AI-აგენტმა პაციენტის სრული შემთხვევები დამოუკიდებლად დაამუშავა — სიმულატორში, წარსულ ჩანაწერებზე

MIRA სამედიცინო AI-ის განსხვავებული ტიპია: ერთი კითხვის პასუხის ნაცვლად ის სიმულირებულ საავადმყოფოს ჩანაწერში მთელ შემთხვევას მართავს — იღებს ანამნეზს, ნიშნავს და კითხულობს კვლევებს, სვამს დიაგნოზს და წერს დანიშნულებებს. საჯარო მონაცემთა ბაზიდან აღებულ 574 წარსულ შემთხვევაზე, რვა წინასწარ შერჩეული დიაგნოზის ფარგლებში, ავტორები აღწერენ ექიმებზე მაღალ დიაგნოსტიკურ სიზუსტეს და მეტწილად გაიდლაინებთან შესაბამის, მედიკამენტურად უსაფრთხო გადაწყვეტილებებს. მაგრამ ყველა შეზღუდვა მნიშვნელოვანია: სისტემა მუშაობდა იზოლირებულ სიმულატორში, წარსულ ჩანაწერებზე და მხოლოდ ტექსტით; უპირატესობის დიდი ნაწილი მკაფიო კვლევითი შედეგების მქონე დაავადებებზე მოდიოდა; ის ექიმებზე დაახლოებით ორჯერ მეტ სისხლის მაჩვენებელს იყენებდა; რამდენიმე შედეგი ისტორიულ ჩანაწერთან შესაბამისობით ფასდებოდა. რეალური წინსვლა არის აგენტი, რომელიც მთელ სამუშაო პროცესში მოქმედებს — და არა მტკიცება, რომ მანქანა უკვე ექიმზე უკეთ სვამს დიაგნოზს. თავად ავტორები ამბობენ, რომ განზოგადებას, უსაფრთხოებას და მმართველობას ჯერ კიდევ სჭირდება პერსპექტიული რეალური კვლევები.

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

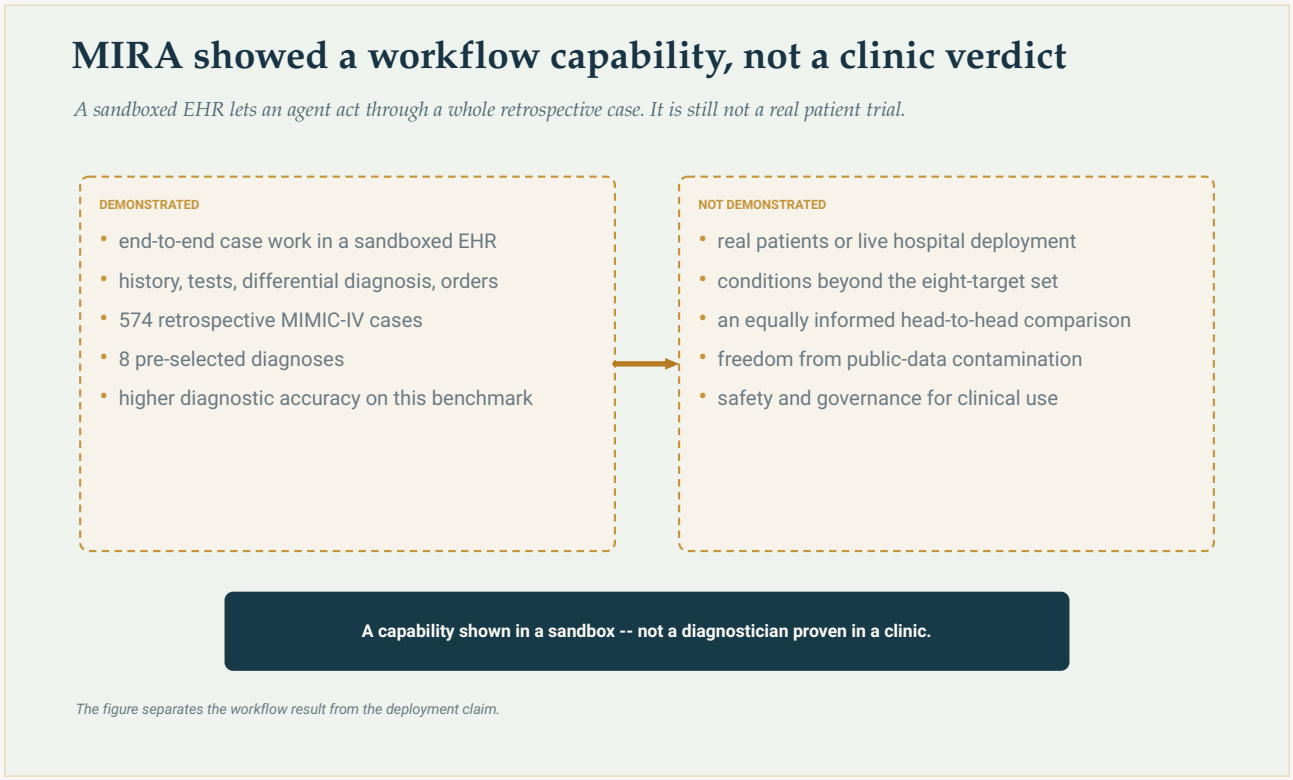
კითხვაზე პასუხი და მთელი კლინიკური შემთხვევი მართვა ერთი და იგივე არ არის

