The
CLEAN PAPER

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

AI agent ilishughulikia patient cases nzima yenyewe — katika simulator, kwa records za zamani

MIRA ni aina mpya ya medical AI: badala ya kujibu swali moja, inashughulikia case nzima ndani ya simulated hospital record — kuchukua history, ku-order na kusoma tests, kufikia diagnosis na kuandika orders. Katika retrospective cases 574 kutoka public database, katika diagnoses nane pre-selected, waandishi wanaripoti iliwashinda physicians kwa diagnostic accuracy na kufanya decisions nyingi guideline-concordant na medication-safe. Kila qualifier katika sentence hiyo ni muhimu: ilifanya kazi sandbox kwenye past records, text only; edge nyingi zilitoka kwa conditions zenye clear-cut test results; ili-order karibu mara mbili blood tests kuliko doctors; na outcomes kadhaa zilipimwa dhidi ya kile original chart ilirekodi. Advance halisi ni agent inayotenda across whole workflow — si proof kwamba machine sasa diagnose better than doctor. Waandishi wanasema hivyo: generalization, safety na governance bado zinahitaji prospective real-world studies.

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

Kujibu swali si sawa na kuendesha case nzima

Hadithi nyingi za AI ya kitabibu zinahusu modeli inayojibu. Unaipa vignette au swali la mtihani, inakurudishia utambuzi au paragraph ya ushauri, na inapata alama nzuri. Ni muhimu, lakini narrow: consultant mwenye akili ambaye hagusi chart.

MIRA imejengwa kufanya kitu tofauti, na hapo ndipo habari halisi ilipo. Badala ya kujibu swali moja, inashughulikia kesi nzima: inasoma rekodi ya mgonjwa, inaamua ni historia gani bado inahitajika, inaagiza vipimo vya maabara, picha za kitabibu na microbiolojia, inasoma matokeo, inapunguza orodha ya utambuzi unaowezekana, kisha inaandika maagizo yanayofuata — dawa, kulazwa hospitalini au rufaa ya upasuaji. Inafanya hayo kwa kutenda *ndani* ya rekodi ya afya ya kielektroniki, kama mhudumu wa afya, badala ya kutoa maandishi huru ambayo binadamu atalazimika kuyaingiza mwenyewe.

Hiyo ni hatua halisi: kutoka mfumo inayoshauri kwenda mfumo inayotenda katika workflow. Pia ndiyo aina ya hatua inayobanwa katika coverage kuwa “AI outperforms madaktari.” Kwa hiyo inafaa kuwa precise kuhusu kilichojaribiwa, na wapi.

Kwa sentensi moja: MIRA ilishughulikia kesi nzima yenyewe na, katika kipimo hiki cha kulinganisha, ikapata usahihi mkubwa wa utambuzi kuliko madaktari — lakini ilifanya hivyo katika mazingira ya majaribio yaliyotengwa, kwa rekodi za zamani, katika aina nane za utambuzi zilizochaguliwa mapema, na kwa mawasiliano ya maandishi tu. Sehemu kubwa ya faida yake ilitokana na hali ambazo vipimo vilitoa jibu lililo wazi. Huo ni uwezo mpya wa kweli. Lakini jedwali la alama si kliniki.

MIRA showed a workflow capability, not a clinic verdict

A sandboxed EHR lets an agent act through a whole retrospective case. It is still not a real patient trial.

DEMONSTRATED

- end-to-end case work in a sandboxed EHR
- history, tests, differential diagnosis, orders
- 574 retrospective MIMIC-IV cases
- 8 pre-selected diagnoses
- higher diagnostic accuracy on this benchmark

NOT DEMONSTRATED

- real patients or live hospital deployment
- conditions beyond the eight-target set
- an equally informed head-to-head comparison
- freedom from public-data contamination
- safety and governance for clinical use

A capability shown in a sandbox -- not a diagnostician proven in a clinic.

The figure separates the workflow result from the deployment claim.

