衡量的标准**。Julie Jacko 教授（爱丁堡大学）指出，MIRA 的若干关键结局是相对于底层数据集中记录的内容来定义的——这意味着系统因**复现了被记录的临床行为**而获得奖励，不一定是展示了最佳诊疗。历史病历变成了答案键。这是构建基准测试的合理方式，但它衡量的是与执行内容的一致，而非某种绝对意义上的正确性。公平而言，这一批评对治疗对齐指标冲击最大——MIRA 的医嘱是否与病历匹配？——而标题诊断准确率是以出院诊断为标准衡量的，这更接近真实结局而非文档回响。

第二，**信息的不对称**（前文已述）：多得多的血液检查（约 51% 对 28% 的分析物）是更多的证据。部分准确率差距很可能是被购买的，而非推理的。

第三，**它运行的环境**。这是沙盒，回顾性病例，八种预选诊断，仅限文本。真实的临床评估，如 Dominic Oliver 博士（牛津大学）所言，不仅取决于患者说了什么，还取决于他们怎么说——伴随体格检查、观察到的行为和身体语言，而这些是一个阅读已完成病历的纯文本代理永远看不到的。并且一个疾病不在八种选定诊断之列的患者，根本不在本研究能够讨论的范围内。

还有一个更微妙的担忧被评审人提出：**数据污染**。 MIMIC-IV 是公开的且被大量书写，因此一个在开

放互联网上训练的语言模型可能已经见过论文、病例讨论或数据本身。Midhun Parakkal Unni 博士（谢菲尔德大学）直接指出了这一点——如果部分答案存在于训练数据中，则部分表现是记忆而非临床推理，只有独立复现才能区分两者。值得注意的是，作者没有回避这一点：他们写道，其结果「可以审慎地解释为可能的上限」，并且「可能高估了对其他公开病例的泛化能力」——一篇给自己的标题设置上限的论文。

这篇论文没有证明什么

- 它**没有**证明 AI 在真实临床环境中比医生更好。这是在沙盒中，基于回顾性数据，仅有八种诊断，仅限文本。
- 它**没有**展示真实世界的患者结局——没有死亡率、没有并发症、没有患者满意度。
- 它**没有**证明更广泛的诊断能力。一个不在八种诊断之列的病例不在研究范围。
- 它**没有**证明自主 AI 在诊所中是安全的。说沙盒中无伤害与说它可以被部署是两回事。
- 它**没有**完全排除数据污染。作者承认了这一点。

证据有多强?

对于核心主张——在沙盒中，MIRA 能够处理整个临床工作流程并在受控回顾性病例上达到高诊断准确率——证据作为能力演示是强有力的。这是一个真实的工程成就：在 FHIR 上运行的代理、超过八万种行动、跨八种诊断的工作流程。那依然是令人印象深刻的。

但作为性能主张——AI 在诊断上击败了医生——应该更加小心地看待。优势被集中在了检验结果最清晰的疾病上，且 MIRA 比医生多获取了大量廉价实验室数据。差距至少部分是由信息而非智力驱动的。此外，该论文没有测量真实患者结局，仅测量了病历上的准确率。

最诚实的框架：一项令人印象深刻的能力演示，而非一项临床验证。

清洁摘要

MIRA 是一个在沙盒化电子健康档案中运行并独立处理完整临床病例的 AI 代理。在 MIMIC-IV 基准测试（574 个病例，8 种诊断）上，它在诊断准确率上超越了医生——独自运行时达到 88.9%，在共享的 311 病例子集上为 87.8%，而资深医生为 78.1%，混合资历团队为 71.1%。但其优势集中在具有决定性强检验结果的疾病上，且 MIRA 比人类获取了更多的实验室数据。性能主张令人震惊，但沙盒、有限的诊断范围、信息不对称和数据污染风险都限制了泛化能力。这是一个真正的工程步骤，而非 AI 已准备好无监督执业的证明。

No-BS 检查

论文所展示的：一个 AI 代理能够端到端地处理临床工作流程，并在八种预选诊断上超越医生，在沙盒中，基于回顾性病历。**合理但未经证实的：**相同的表现会在真实诊所中成立；AI 实际上在推理，而非记忆；差距在同类信息下依然能保持。**它没有展示的：**真实世界患者结局、广泛的诊断覆盖范围、临床安全性、或已部署系统的有效性。**主要局限：**沙盒、回顾性数据、8 种诊断、文本仅限、信息不对称、可能的数据污染、MIMIC-IV 的特定性。**普通读者应该抱有多大信心？**高度信心认为这是一个有能力的代理。中度信心认为诊断击败是干净的。低度信心认为这是「更好的医生」。恰当的立场：令人印象深刻的工程技术，而非经过验证的临床工具。

来源

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