სამედიცინო AI-ზე ამბების უმეტესობა იმ მოდელებზეა, რომლებიც *პასუხობენ*. აწვდით კლინიკურ აღწერას ან საგამოცდო კითხვას, მოდელი აბრუნებს დიაგნოზს ან რჩევის აბზაცს და მაღალ ქულას იღებს. ეს სასარგებლოა, მაგრამ ვიწრო ამოცანაა: ჭკვიანი კონსულტანტი, რომელიც პაციენტის ჩანაწერს თვითონ არასოდეს მართავს.

MIRA სხვა რამისთვის შეიქმნა, და სიახლეც სწორედ ეს განსხვავებაა. ერთი კითხვის პასუხის ნაცვლად ის მთელ შემთხვევას ამუშავებს: კითხულობს პაციენტის ჩანაწერს, წყვეტს კიდევ რა ანამნეზი სჭირდება, ნიშნავს ლაბორატორიულ, გამოსახულებით და მიკრობიოლოგიურ კვლევებს, კითხულობს შედეგებს, ავიწროებს დიფერენციულ დიაგნოზს და შემდეგ საჭირო დანიშნულებებს წერს — რეცეპტებს, ჰოსპიტალიზაციას, ქირურგთან მიმართვას. ამას აკეთებს ელექტრონული სამედიცინო ჩანაწერის **შიგნით მოქმედებების შესრულებით**, ისე როგორც კლინიცისტი, და არა უბრალოდ თავისუფალი ტექსტის გენერირებით, რომლის ხელით გადატანაც ადამიანს მოუწევს.

ეს რეალური ნაბიჯია: რჩევის გამცემი სისტემიდან სამუშაო პროცესში მოქმედ სისტემამდე. თუმცა სწორედ ასეთი ნაბიჯი იკუმშება მედიაში მარტივ ფრაზად — „AI ექიმებს ჯობნის“. ამიტომ მნიშვნელოვანია ზუსტად ვთქვათ, რა გამოცადეს და სად.

ერთ წინადადებაში: MIRA-მ დამოუკიდებლად დაამუშავა სრული შემთხვევები და ამ ბენჩმარკზე დიაგნოსტიკური სიზუსტით ექიმებს აჯობა — მაგრამ ეს გააკეთა იზოლირებულ სიმულატორში, წარსულ ჩანაწერებზე, წინასწარ შერჩეულ რვა დიაგნოზზე, მხოლოდ ტექსტური კომუნიკაციით; უპირატესობის მნიშვნელოვანი ნაწილი კი იმ დაავადებებზე მოდიოდა, სადაც კვლევის შედეგები განსაკუთრებით მკაფიოა. წინსვლა რეალურია. ბენჩმარკის ცხრილი კლინიკა არ არის.



MIRA-მ იზოლირებულ EHR გარემოში თავიდან ბოლომდე კლინიკური სამუშაო პროცესის წარმართვის უნარი აჩვენა. კვლევამ არ აჩვენა რეალურ კლინიკაში დანერგვა, ფართო დიაგნოსტიკური დაფარვა ან ექიმებთან თანაბარი ინფორმაციის პირობებში სამართლიანი პირდაპირი შედარება.

რა გააკეთეს ავტორებმა

MIRA — Medical Intelligence for Reasoning and Action — OpenAI-ის მოდელებზე აგებული ავტონომიური აგენტია. საუბრისა და მოქმედებების ნაწილი **GPT-4o**-ზე მუშაობს, ხოლო ცალკე დაგეგმვის საფეხური OpenAI-ის **o1** მსჯელობის მოდელს იყენებს. სისტემა მოთავსებულია სპეციალურად შექმნილ, სტანდარტებთან თავსებად ჩარჩოში — HL7 FHIR-ზე, მონაცემთა სტანდარტზე, რომელსაც რეალური საავადმყოფოები ჩანაწერების გადასაცემად იყენებენ — და 80,000-ზე მეტი შესაძლო კლინიკური მოქმედების ინსტრუმენტარი აქვს. ამ იზოლირებულ გარემოში MIRA-ს შეუძლია ანამნეზის ამოღება, ლაბორატორიული, გამოსახულებითი და მიკრობიოლოგიური კვლევების დანიშვნა და ინტერპრეტაცია, დიფერენციული დიაგნოზის შედგენა და მკურნალობის გეგმის ფორმირება — მედიკამენტის დანიშვნა, ოპერაციის დაგეგმვა ან ჰოსპიტალიზაციის გადაწყვეტილება. ერთი დეტალი თავიდანვე აღნიშვნის ღირსია: MIRA-ს პასუხებს ავტომატურად ისევ GPT-4o აფასებდა — იგივე მოდელური ოჯახის სისტემა, რომელსაც თავად MIRA იყენებს. რეცენზენტის შენიშვნის შემდეგ ავტორებმა ეს შეფასება სერტიფიცირებული ექიმის შემოწმებითაც გაამყარეს.