MIRA ilionyesha uwezo wa kuendesha mtiririko mzima wa kazi ndani ya EHR ya majaribio iliyotengwa. Haikuonyesha matumizi ya kliniki katika mazingira halisi, uwezo mpana wa utambuzi, wala ulinganisho wa moja

kwa moja na madaktari waliokuwa na taarifa sawa kabisa.

Original Aurora diagram — The Clean Paper CC BY 4.0

Walichofanya waandishi

MIRA — Medical Intelligence for Reasoning and Action — ni mfumo wakala wa AI unaojiendesha uliotengenezwa kwa kutumia modeli za OpenAI. Sehemu inayozungumza na kutenda hutumia **GPT-4o**, huku hatua tofauti ya kupanga ikitumia modeli ya hoja **o1**. Vyote vimefungwa ndani ya mfumo maalum unaofuata viwango, uliojengwa juu ya HL7 FHIR, kiwango cha data kinachotumiwa na hospitali halisi kubadilishana rekodi, pamoja na seti ya zaidi ya shughuli 80,000 za kliniki. Ndani ya mazingira hayo MIRA inaweza kuchukua historia ya mgonjwa, kuagiza na kutafsiri vipimo vya maabara, picha na microbiolojia, kuunda utambuzi tofautishi na kupanga matibabu — ikiwa ni pamoja na kuagiza dawa, kupanga upasuaji na kulazwa hospitalini. Jambo moja muhimu ni kwamba vihakiki otomatiki vilivyopima majibu ya MIRA vilitumia GPT-4o pia, yaani familia ileile ya modeli iliyokuwa ikijaribiwa; baada ya mkaguzi rika kuibua suala hilo, waandishi waliongeza ukaguzi kutoka kwa daktari aliyeidhinishwa na bodi.

Seti ya majaribio ilitokana na **MIMIC-IV**, hifadhidata kubwa ya EHR inayopatikana kwa umma na iliyoondolewa taarifa za utambulisho. Kutoka karibu wagonjwa 300,000 waliotibiwa katika Beth Israel Deaconess Medical Center kati ya 2008 na 2019, waandishi walichagua **kesi 574** zenye **aina nane za utambuzi lengwa** — magonjwa ya tumbo na dharura za tiba ya ndani, ikiwemo appendicitis, pancreatitis, pneumonia na maambukizi ya njia ya mkojo. Katika kila kesi, MIRA ilianza na taarifa chache na ililazimika kuamua hatua inayofuata moja baada ya nyingine, kama katika uchunguzi wa kawaida wa mgonjwa, lakini ikifanya kazi dhidi ya rekodi ambayo matokeo yake ya mwisho yalikuwa tayari yanajulikana.

utendaji yake kisha ililinganishwa na physicians kwenye cases hizo.

> **“Autonomous agent katika sandboxed EHR” inamaanisha nini hasa**
>
> Maneno matatu yanafanya kazi kubwa hapa.
> **Agent** hapa haimaanishi chatbot inayojibu tu. Mfumo huendesha mzunguko wa hatua: huangalia hali ya sasa, huchagua hatua inayofuata — kwa mfano kuagiza kipimo au kuuliza sehemu fulani ya historia — husoma matokeo, kisha huchagua hatua nyingine hadi kufikia utambuzi na mpango wa matibabu. “Intelligence” hutoka kwa modeli ya lugha; *agency* hutoka kwa programu inayozunguka modeli, inayobadilisha maandishi yake kuwa shughuli zinazoruhusiwa katika EHR na kurudisha matokeo kwa modeli.
> **Sandboxed EHR** ni nakala ya mfumo wa rekodi za kitabibu iliyotengwa kwa majaribio, si mfumo wa hospitali unaotumika kwa wagonjwa halisi. MIRA inaweza “kuagiza” kipimo na kupata matokeo ambayo mgonjwa huyo alipata katika maisha halisi kwa sababu kesi ni ya zamani na jibu tayari lipo kwenye rekodi. Hakuna hatua yake inayomgusa mgonjwa halisi au kliniki halisi.
> **Autonomous** inamaanisha inakamilisha case kutoka mwanzo hadi mwisho bila binadamu

