ИИ-агент самостоятельно вёл целые клинические случаи — в симуляторе, на прошлых медицинских записях

MIRA — новый тип медицинского ИИ: вместо ответа на один вопрос система ведёт весь случай внутри симулированной больничной карты — собирает анамнез, назначает и читает исследования, приходит к диагнозу и формирует назначения. На 574 ретроспективных случаях из публичной базы, охватывавших восемь заранее выбранных диагнозов, авторы сообщают о более высокой, чем у врачей, точности диагноза и преимущественно безопасных решениях, согласующихся с рекомендациями. Но каждое уточнение здесь важно: система работала в песочнице на прошлых записях и только с текстом; значительная часть преимущества пришлась на состояния с однозначными тестами; она использовала примерно вдвое больше анализов крови, чем врачи; а несколько результатов оценивали относительно того, что было записано в исходной карте. Настоящий шаг вперёд — агент, действующий в рамках всего рабочего процесса, а не доказательство того, что машина теперь диагностирует лучше врача. Сами авторы подчёркивают: обобщаемость, безопасность и управление ещё требуют проспективных исследований в реальном мире.

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

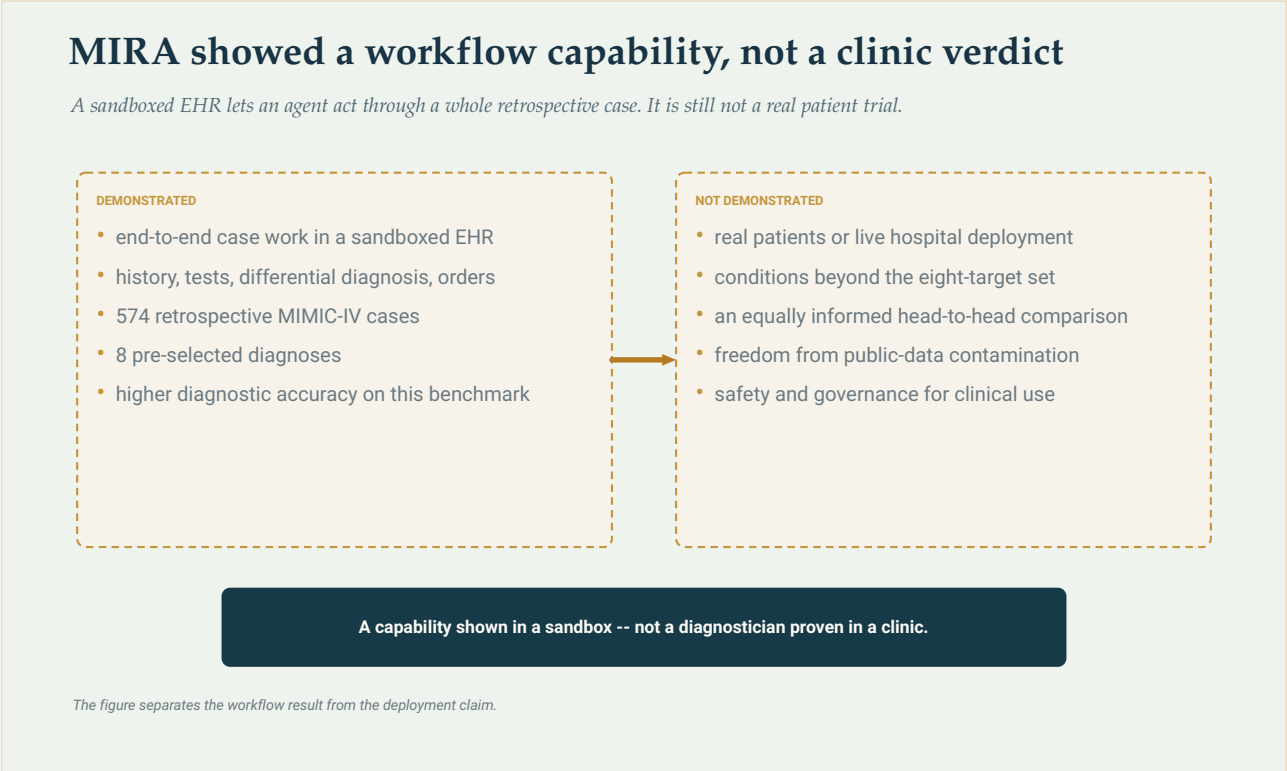
Ответить на вопрос — не то же самое, что вести весь клинический случай