სატესტო შემთხვევები **MIMIC-IV**-დან აიღეს — დიდი, საჯაროდ ხელმისაწვდომი,

დეიდენტიფიცირებული ელექტრონული სამედიცინო ჩანაწერების ბაზიდან. Beth Israel Deaconess Medical Center-ში 2008–2019 წლებში ნამკურნალები დაახლოებით 300,000 პაციენტიდან ავტორებმა შეარჩიეს **574 შემთხვევა**, რომლებიც **რვა სამიზნე დიაგნოზს** მოიცავდა — მუცლის პათოლოგიებს და შინაგანი მედიცინის გადაუდებელ მდგომარეობებს, მათ შორის აპენდიციტს, პანკრეატიტს, პნევმონიას და საშარდე გზების ინფექციას. თითოეულ შემთხვევაში MIRA შეზღუდული საწყისი ინფორმაციით იწყებდა და ეტაპობრივად თვითონ წყვეტდა, შემდეგ რა გაეკეთებინა — რეალური დიაგნოსტიკური პროცესის ფორმით, თუმცა ისტორიულ ჩანაწერში შემთხვევის რეალური დასასრული უკვე ცნობილი იყო.

შემდეგ მისი შედეგები იმავე შემთხვევებზე ექიმების შედეგებს შეადარეს.

რას ნიშნავს სინამდვილეში „ავტონომიური აგენტი იზოლირებულ EHR-ში“

აქ სამი სიტყვა დიდ დატვირთვას ატარებს და მათი გაშლა ღირს.
Agent ნიშნავს, რომ სისტემა უბრალოდ chatbot არ არის, რომელიც ერთ მოთხოვნას პასუხობს. ის ციკლურად მუშაობს: უყურებს მიმდინარე მდგომარეობას, ირჩევს მოქმედებას — მაგალითად კვლევის დანიშვნას ან დამატებითი ანამნეზის კითხვას — იღებს შედეგს და შემდეგ ნაბიჯს ირჩევს, სანამ დიაგნოზსა და გეგმამდე მივა. „ინტელექტი“ ენობრივი მოდელია; *აგენტურობა* კი მის გარშემო არსებული პროგრამული ჩარჩოა, რომელიც მოდელის ტექსტს ნებადართულ EHR მოქმედებებად გარდაქმნის და შედეგებს უკან აწვდის.
Sandboxed EHR ნიშნავს სამედიცინო ჩანაწერების სისტემის სიმულირებულ, გარესამყაროსგან იზოლირებულ ასლს და არა მოქმედ საავადმყოფოს სისტემას. MIRA-ს შეუძლია კვლევა „დანიშნოს“ და მიიღოს ის შედეგი, რომელიც ამ პაციენტს რეალურად ჰქონდა, რადგან შემთხვევა ისტორიულია და პასუხი ჩანაწერში უკვე არსებობს. მისი არც ერთი მოქმედება რეალურ პაციენტს ან კლინიკას არ ეხება.
Autonomous ნიშნავს, რომ აგენტი შემთხვევას თავიდან ბოლომდე ადამიანის უშუალო ჩარევის გარეშე ასრულებს — **იზოლირებული გარემოს შიგნით**. ეს არ ნიშნავს, რომ სისტემას რეალური პაციენტების ზედამხედველობის გარეშე მართვის უფლება მისცეს. კვლევამ მხოლოდ პირველი რამ გამოსცადა.
კიდევ ერთი მნიშვნელოვანი დეტალი: MIRA-ს შეუძლია ნებისმიერი კვლევა „მოითხოვოს“, მაგრამ სისტემა შედეგს მხოლოდ მაშინ უბრუნებს, თუ ეს კვლევა ამ პაციენტს რეალურ ცხოვრებაში **მართლა ჩაუტარდა**. თუ პაციენტს სისხლის ანალიზი ჰქონდა, მოდელი ნამდვილ მნიშვნელობებს იღებს; თუ სკანი არასოდეს ჩატარებულა, პასუხია „N/A — could არა be performed“ და არა გამოგონილი რიცხვი. ანამნეზიც ასე მუშაობს: თუ მოთხოვნილი ინფორმაცია ჩანაწერში არ არის, სიმულირებული პაციენტი უბრალოდ პასუხობს, რომ არ იცის. ამიტომ MIRA ვერ „გამოიძახებს“ გადამწყვეტ კვლევას, რომელიც რეალურ დიაგნოსტიკაში არ ჩატარებულა — მას მხოლოდ იმ მასალაზე შეუძლია მუშაობა, რაც ისტორიულ პროცესში არსებობდა.
ეს განსხვავება მნიშვნელოვანია, რადგან სათაურებში სიტყვა „ავტონომიური“ ყველაზე ადვილად შეიძლება მეორე, ბევრად უფრო ძლიერი მნიშვნელობით

> გაიგონ.

