The Clean Paper

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

# AI 에이전트가 환자 사례 전체를 혼자 처리했다—과거 기록을 이용한 시뮬레이터 안에서

*MIRA* 는 새로운 유형의 의료 *AI* 다. 질문 하나에 답하는 대신 시뮬레이션 병원 기록 안에서 병력을 받고, 검사를 주문·판독하고, 진단하고, 오더를 작성하며 환자 사례 전체를 처리한다. 공개 데이터베이스의 과거 사례 *574* 건과 미리 선택된 *8* 개 진단에서 저자들은 *MIRA* 가 진단 정확도에서 의사보다 높았고 대체로 지침에 부합하며 약물 안전성이 높은 결정을 했다고 보고한다. 하지만 모든 수식어가 중요하다. 과거 기록을 이용한 텍스트 전용 샌드박스였고, 우위의 상당 부분은 검사 결과가 명확한 질환에서 나왔으며, 의사보다 약 두 배 많은 혈액검사 항목을 사용했고, 여러 결과는 원래 차트 기록을 기준으로 채점됐다. 진짜 진전은 워크플로 전체에서 행동하는 에이전트이지 기계가 이제 의사보다 진단을 잘한다는 증명이 아니다. 저자들도 일반화·안전성·거버넌스에는 전향적 실제 임상 연구가 필요하다고 말한다.