Большинство историй о медицинском ИИ рассказывают о модели, которая *отвечает*. Ей дают клиническую задачу или экзаменационный вопрос, а она возвращает диагноз или абзац рекомендаций и набирает высокий балл. Это полезно, но узко: словно умный консультант, который никогда не прикасается к медицинской карте.

MIRA создана для другого, и именно эта разница здесь является настоящей новостью. Вместо того чтобы ответить на один вопрос, она обрабатывает весь случай: читает карту пациента, решает, каких данных анамнеза ещё не хватает, назначает лабораторные, визуализационные и микробиологические исследования, читает результаты, сужает дифференциальный диагноз, а затем формирует дальнейшие назначения — рецепты, госпитализацию, направление на операцию. И делает это, выполняя действия *внутри* электронной медицинской карты, как клиницист, а не просто генерируя текст, который человек должен перенести вручную.

Это реальный шаг: от системы, которая советует, к системе, которая действует в рамках всего рабочего процесса. И именно такой шаг в пересказе легко сплющивается до «ИИ превзошёл врачей». Поэтому важно точно сказать, что именно проверяли — и где.

Версия в одном предложении: MIRA самостоятельно вела случаи от начала до конца и на этом бенчмарке превзошла врачей по точности диагноза — но работала в изолированном симуляторе, на ретроспективных записях, только для восьми заранее выбранных диагнозов, общалась исключительно текстом, а значительная часть её преимущества пришлась на болезни с наиболее однозначными результатами тестов. Прогресс реален. Таблица результатов — не клиника.



MIRA продемонстрировала способность выполнять весь клинический рабочий процесс в «песочнице» электронной медицинской карты. Она не продемонстрировала реального клинического развёртывания, широкого диагностического охвата или справедливого сравнения с врачами при одинаковом объёме информации.

Что сделали авторы

MIRA — Medical Intelligence for Reasoning and Action — это автономный агент на моделях OpenAI: часть, которая общается и выполняет действия, работает на **GPT-4o**, а отдельный этап планирования использует рассуждающую модель OpenAI **o1**. Они встроены в специально созданную, совместимую со стандартами среду на основе HL7 FHIR — стандарта данных, с помощью которого реальные больницы обмениваются медицинскими записями, — с набором более восьмидесяти тысяч возможных клинических действий. Внутри этой «песочницы» MIRA может получить анамнез пациента, назначить и интерпретировать анализы, визуализацию и микробиологическ тесты, сформировать дифференциальный диагноз и план лечения — назначить лекарства, запланировать операцию или госпитализацию. Важная деталь: автоматизированным «судьями», оценивавшими ответы MIRA, были сами **GPT-4o** — то есть та же семья моделей, которую тестировали. После того как рецензент обратил внимание на эту проблему, авторы добавили проверку сертифицированным врачом.

Набор тестов взяли из **MIMIC-IV**, крупной общедоступной деперсонализированной базы электронных медицинских записей. Из примерно 300 000 пациентов, лечившихся

в Beth Israel Deaconess Medical Center между 2008 и 2019 годами, авторы отобрали **574 клинических случая** с **восемью целевыми диагнозами** — абдоминальными патологиями и неотложными состояниями внутренней медицины, среди которых аппендицит, панкреатит, пневмония и инфекция мочевых путей. Для каждого случая MIRA начинала с ограниченной картины и должна была шаг за шагом решать, что делать дальше — по форме это напоминало реальное обследование, но происходило на карте, реальный финал которой уже был записан.