რა აღმოაჩინეს

ბენჩმარკზე **MIRA-მ დიაგნოსტიკური სიზუსტით ექიმებს აჯობა** — მაგრამ სათაური სამ განსხვავებულ რიცხვს ადვილად აერთიანებს, ამიტომ მათი განცალკევებაა საჭიროა.

ყველა **574 შემთხვევაზე** MIRA-მ სწორი დიაგნოზი **88.9%** შემთხვევაში დაასახელა, შეფასება კი MIMIC-IV-ში რეალურად ჩაწერილ გაწერის დიაგნოზს ეყრდნობოდა. ადამიანებთან პირდაპირი შედარებისთვის ექიმებს ყველა 574 შემთხვევა არ გაუვლიათ; მათ იმუშავეს საერთო **311 შემთხვევის ქვეჯგუფზე**, იმავე ჩანაწერებითა და ინსტრუმენტებით, რაც MIRA-ს ჰქონდა. ამ 311 შემთხვევაზე MIRA-ს მაჩვენებელი **87.8%** იყო და მას ორ ცალკე შერჩეულ ექიმთა ჯგუფს ადარებდნენ. პირველი იყო **გამოცდილი ჯგუფი**: ოთხი board-certified ექიმი 7–11-წლიანი გამოცდილებით, შედეგი **78.1%**. მეორე იყო **შერეული გამოცდილების ჯგუფი**: ძირითადად ახალგაზრდა რეზიდენტები და ორი სპეციალისტი, შედეგი **71.1%**. ერთი და იგივე შემთხვევები, ერთი და იგივე ინსტრუმენტები — და შედეგების რიგი იყო: AI პირველი, გამოცდილი ექიმები მეორე, ახალგაზრდა წევრებით დატვირთული გუნდი მესამე. თავად ნაშრომი უფრო ფრთხილად ამბობს, რომ რვავე დაავადებაზე MIRA ექიმების შედეგებს „თანმიმდევრულად უტოლდებოდა და ხშირად აჭარბებდა“.

შემდგომი გადაწყვეტილებებიც მეტწილად გაიდლაინებთან შესაბამისი, მედიკამენტურად უსაფრთხო და ჰოსპიტალიზაციის თვალსაზრისით მისაღები იყო. ავტორები ხუთ უსაფრთხოების კატეგორიაში — წამლების ურთიერთქმედება, თირკმლის ფუნქციაზე დოზის მორგება, ალერგიები, QT-რისკი და ოპიოიდების დანიშვნა — მძიმე მედიკამენტურ შეცდომებს არ აღწერენ; 468 დანიშნულებიდან 467 სწორი იყო. თუმცა ისინი თავადაც აღნიშნავენ, რომ სისტემამ „100%-იან საიმედოობას ვერ მიაღწია“. ზედაპირულად ეს მართლაც შთამბეჭდავია: აგენტი მთლიან შემთხვევას მართავს და დიაგნოზით ექიმებზე უკეთ შედეგს იღებს.

მაგრამ მნიშვნელოვანია არა მხოლოდ გამარჯვება, არამედ **სად და როგორ** გაჩნდა უპირატესობა. დამოუკიდებელმა სპეციალისტებმა ყურადღება სწორედ ამას მიაქციეს. Science Media Centre-ის ექსპერტთა შეფასებაში Dr Wei Xing-მა (University of Sheffield) აღნიშნა, რომ დიაგნოსტიკური უპირატესობა **ძირითადად იმ დაავადებებმა განაპირობა, სადაც კვლევის შედეგები მკაფიოა**, მაგალითად აპენდიციტმა და პანკრეატიტმა, სადაც გადამწყვეტი სკანი ან ლაბორატორიული მაჩვენებელი საკითხს ხშირად აშკარად წყვეტს. პნევმონიასა და საშარდე გზების ინფექციებზე — გადაუდებელი დახმარების ორი ძალიან გავრცელებული მიზეზი — როგორც AI-ს, ისე ექიმებს ყველაზე ცუდი შედეგები ჰქონდათ და მათ შორის სხვაობაც ყველაზე მცირე იყო.