---

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

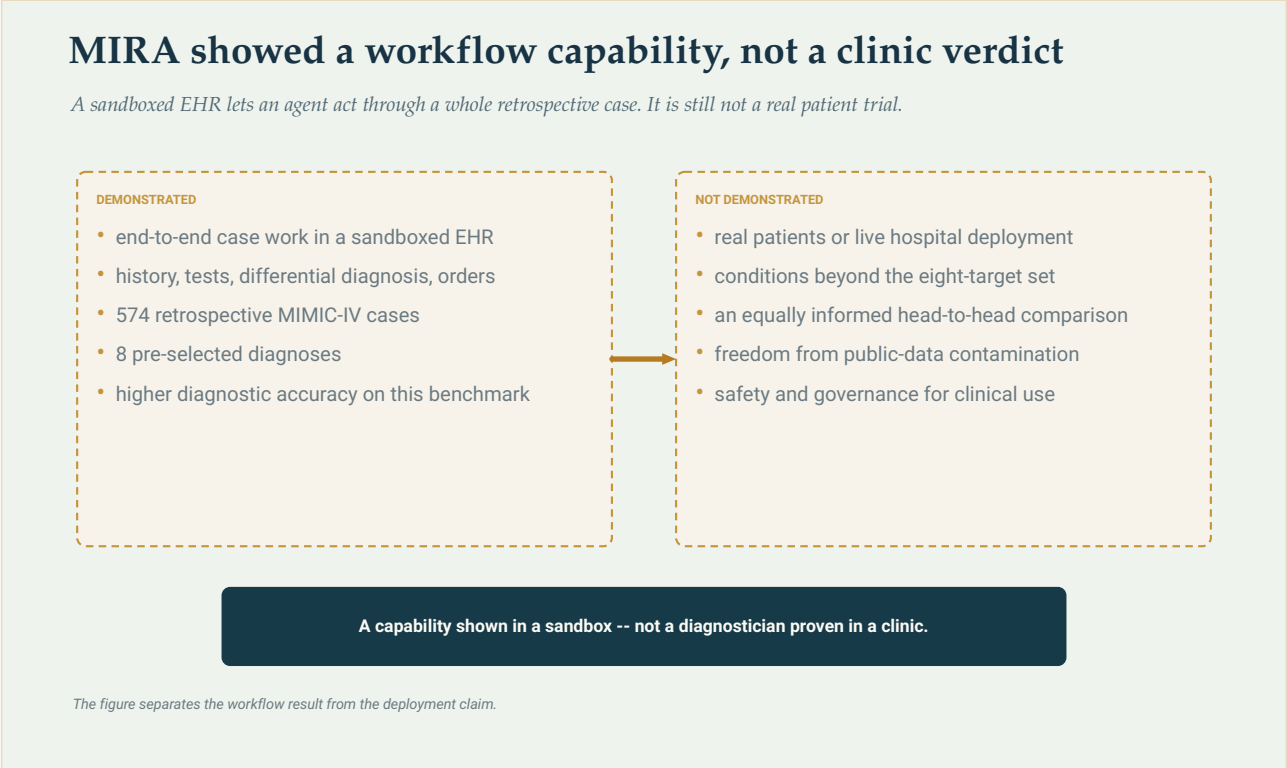
# 질문에 답하는 것과 환자 한 사례를 끝까지 운영하는 것은 다르다

의료 AI 이야기는 대개 답하는 모델에 관한 것이다. 임상 사례나 시험 문제를 주면 진단이나 조언 한 문단을 내놓고 높은 점수를 받는다. 유용하지만 범위는 좁다. 차트를 직접 다루지 않는 영리한 자문가와 비슷하다.

MIRA 는 다른 일을 하도록 만들어졌고, 바로 그 차이가 이 연구의 실제 뉴스다. 질문 하나에 답하는 대신 환자 한 사례 전체를 처리한다. 의무기록을 읽고, 어떤 병력이 더 필요한지 결정하고, 검사실 검사·영상·미생물 검사를 주문하고, 결과를 읽고, 감별진단을 좁힌 뒤, 처방·입원·수술 의뢰 같은 후속 오더까지 작성한다. 사람이 옮겨 적을 자유 텍스트를 만드는 대신 임상의처럼 전자의무기록 안에서 행동을 취한다.

이는 실제로 한 단계의 변화다. 조언만 하는 시스템에서 진료 과정 전체를 직접 처리하는 시스템으로 옮겨간 것이다. 동시에 보도에서는 이런 변화가 쉽게'AI 가 의사를 능가했다' 는 한 문장으로 납작해진다. 그래서 무엇을, 어디에서 시험했는지를 정확히 볼 필요가 있다.

한 문장으로 요약하면 이렇다. MIRA 는 환자 사례 전체를 스스로 처리했고 이 벤치마크에서 진단 정확도가 의사보다 높았다. 하지만 샌드박스에서, 과거 의무기록을 대상으로, 미리 선택된 8 개 진단 범위에서, 텍스트로만 소통하며 시험됐고, 우위의 상당 부분은 검사 결과가 가장 명확한 질환에서 나왔다. 성과는 실제다. 하지만 벤치마크가 곧 진료실은 아니다.




MIRA 는 샌드박스 EHR 안에서 처음부터 끝까지 진료 과정을 처리하는 능력을 보여줬다. 실제 임상 배치, 광범위한 진단 범위, 또는 의사와 동일한 정보를 가진 공정한 일대일 비교를 보여준 것은 아니다.

# 연구진이 한 일

MIRA(Medical Intelligence for Reasoning and Action) 는 OpenAI 모델을 기반으로 만든 자율 에이전트다. 대화하고 행동하는 부분은 **GPT-4o** 에서 실행되고, 별도의 계획 단계에는 OpenAI 의 **o1** 추론 모델이 쓰인다. 이들은 실제 병원이 기록을 주고받을 때 사용하는 데이터 표준인 HL7 FHIR 를 바탕으로 한 맞춤형 표준 준수 프레임워크 안에 들어 있으며, 가능한 임상 행동은 8 만 개가 넘는다. 샌드박스 안에서 MIRA 는 환자 병력을 가져오고, 검사실·영상·미생물 검사를 주문하고 해석하며, 감별진단을 만들고, 약을 처방하고 수술을 일정에 잡거나 입원을 계획하는 등 치료 계획까지 세울 수 있다. 미리 짚고 넘어갈 세부사항도 있다. MIRA 의 답을 자동 채점한 '심사자' 역시 GPT-4o, 즉 시험 대상과 같은 모델 계열이었다. 동료심사자가 이 문제를 제기한 뒤 저자들은 전문의 자격을 가진 의사의 검토를 추가했다.

