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 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

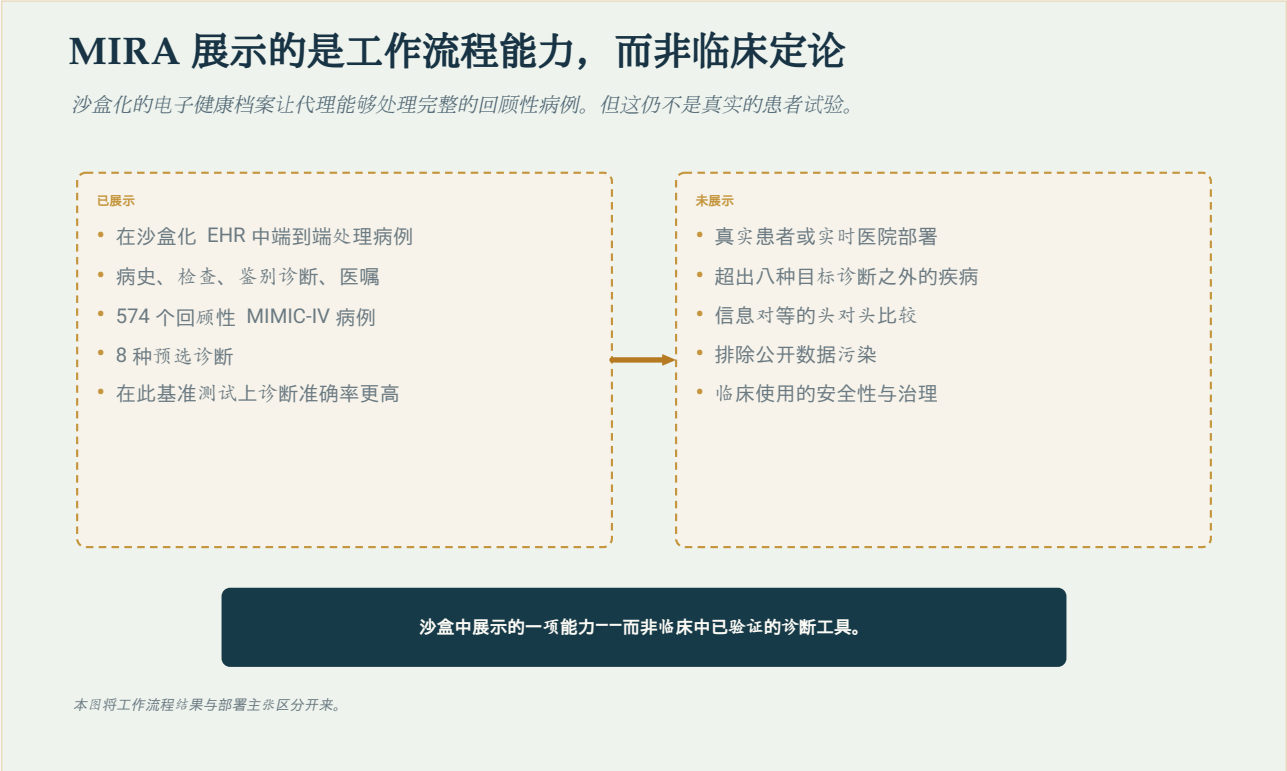
回答问题不等于处理完整病例

大多数医疗 AI 故事讲的都是一个会回答的模型。你给它一段病例简介或一道考试题，它返回一个诊断或一段建议，而且得分不错。这很有用，但范围很窄：它就像一个从不碰病历的聪明会诊医生。

MIRA 的设计目标不同，而这种不同才是真正的新闻。它不是回答一个问题，而是处理完整病例：读取患者病历，判断还需要了解哪些病史，开具实验室、影像和微生物学检查，读取结果，缩小鉴别诊断范围，然后写下随之而来的医嘱——处方、入院、转诊至外科。它通过在电子健康记录内部采取行动来完成这些事，就像临床医生一样，而不是生成供人抄录的自由文本。