ასევე არსებობდა ინფორმაციის გამოყენების ასიმეტრია. MIRA ლაბორატორიულ

მონაცემებს უფრო ინტენსიურად იყენებდა: რუტინულ კლინიკურ პროცესში ხელმისაწვდომი ანალიზების დაახლოებით **51%**-ს, მაშინ როცა board-certified ექიმები დაახლოებით **28%**-ს — მედიანურად, თითო შემთხვევაზე დაახლოებით შვიდი დამატებითი სისხლის მაჩვენებელი. ავტორები სწორად აღნიშნავენ, რომ ეს მაინც *ნაკლებია* იმაზე, რასაც თავად მონაცემთა ნაკრებში რუტინული პრაქტიკა აჩვენებს, ანუ სტრატეგია „ყველაფერი დავნიშნოთ“ არ ყოფილა; ძვირადღირებული cross-sectional imaging-ის სისტემური მატებაც არ დაფიქსირებულა. მაგრამ მეტი ინფორმაცია თავისთავადაც შეიძლება ზრდიდეს დიაგნოსტიკურ სიზუსტეს. ამიტომ ეს თანაბრად ინფორმირებულ მონაწილეებს შორის სრულიად ერთნაირი პირობების შეჯიბრი არ არის. ექიმიც, რომელსაც ყველა იაფი სისხლის ანალიზის დანიშვნა შეეძლებოდა ფასის, დისკომფორტისა და დაყოვნების გათვალისწინების გარეშე, ქაღალდზე უფრო ზუსტი გამოჩნდებოდა.

რატომ არ ნიშნავს „ექიმებს აჯობა“ იმას, რომ „უკეთესი ექიმია“

ბენჩმარკში გამარჯვების ზედმეტად ფართოდ წაკითხვა ადვილია. „MIRA-მ მეტი ქულა აიღო“-სა და „MIRA უკეთესია“-ს შორის სამი მნიშვნელოვანი უფსკრულია.

პირველი: **რას ადარებდნენ მის პასუხებს.** Professor Julie Jacko-მ (University of Edinburgh) აღნიშნა, რომ MIRA-ს რამდენიმე ძირითადი შედეგი underlying dataset-ში დოკუმენტირებულ მოქმედებებს ეყრდნობა — ანუ სისტემა ჯილდოვდება **ჩაწერილი კლინიკური ქცევის გამეორებისთვის** და არა აუცილებლად ოპტიმალური მკურნალობის დემონსტრირებისთვის. ისტორიული ჩანაწერი პასუხების გასაღებად იქცევა. ბენჩმარკისთვის ეს გონივრული არჩევანია, მაგრამ ის ზომავს შესაბამისობას იმასთან, რაც გაკეთდა და არა აბსოლუტურ კლინიკურ სისწორეს. სამართლიანობისთვის უნდა ითქვას, რომ ეს პრობლემა ყველაზე მეტად მკურნალობის შესაბამისობის მეტრიკებს ეხება — დაემთხვა თუ არა MIRA-ს დანიშნულება ჩანაწერს — ხოლო მთავარი დიაგნოსტიკური სიზუსტე გაწერის დიაგნოზს ედრება, რაც უბრალო დოკუმენტაციის გამეორებაზე უფრო შედეგზე ორიენტირებული სამიზნეა.

მეორე: უკვე ნახსენები **ინფორმაციის ასიმეტრია**. გაცილებით მეტი სისხლის ანალიზი — ხელმისაწვდომი მაჩვენებლების დაახლოებით 51% 28%-ის წინააღმდეგ — ნიშნავს მეტ მტკიცებულებას. დიაგნოსტიკური სხვაობის ნაწილი შეიძლება დამატებითი ინფორმაციით იყოს მიღებული და არა მხოლოდ უკეთესი reasoning-ით.

მესამე: **სად ჩატარდა შეფასება.** ეს იყო იზოლირებული გარემო, წარსული შემთხვევები, რვა წინასწარ შერჩეული დიაგნოზი და მხოლოდ ტექსტი. როგორც Dr Dominic Oliver-მა (University of Oxford) აღნიშნა, რეალური კლინიკური შეფასება დამოკიდებულია არა მხოლოდ იმაზე, რას ამბობს პაციენტი, არამედ როგორ ამბობს, ასევე ფიზიკურ გამოკვლევაზე, ქცევაზე და სხეულის ენაზე — ამას დასრულებული ჩანაწერის მკითხველი ტექსტური აგენტი ვერ ხედავს. ხოლო პაციენტი, რომლის პრობლემა ამ რვა დიაგნოზში არ შედის, საერთოდ სცდება კვლევის დასკვნების ფარგლებს.

რეცენზენტებმა უფრო ჩუმი პრობლემაც წამოჭრეს: **სასწავლო მონაცემებით**

დაბინძურების რისკი. MIMIC-IV საჯაროა და მასზე ბევრი რამაა დაწერილი, ამიტომ ღია ინტერნეტზე გაწვრთნილ ენობრივ მოდელს შესაძლოა უკვე ენახა სტატიები, შემთხვევების განხილვები ან თავად მონაცემები. Dr Midhun Parakkal Unni-მ (University of Sheffield) ეს პირდაპირ აღნიშნა: თუ პასუხების ნაწილი სასწავლო მონაცემებში იყო, შედეგის ნაწილი შეიძლება მეხსიერებიდან მოდიოდეს და არა კლინიკური reasoning-იდან; ამის გარჩევა დამოუკიდებელ გამეორებას სჭირდება. ავტორები საკითხს არ უარყოფენ: წერენ, რომ შედეგები „ფრთხილად შეიძლება შეფასდეს, როგორც შესაძლო ზედა ზღვარი“ და შესაძლოა „გადააჭარბოს სხვა საჯარო შემთხვევებზე განზოგადების უნარს“ — ანუ ნაშრომი საკუთარ სათაურისეულ რიცხვებს თვითონვე უწესებს ჭერს.