Затем её результаты сравнили с работой врачей на тех же случаях.

Что на самом деле означает «автономный агент в песочнице EHR»

Здесь три слова несут очень много смысла, поэтому их стоит разобрать.

Агент означает, что система — не чатбот, отвечающий на один запрос. Она работает циклом: смотрит на текущее состояние, выбирает действие (назначить этот тест, спросить этот факт анамнеза), получает результат, выбирает следующее действие — и так до диагноза и плана. «Интеллект» здесь даёт языковая модель; *агентность* обеспечивает программная обвязка вокруг неё, которая превращает текст модели в разрешённые операции в EHR и возвращает результаты обратно.

Песочница EHR означает симулированную, изолированную копию системы медицинских записей, а не живую больничную систему. MIRA может «назначить» тест и получить результат, который у этого пациента действительно был, потому что случай исторический и ответ уже есть в карте. Ни одно её действие не затрагивает реального пациента или реальную клинику.

Автономный означает, что система завершает случай от начала до конца без человека в цикле — *внутри песочницы*. Это **не** означает, что ей разрешили без надзора вести настоящих пациентов. Это очень разные утверждения, и проверяли только первое.

Ещё одна важная деталь о песочнице: MIRA может *заказать* любой тест, но система способна вернуть результат только тогда, когда этот тест **действительно проводили** данному пациенту. Если она запрашивает анализ крови, который пациент реально сдавал, то получает реальные значения; если просит сканирование, которого никто не назначал, получает «N/A — невозможно выполнить», а не выдуманное число. С анамнезом так же: если в карте нет ответа на вопрос MIRA, симулированный пациент просто говорит, что не знает. Поэтому MIRA не может волшебным образом получить решающий тест, которого в реальном случае никогда не делали: она ограничена тем, что содержал фактический процесс обследования.

Это различие важно, потому что слово «автономный» проще всего ошибочно прочитать во втором смысле.

Что они обнаружили

На бенчмарке **MIRA превзошла врачей по точности диагноза** — но за громкой формулировкой стоят три разных числа, которые легко перепутать.

Самостоятельно, на всех **574 случаях**, MIRA назвала правильный диагноз в **88.9%** случаев — правильность определяли по выписному диагнозу, реально записанному для каждого пациента в MIMIC-IV. В прямом сравнении с людьми врачи работали не со всеми 574 случаями, а с общей **подвыборкой из 311 случаев**, имея те же записи и инструменты, что и MIRA. На этих 311 случаях MIRA получила **87.8%**. Её сравнивали с двумя отдельно набранными группами. Первая — **старшая группа**: четыре сертифицированных врача с 7–11 годами опыта, результат **78.1%**. Вторая — **группа смешанного уровня опыта**: преимущественно младшие резиденты плюс два специалиста, то есть состав, более похожий на реальное отделение неотложной помощи, — **71.1%**. Те же случаи, те же инструменты — и порядок получился таким: ИИ первый, опытные врачи вторые, команда с преимущественно молодыми врачами третья. Осторожная формулировка самой статьи: для всех восьми болезней MIRA была «стабильно эквивалентна врачам и часто превосходила их».

Дальнейшие решения тоже выглядели хорошо: в основном соответствовали клиническим рекомендациям, были безопасными в отношении лекарств и уместными в отношении госпитализации. Авторы не зафиксировали ошибок высокой тяжести в пяти категориях лекарственной безопасности — взаимодействия препаратов, дозирование при нарушении функции почек, аллергии, риск удлинения QT и назначение опиоидов — и сообщили о правильных назначениях в 467 из 468 случаев, одновременно прямо отметив, что система «не достигла 100% надёжности». Если воспринимать эти числа буквально, результат впечатляет: агент ведёт весь случай и опережает врачей по диагнозу.