这是实实在在的一步：从提供建议的系统，走向贯穿整个工作流程并采取行动的系统。但在报道中，这一步也恰恰最容易被压扁成“AI 胜过医生”。因此，有必要准确说明测试了什么，以及测试范围在哪里。

一句话概括：MIRA 能够独立处理完整病例，而且在这项基准测试中，诊断准确率超过了医生——但它是在沙盒中、基于回顾性病历、针对预先选定的八种诊断、仅通过文字交流完成的，并且它的大部分优势来自检查结果最明确的疾病。进步是真实的，计分板却不是诊所。



MIRA 在沙盒化 EHR 内展示了端到端工作流程能力。它没有展示真实世界的临床部署能力、广泛的诊断覆盖范围，也没有与医生进行一场信息完全对等的公平正面对比。

作者做了什么

MIRA——Medical Intelligence for Reasoning and Action（用于推理和行动的医疗智能）——是一个基于 OpenAI 模型构建的自主智能体：负责对话和行动的部分运行在 **GPT-4o** 上，另一个独立的规划步骤使用 OpenAI 的 **o1** 推理模型。这些组件运行在一个定制的、符合标准的框架中——该框架基于 HL7 FHIR，这是现实中的医院用于传递病历的数据标准——并配备了八万多个可能的临床操作工具。在这个沙盒里，MIRA 可以调取患者病史，开具并解读实验室、影像和微生物学检查，生成鉴别诊断，并制定治疗计划——开药、安排外科手术、计划入院。这里先指出一个值得注意的细节：为 MIRA 的回答打分的自动“评委”本身就是 GPT-4o，也就是正在接受测试的同一模型家族；在同行评审者提出这一担忧后，作者又让一名获得委员会认证的医生进行了复核。

测试集来自 **MIMIC-IV**，这是一个公开的大型去标识化 EHR 数据库。作者从 2008 年至 2019 年间在 Beth Israel Deaconess Medical Center 接受治疗的大约 300,000 名患者中，选取了涵盖 **八种目标诊断** 的 **574 个患者病例**——包括腹部病变和内科急症，其中有阑尾炎、胰腺炎、肺炎和尿路感染。对于每个病例，MIRA 都从有限的信息开始，必须逐步判断下一步该做什么——形式上与真实检查过程相同，但它面对的是一份真实结局已经记录在案的病历。

随后，研究者将它在这些相同病例上的表现与医生进行了比较。

> **“沙盒化 EHR 中的自主智能体”究竟是什么意思**
>
> 这里的三个词包含了很多含义，值得拆开来说。
> **智能体**意味着这个系统不是一个回答提示词的聊天机器人。它会运行一个循环：查看当前状态，选择一项行动（开这项检查、询问那段病史），查看结果，再选择下一项行动——直到得出诊断和方案。“智能”来自语言模型；自主行动能力来自围绕模型搭建的支架，它把模型的文本转换成获准执行的 EHR 操作，再将结果反馈回来。
> **沙盒化 EHR** 意味着这是一个模拟的、隔离的病历系统副本，而不是医院正在使用的真实系统。MIRA 可以“开具”一项检查，并收到该患者实际做过的结果，因为病例来自历史记录，而且答案已经被记录下来。它做的任何事都不会触及真实患者或真实诊所。
> **自主**意味着它在没有人类参与的情况下端到端完成病例——但仅限于沙盒内。这**不**意味着它被放出去，在没有监督的情况下管理患者。这是两种截然不同的说法，而测试的只有第一种。
> 还需要说明沙盒的一点，因为人们很容易对此形成错误想象：MIRA 可以开具它想要的任何检查，但只有当患者现实中**确实接受过**这项检查时，系统才能返回结果。要求一项患者做过的血液检查，你会得到真实数值；要求一项没人开过的扫描，你得到的则是“N/A — 无法执行”，绝不会得到编造的数字。病史也是同样：如果病历没有 MIRA 询问的内容，模拟患者只会说自己不知道。因此，MIRA 不能召唤出一项从未做过的决定性检查——它只能利用真实检查过程碰巧包含的内容。
> 这个区别很重要，因为标题中的那个词——“自主”——最容易被理解成后者。