ეს ყველაფერი MIRA-ს ნაკლებად საინტერესოს არ ხდის. უბრალოდ სწორი ინტერპრეტაცია უფრო ზუსტია: წარსულ ჩანაწერებზე აგებულ ბენჩმარკში აგენტმა, რომელსაც მთელი ჩანაწერის ფარგლებში მოქმედება შეეძლო, მკაფიო კვლევითი ნიშნების მქონე დიაგნოზებზე ექიმებზე უკეთ იმუშავა — ნაწილობრივ მეტი კვლევის გამოყენებით, ნაწილობრივ ისტორიულ ჩანაწერთან შესაბამისობით. ეს რეალური შესაძლებლობის დემონსტრაციაა და არა განაჩენი, რომ მანქანა ახლა ექიმზე უკეთ სვამს დიაგნოზს.

რას არ ამტკიცებს ეს

- კვლევა **არ** აჩვენებს, რომ MIRA რეალურ პაციენტებზე მუშაობს. ყველა შემთხვევა წარსული და სიმულირებული იყო; სისტემას არც ერთი რეალური პაციენტი არ უმართავს.
- ის **არ** აჩვენებს მუშაობას რვა დიაგნოზის მიღმა. წინასწარ შერჩეულ ჩამონათვალს გარეთ არსებული მდგომარეობები — მედიცინის რთული, არადიფერენცირებული უმრავლესობა — არ გამოუცდიათ.
- ის **არ** აჩვენებს, რომ სისტემა უსაფრთხოა დასანერგად. თავად ავტორები წერენ, რომ განზოგადება, უსაფრთხოება და მმართველობა პერსპექტიულ, რეალურ კვლევებს მოითხოვს.
- ის **არ** ადგენს ექიმებთან სრულად სამართლიან პირდაპირ შედარებას. MIRA ბევრად მეტ სისხლის ანალიზს იყენებდა — ხელმისაწვდომი მაჩვენებლების დაახლოებით 51% 28%-ის წინააღმდეგ — ხოლო მკურნალობის რამდენიმე შედეგი ნაწილობრივ ისტორიულ ჩანაწერთან შესაბამისობით ფასდებოდა და არა საუკეთესო შესაძლო მოვლის აბსოლუტური სტანდარტით.
- ის **არ** გამორიცხავს დამახსოვრებას. საჯარო MIMIC-IV შესაძლოა მოდელის სასწავლო მონაცემებს გადაფარავდეს; ამის გამორიცხვას დამოუკიდებელი replication სჭირდება.
- ის **არ** ნიშნავს „AI ექიმებს ჩაანაცვლებს“. ეს გადაწყვეტილების მხარდამჭერი აგენტია, რომელსაც ჩანაწერში *მოქმედება* შეუძლია; ექსპერტების ხედვით რეალური გამოყენება კლინიცისტებთან პარტნიორობაში უნდა მოხდეს, სადაც საბოლოო უფლებამოსილება ადამიანს რჩება და ის აწვდის ყველაფერს, რასაც ტექსტური ჩანაწერი ვერ შეიცავს.

რამდენად ძლიერია მტკიცებულება?

განცხადება ორ ნაწილად გავყოთ, რადგან მათ უკან მტკიცებულების ძალა ძალიან განსხვავებულია.

როგორც შესაძლებლობის დემონსტრაცია — რომ ენობრივ მოდელზე აგებულ აგენტს შეუძლია EHR-ის იზოლირებულ გარემოში მთელი კლინიკური შემთხვევის წარმართვა, ანამნეზის, კვლევების, დიაგნოზისა და დანიშნულებების ერთმანეთზე გადაბმა — ნაშრომი ნამდვილად ახალი და საკმაოდ დამაჯერებელია. სწორედ ეს ნაწილია ახალი და სწორედ მასზე ღირს ყურადღება.

როგორც უპირატესობის მტკიცება — რომ MIRA *ექიმებზე უკეთესია* — მტკიცებულება ბევრად უფრო შეზღუდულია. შედეგი ეხება ერთ კონკრეტულ retrospective საკონტროლო ნიშნული-ს, რვა დიაგნოზს, ინფორმაციის ასიმეტრიას — უფრო მეტ კვლევას — და პასუხების გასაღებად ნაწილობრივ ისტორიულ ჩანაწერს; ამასთან, სასწავლო მონაცემების შესაძლო გადაფარვაც არსებობს. ეს საკმარისია სათქმელად: „აგენტი ამ ბენჩმარკზე შთამბეჭდავად მუშაობდა.“ ეს საკმარისი არ არის სათქმელად: „აგენტი ექიმზე უკეთესი დიაგნოსტია.“ ავტორებიც მეორე მტკიცებას არ აკეთებენ.