Но *форма* этой победы не менее важна, чем сам факт. Независимые специалисты, читавшие статью, обратили внимание на то, откуда происходило преимущество. В экспертной панели Science Media Centre доктор Wei Xing (University of Sheffield) отметил, что главный результат — преимущество ИИ по точности диагноза — **в основном обусловлен состояниями с чёткими результатами тестов**, такими как аппендицит и панкреатит, где решающий снимок или лабораторное значение быстро закрывает вопрос. Для пневмонии и инфекций мочевых путей — двух очень распространённых причин обращения за неотложной помощью — *и* ИИ, *и* врачи показали худшие результаты, а разрыв между ними был наименьшим.

Была и асимметрия в том, сколько информации использовали обе стороны, и она видна из чисел самой статьи. MIRA значительно активнее полагалась на лабораторию: использовала около **51%** лабораторных показателей, доступных в обычной практике, против примерно **28%** у сертифицированных врачей — медианно на семь дополнительных показателей крови на один случай. Авторы осторожно подчёркивают, что это всё равно *меньше* фактического уровня рутинного использования тестов в наборе данных, а не стратегия «заказать всё», и не обнаружили систематического увеличения более дорогой томографической визуализации. Однако больше информации само по себе

может повысить точность диагноза. Поэтому это не совсем сравнение двух одинаково информированных участников — на что Xing прямо обратил внимание. Врач, которому позволить без ограничений назначать все дешёвые анализы крови, не учитывая стоимость, дискомфорт или задержку, тоже будет выглядеть точнее на бумаге.

Почему «победить врачей» не означает «быть лучшим врачом»

Победу на бенчмарке легко переоценить. Между «MIRA набрала больше» и «MIRA лучше» стоят как минимум три вещи.

Во-первых, **с чем сравнивали ответы**. Профессор Julie Jacko (University of Edinburgh) заметила, что несколько ключевых показателей MIRA определены относительно того, что было задокументировано в исходном наборе данных. То есть систему вознаграждают за **воспроизведение зафиксированного клинического поведения**, а не обязательно за демонстрацию оптимальной помощи. Историческая карта становится ключом с ответами. Для бенчмарка это разумный подход, но он измеряет соответствие тому, что было сделано, а не абсолютную правильность. Справедливости ради, эта проблема больше всего касается показателей *согласованности лечения* — совпадают ли назначения MIRA с картой, — тогда как основную диагностическую точность оценивали по выписному диагнозу, что ближе к реальному результату, чем простое повторение документации.

Во-вторых, уже упомянутая **асимметрия информации**: значительно больше анализов крови — около 51% доступных показателей против 28% — означает больше доказательств. Вероятно, часть разрыва в точности не «выведена рассуждением», а *куплена* дополнительными тестами.

В-третьих, **где именно работала система**. Это была песочница с ретроспективными случаями, восемью заранее отобранными диагнозами и исключительно текстовым интерфейсом. Реальная клиническая оценка зависит не только от того, *что* говорит пациент, но и от того, *как* он это говорит, а также от физикального осмотра, наблюдаемого поведения и языка тела — ничего из этого текстовый агент, читающий уже завершённую карту, не видит. А пациент, чья проблема не относится к восьми выбранным диагнозам, просто находится за пределами того, о чём может говорить это исследование.

Есть и более тихая проблема, поднятая рецензентами: **загрязнение обучающими данными**. MIMIC-IV — публичная база, о которой много пишут, поэтому языковая модель, обученная на открытом интернете, могла уже видеть статьи, разборы случаев или сами данные. Доктор Midhun Parakkal Unni (University of Sheffield) прямо указал на это: если часть ответов уже была в обучающих данных, то определённая доля результата может быть памятью, а не клиническим рассуждением, и отличить одно от другого может только независимое повторение. Важно, что авторы не отмахиваются от этой проблемы: они пишут, что результаты «можно осторожно интерпретировать

как возможную верхнюю границу» и что они «могут завышать обобщаемость на другие публичные случаи». Редкий случай, когда сама статья ставит потолок над собственным громким результатом.