他们发现了什么

在这项基准测试中，**MIRA 的诊断准确率超过了医生**——但标题掩盖了三个不同的数字，而它们很容易被混为一谈，所以有必要分开来看。

MIRA 独立处理全部 **574 个病例**时，有 **88.9%** 的时间给出了正确诊断——评分依据是 MIMIC-IV 中每位患者实际记录的出院诊断。至于与人类进行正面对比，医生并没有处理全部 574 个病例，而是处理了一个共享的 **311 个病例子集**，使用与 MIRA 相同的病历和工具。在这 311 个相同病例上，MIRA 得分 **87.8%**，研究者将它与两个单独招募的群体进行了比较。第一组是**资深组**：四名获得委员会认证、拥有 7 至 11 年经验的医生，得分 **78.1%**。第二组是**混合资历组**：主要由初级住院医师和两名专科医生组成，更接近真实急诊科的人员构成，得分 **71.1%**。病例相同，工具相同——最终顺序却是 AI 第一，经验丰富的医生第二，以初级医生为主的团队第三。论文自己的谨慎表述是，在八种疾病中，MIRA 的表现“始终不逊于医生，而且经常超过”医生。

它后续的决策也经得起检验：总体上符合指南，用药安全，入院安排恰当。作者报告称，在五类安全问题（药物相互作用、肾脏剂量、过敏、QT 风险和阿片类药物处方）中，没有出现高严重度用药错误；468 个病例中有 467 个病例的处方正确——同时也指出该系统“没有达到 100% 的可靠性”。如果照字面看，这是一个引人注目的结果：一个能够运行完整病例的智能体，在诊断上胜过了医生。

但胜利的形状与胜利本身同样重要。独立专科人士阅读论文后指出了这种优势的实际来源。在科学媒体中心的专家小组中，Wei Xing 医生（谢菲尔德大学）指出，标题中的数字——AI 在诊断准确率上胜过医生——**主要由检查结果明确的疾病驱动**，例如阑尾炎和胰腺炎，因为一项决定性的扫描或实验室数值就能解决问题。对于肺炎和尿路感染——这两种疾病是人们实际前往急诊科的最常见原因之一——*AI* 和医生的表现都最差，而且两者之间的差距最小。

双方进行比较时还存在一种不对称，而这点就体现在论文自己的数字中。MIRA 更依赖实验室检查：它利用了常规医疗中可用实验室分析物的大约 **51%**，而获得委员会认证的医生约使用 **28%**——每个病例的中位数多出七项血液参数。作者谨慎地将这一点描述为低于数据集自身的常规医疗基线，而不是“什么都开”的策略，并报告称高成本横断面影像检查没有系统性增加。不过，更多信息本身就可能带来更高的诊断准确率——因此，这并不完全是信息量相同的选手之间的同类对决，这也是 Xing 直接指出的差距。如果允许临床医生开具所有便宜的血液检查，却不必权衡费用、不适或延误，他们在纸面上的表现也会更好。

“胜过医生”不等于“更好的医生”

基准测试获胜很容易被过度解读。从“MIRA 得分更高”到“MIRA 更好”之间，隔着三件事。

第一，**它是依据什么评分的**。Julie Jacko 教授（爱丁堡大学）指出，MIRA 的几项关键结果是相对于底层数据集中记录的内容定义的——这意味着系统获得的奖励是**复现记录中的临床行为**，而不一定是展示最佳医疗。历史病历成了答案依据。这是构建基准测试的一种合理方式，但它衡量的是与既有做法的一致性，而不是某种绝对意义上的正确性。公平地说，这一问题对治疗一致性指标的影响最大——MIRA 的医嘱是否与病历一致？——而标题中的诊断准确率是以出院诊断为依据的，比单纯复述记录更接近真实结局。

第二，就是前面提到的**信息不对称**：血液检查多得多（可用分析物约 51%，而不是 28%）就意味着证据更多。准确率差距中的一部分，很可能是用资源买来的，而不是推理出来的。