시험 세트는 대규모 공개 비식별 EHR 데이터베이스인 **MIMIC-IV** 에서 가져왔다. 2008~2019 년 Beth Israel Deaconess Medical Center 에서 치료받은 약 30 만 명 가운데 연구진은 **574 개 환자 사례**를 골랐다. 대상은 **8 개 진단**으로, 충수염·췌장염·폐렴·요로감염 등을 포함한 복부 질환과 내과 응급질환이었다. 각 사례에서 MIRA 는 제한된 정보만 가진 상태로 시작해 다음에 무엇을 할지 단계별로 결정해야 했다. 실제 진단 과정과 형태는 비슷하지만, 이미 실제 결말이 기록된 과거 사례를 대상으로 한 것이다.

그 다음 같은 사례에서 의사들과 성능을 비교했다.

> **' 샌드박스 EHR 의 자율 에이전트' 가 실제로 뜻하는 것**
>
> 여기서는 세 단어가 매우 많은 의미를 떠맡고 있으므로 풀어볼 가치가 있다.
> 여기서 **에이전트 (agent)** 라는 말은 단순히 프롬프트에 답하는 챗봇이 아니라는 뜻이다. 현재 상태를 보고, 행동 (이 검사를 주문하라, 저 병력을 물어보라) 을 고르고, 결과를 본 뒤 다음 행동을 선택하는 순환을 진단과 계획에 도달할 때까지 반복한다. 언어 모델이 추론을 담당하고, 그 출력을 허용된 EHR 작업으로 바꾼 뒤 결과를 다시 모델에 돌려주는 주변 구조가 자율적 행동을 가능하게 한다.
> **샌드박스 EHR** 은 실제 병원 시스템이 아니라 벽으로 격리된 의료기록 시스템의 시뮬레이션 복사본이라는 뜻이다. 사례가 과거 기록이고 결과가 이미 저장돼 있기 때문에 MIRA 가 검사를' 주문' 하면 그 환자가 실제로 받았던 검사 결과를 돌려받을 수 있다. MIRA 가 하는 어떤 행동도 실제 환자나 실제 진료 현장에는 영향을 주지 않는다.
> **자율 (autonomous)** 은 사람의 개입 없이 사례를 처음부터 끝까지 처리한다는 뜻이다. 단, 어디까지나 샌드박스 안에서 그렇다. 환자를 감독 없이 맡겨 관리하게 했다는 뜻은 **아니다**. 두 주장은 전혀 다르고, 이번 연구가 시험한 것은 첫 번째뿐이다.
> 샌드박스를 잘못 상상하기 쉬워 한 가지를 더 덧붙이면, MIRA 는 원하는 검사를 무엇이든 주문할 수 있지만 그 검사가 실제 환자에게 **실제로 시행됐던 경우에만** 결과를 받을 수 있다. 실제로 시행된 혈액검사를 요청하면 실제 수치를 받지만, 현실에서 아무도 주문하지 않았던 스캔을 요청하면 꾸며낸 숫자 대신'N/A — 시행할 수 없음' 이 돌아온다. 병력도 같다. 기록에 없는 내용을 물으면 시뮬레이션 환자는 모른다고 답한다. 따라서 MIRA 는 실제로 하지 않았던 결정적 검사를 마음대로 소환할 수 없고, 현실의 진단 과정에 포함됐던 정보 범위 안에서만 일한다.
> 이 구분이 중요한 이유는 제목에서 가장 눈에 띄는' 자율' 이라는 말이 두 번째 의미로 읽힐 가능성이 가장 높기 때문이다.

# 무엇을 발견했나

벤치마크에서 **MIRA 의 진단 정확도는 의사보다 높았다.** 하지만 이 제목 아래에는 서로 다른 세 숫자가 들어 있고 쉽게 뒤섞일 수 있으므로 분리해서 봐야 한다.