Всё это не делает MIRA менее интересной. Оно лишь уточняет корректный вывод: на ретроспективном бенчмарке агент, способный действовать по всей карте, превзошёл врачей прежде всего на диагнозах с чёткими тестами — отчасти заказывая больше анализов, отчасти воспроизводя зафиксированный в карте ход. Это реальная демонстрация возможности, а не вердикт о том, что машины теперь диагностируют лучше врачей.

Чего эта работа *не* доказывает

- Она **не** показывает, что MIRA работает на реальных пациентах. Все случаи были ретроспективными и симулированными; ни одного пациента она не вела.
- Она **не** показывает, что система работает за пределами восьми диагнозов. Состояния вне заранее выбранного набора — то есть значительная часть сложной, недифференцированно медицины — не тестировались.
- Она **не** показывает, что систему безопасно развёртывать. Сами авторы пишут, что обобщаемость, безопасность и управление нужно проверить в проспективных исследованиях в реальном мире.
- Она **не** устанавливает справедливого прямого сравнения с врачами. MIRA использовала намного больше анализов крови (около 51% доступных показателей против 28%), а несколько результатов лечения частично оценивали по тому, что было записано в исторической карте, а не по независимому эталону наилучшей помощи.
- Она **не** доказывает отсутствие запоминания. Публичные данные MIMIC-IV могли пересекаться с обучающими данными модели; исключить это может только независимая репликация.
- Она **не** означает «ИИ заменяет врачей». Это система поддержки решений, способная *действовать* по всей карте; консенсус независимых комментаторов состоит в том, что реальное использование должно происходить в партнёрстве с клиницистами, которые сохраняют полномочия и добавляют всё то, чего нет в текстовой записи.

Насколько убедительны доказательства?

Здесь стоит разделить утверждение на две части, потому что сила доказательств для них очень различается.

Как демонстрация возможности — что агент на языковой модели может провести весь клинический случай в песочнице EHR, последовательно объединяя анамнез, тесты, диагноз и назначения, — работа действительно нова и достаточно убедительна. Именно это здесь новое, и именно на это стоит обратить внимание.

Как утверждение о превосходстве — что MIRA *лучше врачей* — доказательства имеют чёткие границы и требуют осторожного прочтения. Результат справедлив для конкретного ретроспективного бенчмарка с восемью диагнозами, асимметрией информации (больше тестов), ключом ответов из исторической карты и реальной возможностью загрязнения обучающими данными. Этого достаточно, чтобы сказать: «агент впечатляюще выступил на этом бенчмарке». Этого недостаточно, чтобы сказать: «агент — лучший диагност, чем врач», — и авторы второго не утверждают.

Самая полезная позиция — не «ИИ победил врачей» и не «это всего лишь игрушка». Точнее так: новый тип медицинского ИИ — тот, который *действует* по всей карте, а не просто отвечает на вопросы, — хорошо показал себя на сложном ретроспективном бенчмарке и теперь должен доказать себя там, где его ещё не испытывали: проспективно, на реальных недифференцированных пациентах и при надлежащем управлении.

Почему это важно

За последние несколько лет интересный вопрос в медицинском ИИ незаметно изменился. Раньше он звучал: «может ли модель правильно поставить диагноз?» — и модели всё чаще отвечали «да» на всё более чистых тестовых наборах. MIRA обозначает переход к более сложному вопросу: «может ли модель *выполнять работу* — собирать данные, назначать, интерпретировать, решать и действовать — в рамках всего рабочего процесса?» Это более полезный и честный вопрос, потому что именно в действии живут и сложность, и риск.

Именно поэтому рамка важна. Напрашивающийся заголовок — *ИИ превзошёл врачей* — указывает на наименее новую и слабее всего поддержанную часть результата. Действительно новое здесь скромнее по формулировке, но важнее по последствиям: агент, способный двигаться через всю карту. Если эта возможность выдержит дальнейшую проверку, она, вероятно, изменит **рабочий процесс** задолго до того, как изменится роль человека, который им управляет. Врач не исчезает; первое место, где окажется такой инструмент, — назначения, интерпретация, отслеживание результатов, то есть сам рабочий процесс.