第三，**它运行在哪里**。这是一项在沙盒中、基于回顾性病例、针对预先选定的八种诊断、仅以文字进行的测试。正如 Dominic Oliver 医生（牛津大学）所说，真实临床评估不仅取决于患者说了什么，也取决于他们怎么说——还包括体格检查、观察到的行为和肢体语言；这些全都不是一个只读文字、阅

读已经完成的病历的智能体能够看到的。而如果患者的问题不属于选定的八种诊断之一，那就完全超出了这项研究能够说明的范围。

评审者还提出了一个不那么显眼的担忧：**数据污染**。MIMIC-IV 是公开数据集，相关内容也被大量撰写，因此，一个在开放互联网数据上训练的语言模型可能已经看过相关论文、病例讨论，甚至数据本身。谢菲尔德大学的 Midhun Parakkal Unni 医生直接提出了这一点——如果部分答案出现在训练数据中，那么表现的一部分来自记忆，而不是临床推理；只有独立复现才能区分二者。值得注意的是，作者没有敷衍这一问题：他们写道，研究结果 “可以谨慎地解释为一个可能的上限”，并且 “可能高估了对其他公开病例的泛化能力”——这是一篇主动给自己的标题数字设限的论文。

这些都不会让 MIRA 变得不那么有趣。它们只是意味着，正确的解读应当经过校准：在一项回顾性基准测试中，一个能够贯穿整份病历采取行动的智能体，在检查结果明确的诊断上胜过了医生——部分原因是它开了更多检查，部分原因是它与记录中的病历相一致。这是一次真实的能力展示，但不是 “机器现在诊断得比医生更好” 的判决。

这项研究**不能**证明什么

- 它**不能**说明 MIRA 能用于真实患者。每个病例都是回顾性且经过模拟的，没有任何患者曾由它管理。
- 它**不能**说明 MIRA 能处理八种诊断之外的疾病。预先选定范围之外的疾病——医学中混乱、尚未分化的大多数情况——没有接受测试。
- 它**不能**说明 MIRA 可以安全部署。作者自己写道，泛化能力、安全性和治理仍需要前瞻性的真实世界研究。
- 它**不能**证明与医生进行的是公平的正面对比。它使用了多得多的血液检查（可用分析物约 51%，而不是 28%），而且几项治疗结果的评分部分依据记录中的病历，而不是最佳治疗的真实标准答案。
- 它**不能**说明结果没有受到记忆影响。公开的 MIMIC-IV 数据可能与模型训练数据重叠，独立复现需要排除这一点。
- 它**不**意味着 “AI 取代医生”。它是一种能够贯穿整份病历采取行动的决策支持工具；评审者的共识是，现实中的使用将与临床医生合作，由临床医生保留决策权，并提供文字病历无法包含的全部信息。

证据有多强?

把这个主张拆成两半，因为两半的证据很不一样。

作为能力展示——也就是证明语言模型智能体能够在 EHR 沙盒中运行完整的临床病例，将病史、检查、诊断和医嘱端到端串联起来——这项工作确实新颖，也相当有说服力。这是其中真正新颖的部分，也是值得关注的部分。

作为优越性主张——也就是 MIRA 比医生更好——证据是有边界的，应当谨慎解读。这个结论只在一项具体的回顾性基准测试、八种诊断、存在信息不对称（检查更多）、答案取自历史病历的情况下成立，同时还存在训练数据污染的可能性。这足以说明 “该智能体在这项基准测试中的表现令人印象深刻”，

却不足以说明“该智能体是比医生更好的诊断者”；作者也没有提出后一种主张。

最有用的立场既不是“AI 胜过医生”，也不是“不过是玩具”。应该这样理解：一种新型医疗 AI——它不是回答问题，而是贯穿病历采取行动——在一项艰难的回顾性基准测试中表现良好；现在，它必须在尚未接受检验的地方证明自己：面对真实的、尚未分化的患者，在前瞻性研究和治理框架下接受测试。