전체 **574 개 사례**에서 MIRA 가 올바른 진단을 낸 비율은 **88.9%** 였다. 정답은 MIMIC-IV 에 실제 기록된 퇴원 진단을 기준으로 했다. 사람과의 직접 비교에서는 의사들이 574 건 모두를 처리한 것이 아니라, MIRA 와 동일한 기록과 도구를 사용해 공통 **311 개 사례**를 처리했다. 이 311 건에서 MIRA 는 **87.8%** 를 기록했고, 별도로 모집한 두 집단과 비교됐다. 첫 번째는 7~11 년 경력의 전문의 4 명으로 구성된 **고경력 집단**으로 **78.1%** 였다. 두 번째는 주로 초년 레지던트와 전문의 2 명으로 구성된 **혼합 경력 집단**으로, 실제 응급실 인력 구성에 더 가까웠고 **71.1%** 였다. 같은 사례와 같은 도구에서 AI 가 1 위, 숙련 의사가 2 위, 주니어 중심 팀이 3 위였다. 논문은 8 개 질환 전체에서 MIRA 가 의사와' 일관되게 동등했고, 자주 능가했다' 고 조심스럽게 표현한다.

후속 결정도 상당히 잘 버텼다. 대체로 진료지침에 부합했고 약물 안전성과 입원 적절성도 높았다. 저자들은 약물 상호작용, 신장기능에 따른 용량, 알레르기, QT 위험, 오피오이드 처방이라는 5 개 안전 범주에서 고중증도 약물 오류가 없었고 468 건 중 467 건의 처방이 정확했다고 보고한다. 동시에 시스템이'100% 신뢰도에 도달하지는 않았다' 고 적었다. 표면적으로만 보면 놀라운 결과다. 사례 전체를 운영한 에이전트가 진단에서 의사보다 앞섰다.

하지만 승리의 형태는 승리 자체만큼 중요하다. 논문을 읽은 독립 전문가들은 우위가 어디서 나왔는지 짚었다. Science Media Centre 의 전문가 패널에서 셰필드대의 Wei Xing 박사는 AI 가 진단 정확도에서 의사를 이겼다는 대표 수치가 **충수염이나 췌장염처럼 검사 결과가 명확한 질환에서 주로 나온 것**이라고 지적했다. 결정적인 스캔이나 검사 수치가 답을 가르는 경우다. 반대로 실제 응급실 방문의 흔한 원인인 폐렴과 요로감염에서는 AI 와 의사 모두 가장 낮은 성능을 보였고 격차도 가장 작았다.

두 쪽이 게임을 한 방식에도 비대칭이 있었고, 이는 논문 수치에 그대로 드러난다. MIRA 는 검사실 정보를 더 적극적으로 사용했다. 일상 진료에서 이용 가능한 검사 분석항목의 약 **51%** 를 활용한 반면 전문의는 약 **28%** 였다. 사례당 중앙값으로 혈액검사 항목 7 개가 더 많았다. 저자들은 이것이 '가능한 검사를 전부 주문하는' 전략이 아니라 데이터셋의 실제 일상진료 기준보다도 낮은 수준이라고 설명하고, 더 비싼 단층 영상검사가 체계적으로 늘지는 않았다고 보고한다. 그래도 정보가 많으면 그 자체로 진단 정확도가 올라갈 수 있다. 따라서 똑같이 정보를 가진 두 참가자의 완전히 공정한 비교는 아니며, Xing 도 바로 이 점을 지적했다. 비용·불편·지연을 고려하지 않고 저렴한 혈액검사를 마음껏 주문할 수 있는 임상의도 종이 위에서는 더 정확해 보일 수 있다.

# ' 의사를 이겼다' 와' 더 좋은 의사다' 는 다르다

벤치마크 승리는 과대해석하기 쉽다. 'MIRA 점수가 더 높았다' 에서'MIRA 가 더 낫다' 로 가는 사이에는 세 가지가 놓여 있다.

첫째는 **무엇을 정답으로 채점했는가**다. 에든버러대 Julie Jacko 교수는 MIRA 의 몇몇 핵심 결과가 원래 데이터셋에 기록된 내용을 기준으로 정의돼 있다고 지적했다. 즉 시스템은 반드시 최적의 진료를 보여줘서가 아니라 **기록된 임상 행동을 재현하면** 보상을 받는다. 역사적 차트가 답안지가 되는 셈이다. 벤치마크를 만드는 합리적인 방법이지만, 절대적 의미의 올바름보다 실제로 수행된 진료와의 일치도를 잰다. 다만 이

문제는'MIRA 의 오더가 기록과 일치했는가' 를 보는 치료 일치도 지표에 가장 크게 해당한다. 대표적인 진단 정확도는 퇴원 진단을 기준으로 하므로 단순한 기록 모방보다는 실제 결과에 더 가깝다.