И это правильно меняет бремя доказательства. Победа в таблице ретроспективных случаев — причина провести проспективное испытание, а не его замена. Авторы говорят то же самое. Честное прочтение MIRA — это приглашение к следующему исследованию, проведённому по стандарту, который медицина уже знает: на пациентах, а не только на бенчмарках.

Краткий итог

MIRA — автономный ИИ-агент, работающий в изолированной электронной медицинской карте: он может собирать анамнез, назначать и интерпретировать лабораторные, визуализационные и микробиологические исследования, устанавливать диагноз и формировать план лечения. На 574 ретроспективных случаях из публичной базы MIMIC-IV, охватывавших восемь заранее выбранных диагнозов, система превзошла врачей по точности диагноза (88.9% в целом; в прямом сравнении 87.8% против 78.1%) и преимущественно принимала решения, согласующиеся с рекомендациями и безопасные в отношении лекарств. Но оценка была симуляцией на прошлых записях и только в тексте; значительная часть преимущества пришлась на состояния с однозначными тестами; MIRA использовала гораздо больше анализов крови, чем врачи (около 51% доступных показателей против 28%); несколько результатов лечения оценивали относительно того, что было записано в исходной карте; а публичность набора данных создаёт реальный риск загрязнения обучающими данными — то, что сами авторы называют возможной верхней границей своих результатов. Настоящий шаг вперёд — агент, действующий по всему рабочему процессу, а не отвечающий на изолированные вопросы. Авторы прямо пишут, что обобщаемость, безопасность и управление ещё требуют проспективных исследований в реальном мире. Именно так результат и стоит читать: впечатляющая демонстрация возможности, а не доказательство того, что ИИ диагностирует лучше врача, и не система, готовая к реальной клинике.

Проверка без прикрас

Что показывает статья: Агент на языковой модели (MIRA) может провести весь клинический случай от начала до конца в песочнице EHR — анамнез, тесты, диагноз, назначения — и на ретроспективном бенчмарке из 574 случаев и восьми диагнозов превзошёл врачей по точности диагноза (88.9% в целом; 87.8% против 78.1% в сравнении с сертифицированными врачами), не допустив ошибок высокой тяжести в назначении лекарств.

Что правдоподобно, но не доказано: Что эта возможность принесёт пользу реальным недифференцированным пациентам; что преимущество в точности отражает лучшее рассуждение, а не большее число назначенных тестов, согласование с исторической картой или запомненные публичные данные.

Чего она не показывает: Что MIRA работает за пределами восьми диагнозов; что её безопасно развёртывать; что она побеждает врачей в справедливом сравнении при одинаковой информации; что она заменяет клиницистов; что результаты гарантированно свободны от загрязнения обучающими данными.

Главные ограничения: Ретроспективная, симулированная и только текстовая оценка;

восемь заранее выбранных диагнозов; асимметрия информации (значительно больше анализов крови — около 51% доступных показателей против 28%, причём речь идёт о дешёвых анализах крови, а не более дорогой визуализации); показатели лечения частично оценивали по исторической карте; публичный бенчмарк MIMIC-IV, который языковая модель могла видеть во время обучения — и сами авторы называют это возможной верхней границей результативности.

Какой уверенности заслуживает вывод для обычного читателя? Высокой в том, что MIRA — реальная и новая демонстрация возможности: это агент, который *действует* по всей карте, а не просто чатбот. Низкой в отношении утверждения, что он *лучший диагност, чем врач*: оно ограничено одним ретроспективным бенчмарком и смешано с эффектом большего числа тестов, оценкой по карте и возможным загрязнением данных. Собственный вывод авторов здесь лучший: прежде чем это будет что-то значить для помощи пациентам, нужно проспективное исследование в реальном мире.

Источники

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