ყველაზე სასარგებლო პოზიცია არც „AI ექიმებს სჯობნის“ არის და არც „ეს უბრალოდ სათამაშოა“. უფრო ზუსტად: სამედიცინო AI-ის ახალმა ტიპმა — რომელიც კითხვებზე პასუხის ნაცვლად ჩანაწერში *მოქმედებს* — რთულ წარსულ ბენჩმარკზე კარგად იმუშავა და ახლა უნდა დაამტკიცოს თავი იქ, სადაც ჯერ არ გამოუცდიათ: რეალურ, წინასწარ არაშერჩეულ პაციენტებზე, პერსპექტიულად და მკაფიო მმართველობის პირობებში.

რატომ აქვს მნიშვნელობა

ბოლო რამდენიმე წელიწადში სამედიცინო AI-ის მთავარი კითხვა ჩუმად შეიცვალა. ადრე კითხვა იყო: „შეუძლია მოდელს სწორი დიაგნოზის დასმა?“ — და მოდელები სულ უფრო სუფთა ტესტურ ნაკრებებზე ხშირად პასუხობდნენ „კი“. MIRA უფრო რთულ კითხვაზე გადასვლას აღნიშნავს: „შეუძლია მოდელს *საქმის გაკეთება* — ინფორმაციის შეგროვება, კვლევების დანიშვნა, შედეგების ინტერპრეტაცია, გადაწყვეტილება და მოქმედება — მთელი სამუშაო პროცესის განმავლობაში?“ ეს უფრო სასარგებლო და უფრო გულწრფელი კითხვაა, რადგან სწორედ მოქმედებაშია თავმოყრილი როგორც სირთულე, ისე რისკი.

ამიტომაა მნიშვნელოვანი სწორად ჩამოყალიბება. სათაურის ავტომატური რეფლექსი — *AI ექიმებს სჯობნის* — შედეგის ყველაზე ნაკლებად ახალ და ყველაზე ნაკლებად გამყარებულ ნაწილს უსვამს ხაზს. ნამდვილად ახალი რამ უფრო მცირე, მაგრამ უფრო მნიშვნელოვანი ცვლილებაა: აგენტი, რომელსაც მთელი ჩანაწერის პროცესში გადაადგილება შეუძლია. თუ ეს შესაძლებლობა რეალურ გარემოშიც გამძლე აღმოჩნდება, ის პირველ რიგში **სამუშაო პროცესს** შეცვლის და არა იმას, ვინ არის საბოლოოდ პასუხისმგებელი. ექიმი არ ქრება; სწორედ კვლევების დანიშვნა, შედეგების ინტერპრეტაცია

და მათი მიდევნებაა ადგილი, სადაც ასეთი ინსტრუმენტი პირველად ჩნდება.

და ეს მტკიცების ტვირთსაც სწორ ადგილას აბრუნებს. წარსულ შემთხვევებზე ლიდერბორდის მოგება მიზეზია პერსპექტიული კვლევის ჩასატარებლად და არა მისი შემცვლელი. ავტორებიც სწორედ ამას ამბობენ. MIRA-ს გულწრფელი წაკითხვა შემდეგი კვლევის მოწვევაა — იმ სტანდარტით, რომელიც სფერომ უკვე იცის: პაციენტებზე გაზომილი და არა მხოლოდ ბენჩმარკებზე.