둘째는 앞서 말한 **정보의 비대칭**이다. 이용 가능한 혈액검사의 약 51% 를 쓴 것과 28% 를 쓴 것은 증거량 자체가 다르다. 정확도 차이의 일부는 더 나은 추론이 아니라 더 많은 검사를 통해 얻은 것일 수 있다.

셋째는 **어디에서 실행됐는가**다. 이것은 샌드박스에서 과거 사례를 대상으로, 미리 선택한 8 개 진단에 한정해, 텍스트만으로 수행한 시험이었다. Oxford 의 Dominic Oliver 박사가 지적했듯 실제 임상 평가는 환자가 무엇을 말하는지만이 아니라 어떻게 말하는지, 신체검사, 관찰되는 행동과 몸짓에도 의존한다. 이미 완성된 기록을 읽는 텍스트 전용 에이전트는 이런 것을 전혀 보지 못한다. 게다가 환자의 문제가 선택된 8 개 진단 중 하나가 아니라면 이 연구는 그 경우에 대해 말할 수 없다.

심사자들이 제기한 또 다른 우려는 **데이터 오염 (data contamination)** 이다. MIMIC-IV 는 공개돼 있고 수많은 글과 논문에서 다뤄졌으므로, 공개 인터넷으로 학습한 언어 모델이 관련 논문이나 사례 토론, 혹은 데이터 자체를 이미 보았을 가능성이 있다. 셰필드대 Midhun Parakkal Unni 박사는 이를 직접 지적했다. 일부 정답이 학습 데이터에 있었다면 성능의 일부는 임상 추론이 아니라 기억일 수 있고, 둘을 구분하려면 독립적인 재현이 필요하다. 중요한 점은 저자들도 이를 얼버무리지 않았다는 것이다. 결과를 '가능한 상한선으로 조심스럽게 해석할 수 있으며', ' 다른 공개 사례에 대한 일반화 성능을 과대평가할 수 있다' 고 썼다. 논문 스스로 자기 대표 수치에 천장을 설정한 셈이다.

이 모든 것이 MIRA 를 덜 흥미롭게 만드는 것은 아니다. 다만 결과를 어디까지 받아들여야 하는지 범위를 정해 준다. 과거 자료 벤치마크에서 기록 전체를 오가며 행동할 수 있는 에이전트가 검사 결과가 명확한 질환에서 의사보다 높은 진단 성능을 보였다. 그 일부는 더 많은 검사를 주문한 데서 왔고, 일부 평가는 기록된 차트와 얼마나 맞는지를 보았다. 이는 진짜 능력 시연이지, 기계가 이제 의사보다 진단을 더 잘한다는 판결문은 아니다.

# 이 연구가 증명하지 않는 것

- MIRA 가 실제 환자에게서 작동한다는 것을 보여주지 **않는다**. 모든 사례는 과거 자료를 이용한 시뮬레이션이었고 실제 환자를 관리한 적은 없다.
- 8 개 진단을 넘어 작동한다는 것을 보여주지 **않는다**. 미리 선택된 범주 밖의 질환, 즉 실제 의학의 복잡하고 미분화된 대부분은 시험하지 않았다.
- 배치해도 안전하다는 것을 보여주지 **않는다**. 저자들 스스로 일반화, 안전성, 거버넌스에는 전향적 실제 임상 연구가 필요하다고 쓴다.
- 의사와 공정한 일대일 비교를 확립하지 **않는다**. MIRA 는 이용 가능한 혈액검사 항목을 훨씬 많이 사용했고 (약 51% 대 28%), 여러 치료 결과는 최선의 진료라는 객관적 진실보다 기록된 차트를 부분적으로 기준 삼아 채점됐다.
- 결과가 암기 효과와 무관하다는 것을 보여주지 **않는다**. 공개 MIMIC-IV 데이터가 모델 학습에 들어갔을 가능성은 독립적 재현으로 배제해야 한다.
- 'AI 가 의사를 대체한다' 는 뜻이 **아니다**. 이것은 기록 전체에서 행동할 수 있는 의사결정 지원 시스템이다. 전문가들의 공통된 견해는 실제 사용 시 임상의와 협력하며, 임상의가 권한을 유지하고 텍스트 기록에 빠진 요소를 제공해야 한다는 것이다.