为什么这很重要

几年来，医疗 AI 中真正有趣的问题已经悄然改变。过去的问题是“模型能否正确诊断？”——模型在越来越理想化的测试集上不断回答“能”。MIRA 把问题转向了更难的一层：“模型能否完成这项工作——收集、开具、解读、决策、行动——贯穿整个工作流程？”这是一个更有用、也更诚实的问题，因为困难和风险都存在于采取行动之中。

这正是框架如此重要的原因。标题式的条件反射——*AI* 胜过医生——指向了结果中最不新颖、支持也最薄弱的部分。真正的新东西更小，却影响更大：一个能够贯穿整份病历运行的智能体。如果这项能力经受住检验，它首先改变的是**工作流程**，很久之后才可能改变谁负责它。医生并没有被取代；开检查、解读结果、追踪检查结果——也就是工作流程——才是这类工具首先落地的地方。

它还把举证责任重新放回了正确的位置。在回顾性病例上的排行榜胜利，是开展前瞻性试验的理由，而不是试验的替代品。作者也表达了同样的意思。对 MIRA 最诚实的解读，是把它视为对下一项研究的邀请——按照这个领域已经熟悉的标准：在患者身上进行评估，而不是在基准测试上。

简明总结

MIRA 是一个自主 AI 智能体，运行在沙盒化电子健康记录中：它可以采集病史，开具并解读实验室、影像和微生物学检查，得出诊断，并写出治疗计划。在公开 MIMIC-IV 数据库的 574 个回顾性病例上、针对预先选定的八种诊断进行测试时，它的诊断准确率超过了医生（总体 88.9%；正面对比中为 87.8%，医生为 78.1%），并作出了总体上符合指南且用药安全的决策。但这项评估是在过去病历上的模拟，仅使用文字；它的大部分优势来自检查结果明确的疾病；它使用的血液检查远多于医生（可用分析物约 51%，而医生为 28%）；几项治疗结果是依据原始病历的记录评分的；公开数据集还带来了真实的数据污染风险——作者自己也称这可能构成其数字的上限。真正的进步在于，一个智能体能够贯穿整个工作流程采取行动，而不是只回答孤立的问题。作者明确表示，泛化能力、安全性和治理仍需要前瞻性的真实世界研究——正确的解读正是如此：这是一次令人印象深刻的能力展示，不是 AI 诊断能力胜过医生的证明，也不是一个已经准备好进入真实诊所的系统。

不绕弯检查

论文展示了什么：一个语言模型智能体（MIRA）能够在沙盒化 EHR 内端到端运行完整的临床病例——病史、检查、诊断、医嘱——并且在涵盖八种诊断的 574 个回顾性病例基准测试中，诊断准确率

超过了医生（总体 88.9%；与获得委员会认证的医生正面对比时为 87.8%，医生为 78.1%），且没有出现高严重度用药错误。

合理但尚未证明的内容：这项能力能够为真实的、尚未分化的患者带来益处；准确率优势反映的是更好的推理，而不是开具了更多检查、与记录中的病历一致，或记住了公开数据。

它没有展示什么：MIRA 能够处理八种诊断之外的疾病；它可以安全部署；它在公平且信息量相同的比较中胜过医生；它能取代临床医生；结果没有受到训练数据污染。

主要局限：回顾性、模拟、仅使用文字的评估；八种预先选定的诊断；信息不对称（血液检查多得多——可用分析物约 51%，而不是 28%，而且是低成本血液检查，不是影像检查）；治疗结果部分依据历史病历评分；以及一个语言模型可能在训练中见过的公开基准（MIMIC-IV）——作者指出，这可能构成表现的上限。

普通读者应该有多大信心？对于 MIRA 是一次真实且新颖的能力展示——一个能够贯穿病历采取行动、而不只是聊天的智能体——应当有较高信心。对于它是比医生更好的诊断者，信心应当较低：这一主张被限制在一项回顾性基准测试中，而且受到检查开具、依据病历评分以及可能的数据污染等因素的干扰。作者自己的结论是正确的——在这项结果对患者护理产生任何意义之前，还需要开展前瞻性的真实世界研究。

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-

News-Medical (2026) —illustrates the 'outperforms doctors' framing
https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-simulated-EH
aspx