მოკლე შეჯამება

MIRA ავტონომიური AI-აგენტია, რომელიც ელექტრონული სამედიცინო ჩანაწერის იზოლირებულ სიმულატორში მუშაობს: შეუძლია ანამნეზის აღება, ლაბორატორიული, გამოსახულებითი და მიკრობიოლოგიური კვლევების დანიშვნა და ინტერპრეტაცია, დიაგნოზამდე მისვლა და მკურნალობის გეგმის დაწერა. საჯარო MIMIC-IV მონაცემთა ბაზიდან აღებულ 574 წარსულ შემთხვევაზე, რვა წინასწარ შერჩეული დიაგნოზის ფარგლებში, მან დიაგნოსტიკური სიზუსტით ექიმებს აჯობა (88.9% მთლიანობაში; პირდაპირ შედარებაში 87.8% 78.1%-ის წინააღმდეგ) და მეტწილად გაიდლაინებთან შესაბამისი, მედიკამენტურად უსაფრთხო გადაწყვეტილებები მიიღო. მაგრამ შეფასება წარსულ ჩანაწერებზე ტექსტურ სიმულაციაში ჩატარდა; უპირატესობის დიდი ნაწილი მკაფიო კვლევითი შედეგების მქონე დაავადებებზე მოდიოდა; MIRA ექიმებზე ბევრად მეტ სისხლის ანალიზს იყენებდა (ხელმისაწვდომი მაჩვენებლების დაახლოებით 51% 28%-ის წინააღმდეგ); მკურნალობის რამდენიმე შედეგი ისტორიულ ჩანაწერთან შესაბამისობით ფასდებოდა; ხოლო საჯარო მონაცემთა ნაკრები სასწავლო მონაცემებით დაბინძურების რეალურ რისკს ქმნის — რასაც ავტორებიც საკუთარი რიცხვების შესაძლო ზედა ზღვრად განიხილავენ. რეალური წინსვლა ისაა, რომ აგენტი იზოლირებულ კითხვებზე პასუხის ნაცვლად მთელი სამუშაო პროცესის გასწვრივ მოქმედებს. ავტორები პირდაპირ ამბობენ, რომ განზოგადებას, უსაფრთხოებას და მმართველობას ჯერ კიდევ სჭირდება პერსპექტიული, რეალურ სამყაროში ჩატარებული კვლევა. ამიტომ სწორი დასკვნაა: შთამბეჭდავი შესაძლებლობის დემონსტრაცია და არა მტკიცება, რომ AI ექიმზე უკეთ სვამს დიაგნოზს ან უკვე მზადაა კლინიკისთვის.

პირდაპირი შემოწმება

რას აჩვენებს ნაშრომი: ენობრივ მოდელზე აგებულ აგენტ MIRA-ს შეუძლია იზოლირებულ EHR-ში კლინიკური შემთხვევის თავიდან ბოლომდე წარმართვა — ანამნეზი, კვლევები, დიაგნოზი, დანიშნულებები — და რვა დიაგნოზზე აგებულ 574 წარსულ შემთხვევის ბენჩმარკში დიაგნოსტიკური სიზუსტით ექიმებს აჯობა (88.9% მთლიანობაში; პირდაპირ შედარებაში 87.8% board-certified ექიმების 78.1%-ის წინააღმდეგ), მძიმე მედიკამენტური შეცდომების გარეშე.

რა არის შესაძლებელი, მაგრამ დაუმტკიცებელი: რომ ეს შესაძლებლობა რეალურ,

წინასწარ არაშერჩეულ პაციენტებზე სარგებლად გადაიქცევა; ან რომ სიზუსტის უპირატესობა უკეთ reasoning-ს ასახავს და არა მეტ კვლევას, ისტორიულ ჩანაწერთან შესაბამისობას ან საჯარო მონაცემების შესაძლო დამახსოვრებას.

რას არ აჩვენებს: რომ MIRA რვა დიაგნოზის მიღმაც მუშაობს; რომ დასანერგად უსაფრთხოა; რომ თანაბრად ინფორმირებულ ექიმებთან სამართლიან შედარებაში იმარჯვებს; რომ კლინიცისტებს ანაცვლებს; ან რომ შედეგები სასწავლო მონაცემების გადაფარვისგან თავისუფალია.

მთავარი შეზღუდვები: წარსული, სიმულირებული, მხოლოდ ტექსტზე დაფუძნებული შეფასება; რვა წინასწარ შერჩეული დიაგნოზი; ინფორმაციის ასიმეტრია — გაცილებით მეტი სისხლის ანალიზი, დაახლოებით 51% ხელმისაწვდომი მაჩვენებლებისა 28%-ის წინააღმდეგ, თუმცა არა მეტი ძვირადღირებული ვიზუალიზაცია; მკურნალობის შედეგების ნაწილობრივი შეფასება ისტორიული ჩანაწერით; და საჯარო MIMIC-IV ბენჩმარკი, რომელიც ენობრივ მოდელს სწავლებისას შეიძლება ენახა — რასაც ავტორები შედეგების შესაძლო ზედა ზღვრად თავადაც აღნიშნავენ.

რამდენად უნდა ენდოს ამას ზოგადი მკითხველი? მაღალი ნდობაა გამართლებული, რომ MIRA ახალი და რეალური შესაძლებლობის დემონსტრაციაა — აგენტი, რომელიც ჩანაწერში *მოქმედებს* და არა უბრალოდ chatbot. დაბალი ნდობაა გამართლებული მტკიცებისთვის, რომ ის *ექიმზე უკეთესი დიაგნოსტია*: ეს დასკვნა ერთ წარსულ ბენჩმარკს ეხება და მასზე გავლენას ახდენს კვლევების რაოდენობა, ისტორიული ჩანაწერით შეფასება და შესაძლო მონაცემთა დაბინძურება. ავტორების საკუთარი დასკვნა ყველაზე სწორია — პაციენტების მოვლისთვის მნიშვნელობის მისანიჭებლად საჭიროა პერსპექტიული, რეალურ გარემოში ჩატარებული კვლევა.

წყაროები

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