# 근거는 얼마나 탄탄한가?

주장을 두 부분으로 나눠야 한다. 두 부분의 증거 강도가 매우 다르기 때문이다.

**능력 시연으로서**, 즉 언어모델 에이전트가 EHR 샌드박스에서 병력 → 검사 → 진단 → 오더를 잇는 전체 임상 사례를 끝까지 운영할 수 있다는 점은 실제로 새롭고 꽤 설득력 있다. 이것이 이 연구에서 새롭게 등장한 부분이고, 주목할 만한 부분이다.

**우월성 주장으로서**, 즉 MIRA 가 의사보다 낫다는 말은 범위가 좁으며 조심해서 읽어야 한다. 특정 과거 자료 벤치마크, 8 개 진단, 더 많은 검사를 사용하는 정보 비대칭, 역사적 차트에서 가져온 답안지, 학습 데이터 오염 가능성이라는 조건이 붙는다. ' 이 에이전트가 이 벤치마크에서 인상적인 성능을 보였다' 고 말하기에는 충분하다. ' 이 에이전트가 의사보다 더 좋은 진단자다' 라고 말하기에는 충분하지 않으며, 저자들도 두 번째 주장을 하지 않는다.

가장 유용한 태도는'AI 가 의사를 이겼다' 도' 그저 장난감이다' 도 아니다. 질문에 답하는 대신 기록 전체에서 행동하는 새로운 종류의 의료 AI 가 어려운 과거 자료 벤치마크에서 잘 작동했고, 이제 아직 시험하지 않은 곳에서 증명해야 한다는 것이다. 실제의 미분화된 환자들을 대상으로, 전향적으로, 거버넌스 아래에서 말이다.

# 왜 중요한가

지난 몇 년 동안 의료 AI 에서 흥미로운 질문은 조용히 바뀌었다. 예전에는 '모델이 진단을 맞힐 수 있는가?' 였고, 모델들은 점점 더 정제된 시험 세트에서 계속 '그렇다' 고 답했다. MIRA 는 더 어려운 질문으로 옮겨간다. '모델이 일을 할 수 있는가? 진료 과정 전체에서 정보를 모으고, 검사를 주문하고, 해석하고, 결정하고, 행동할 수 있는가?' 이 질문은 더 유용하고 더 정직하다. 실제로 행동하는 단계에 어려움과 위험이 함께 있기 때문이다.

그래서 결과를 어떻게 설명하느냐가 중요하다. 'AI 가 의사를 능가했다' 는 반사적인 제목은 결과 가운데 가장 덜 새롭고 가장 덜 뒷받침된 부분을 가리킨다. 진짜 새로운 것은 더 작지만 더 중요한 것이다. 기록 전체를 오가며 일할 수 있는 에이전트다. 이 능력이 실제 환경에서도 유지된다면 누가 책임자인지를 바꾸기 훨씬 전에 **진료 과정**을 바꿀 것이다. 의사가 사라지는 것이 아니라, 검사를 지시하고 결과를 해석하고 후속 조치를 하는 과정에 이런 도구가 먼저 들어올 가능성이 높다.

그리고 입증 책임을 올바른 방향으로 되돌린다. 과거 자료 벤치마크에서의 승리는 전향적 시험을 해야 할 이유이지 그 시험을 대신하는 것이 아니다. 저자들도 그렇게 말한다. MIRA 를 정직하게 읽으면 다음 연구를 요청하는 결과다. 의료 분야가 이미 알고 있는 기준대로, 벤치마크가 아니라 환자에게서 측정해야 한다.

# 핵심 요약

MIRA 는 실제 병원 시스템과 분리된 **EHR 샌드박스**에서 작동하는 자율 AI 에이전트다. 병력을 받고, 검사실·영상·미생물 검사를 주문·해석하고, 진단한 뒤 치료 계획과 오더를 작성할 수 있다. 공개 MIMIC-IV 데이터베이스의 과거 사례 574 건, 미리 선택된 8 개 진단에서 시험했을 때 진단 정확도는 전체 88.9% 였고

직접 비교에서는 87.8% 대 전문의 78.1% 로 의사보다 높았으며, 결정도 대체로 지침에 부합하고 약물 안전성이 높았다. 하지만 평가는 과거 기록을 이용한 텍스트 전용 시뮬레이션이었다. 우위의 상당 부분은 검사 결과가 명확한 질환에서 나왔고, MIRA 는 의사보다 훨씬 많은 혈액검사 항목을 사용했다 (이용 가능 항목의 약 51% 대 28%). 여러 치료 지표는 원래 차트에 기록된 행동을 기준으로 채점됐고, 공개 데이터셋이라는 점은 학습 데이터 오염 위험을 만든다. 저자들은 그 수치가 가능한 상한선일 수 있다고 직접 적었다. 진짜 진전은 고립된 질문에 답하는 대신 진료 과정 전체에서 행동하는 에이전트를 보여준 것이다. 일반화, 안전성, 거버넌스는 여전히 전향적 실제 연구가 필요하다. 따라서 이는 인상적인 능력 시연이지, AI 가 의사보다 진단을 잘한다는 증명도 실제 진료에 바로 쓸 시스템도 아니다.

# 과장 없이 보면

**논문이 보여주는 것:** 언어모델 에이전트 MIRA 가 샌드박스 EHR 안에서 병력, 검사, 진단, 오더를 포함한 전체 임상 사례를 끝까지 운영할 수 있고, 8 개 진단의 과거 사례 574 건 벤치마크에서 진단 정확도가 의사보다 높았다는 것 (전체 88.9%, 직접 비교에서 87.8% 대 전문의 78.1%). 고중증도 약물 오류도 보고되지 않았다.

**그럴듯하지만 입증되지 않은 것:** 이 능력이 실제의 미분화된 환자에게 이익으로 이어진다는 것, 정확도 우위가 더 많은 검사 주문·기록된 차트와의 일치·공개 데이터 암기보다 더 나은 임상 추론을 반영한다는 것.

**보여주지 않는 것:** 8 개 진단 밖에서도 작동한다는 것, 배치가 안전하다는 것, 동일한 정보를 가진 의사와의 공정한 비교에서 이긴다는 것, 임상의를 대체한다는 것, 학습 데이터 오염이 없다는 것.

**주요 한계:** 과거 자료를 이용한 시뮬레이션·텍스트 전용 평가, 미리 선택한 8 개 진단, 정보 비대칭 (저가 혈액검사에서 이용 가능 분석항목의 약 51% 대 28%; 영상검사 증가가 핵심은 아님), 치료 결과 일부를 역사적 차트에 맞춰 채점한 점, 그리고 언어 모델이 학습 중 봤을 수 있는 공개 벤치마크 MIMIC-IV. 저자들도 이를 가능한 성능 상한선이라고 경고한다.

**일반 독자는 얼마나 신뢰해도 될까?** MIRA 가 단순한 챗봇이 아니라 기록 전체에서 행동하는 진짜 새 능력 시연이라는 데에는 높은 신뢰를 둘 수 있다. MIRA 가 의사보다 더 좋은 진단자라는 주장에는 낮은 신뢰가 적절하다. 그 결론은 하나의 과거 자료 벤치마크에 한정되고 검사 주문량, 차트 기반 채점, 오염 가능성의 영향을 받는다. 저자들의 결론이 가장 정확하다. 환자 진료와 관련된 의미를 주장하려면 전향적 실제 임상 연구가 필요하다.

## 출처

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