> anayehusika moja kwa moja — *ndani ya sandbox*. Haimaanishi iliachiliwa kusimamia wagonjwa bila supervision. Hayo ni madai tofauti sana, na la kwanza tu ndilo lilijaribiwa.
> Kuna kikomo kingine muhimu cha mazingira haya. MIRA inaweza *kuagiza* kipimo chochote, lakini mfumo unaweza kurudisha matokeo tu ikiwa kipimo hicho kilifanywa kwa mgonjwa huyo katika maisha halisi. Ikiomba kipimo cha damu kilichofanyika, hupata thamani halisi; ikiomba picha ambayo haikuwahi kupigwa, hupata "N/A — could not be performed," si namba iliyotungwa. Vivyo hivyo kwa historia: rekodi ikiwa haina taarifa MIRA inayoomba, mgonjwa wa kuigiza hujibu kuwa hajui. Kwa hiyo MIRA haiwezi kuitisha kipimo cha kuamua ambacho hakikufanywa katika uchunguzi wa awali; lazima ifanye kazi na taarifa zilizokuwapo katika rekodi halisi.
> Tofauti hiyo ni muhimu kwa sababu neno la kichwa cha habari — "autonomous" — ndiyo rahisi kusomwa kama dai la pili.

Walichogundua

Kwenye kipimo linganishi, **MIRA iliwashinda physicians kwa usahihi wa utambuzi** — lakini kichwa cha habari inaficha namba tatu tofauti, na ni rahisi kuzichanganya.

Peke yake, katika **kesi 574**, MIRA ilitaja utambuzi sahihi katika **88.9%** ya kesi, ikilinganishwa na utambuzi wa wakati wa kuruhusiwa kutoka hospitalini uliorekodiwa katika MIMIC-IV. Kwa ulinganisho wa moja kwa moja na binadamu, madaktari hawakufanya kesi zote 574 bali seti ileile ya **kesi 311**. Katika seti hiyo MIRA ilipata **87.8%**. Kundi la kwanza la madaktari lilikuwa na madaktari wanne waliothibitishwa na bodi, wenye uzoefu wa miaka 7–11, na lilipata **78.1%**. Kundi la pili lilichanganya viwango vya uzoefu, likiwa na wakazi wengi wachanga pamoja na wataalamu wawili, na lilipata **71.1%**. Kwa hiyo katika kesi zilezile na zana zilezile, MIRA ilikuwa ya kwanza, madaktari wenye uzoefu wa pili, na kundi lenye madaktari wengi wachanga la tatu. Waandishi wenyewe wanasema MIRA ilikuwa "consistently equivalent to, and often exceeded" physicians katika magonjwa yote nane.

Maamuzi yaliyofuata pia yalikuwa mazuri kwa ujumla: mara nyingi yaliendana na miongozo, yalikuwa salama kwa matumizi ya dawa na yalifaa kuhusu kulazwa hospitalini. Waandishi hawakuripoti makosa makubwa ya dawa katika makundi matano ya usalama — mwingiliano wa dawa, dozi kwa wagonjwa wenye matatizo ya figo, mzio, hatari ya QT na uagizaji wa opioid — na maagizo 467 kati ya 468 yalihesabiwa kuwa sahihi. Wakati huo huo wanasisitiza kuwa mfumo "did not achieve 100% reliability." Kwa mtazamo wa kwanza hilo ni jambo la kuvutia: wakala anaendesha kesi nzima na anapata utambuzi sahihi mara nyingi kuliko madaktari katika kipimo hiki.

Lakini namna faida hiyo ilivyopatikana ni muhimu kama alama yenyewe. Wataalamu huru waliokagua makala walibainisha mahali tofauti hiyo ilipotokea. Katika jopo la wataalamu la Science Media Centre, Dk Wei Xing (University of Sheffield) alibainisha kuwa faida kubwa zaidi ilionekana katika hali zenye **vipimo vinavyotoa jibu lililo wazi**, kama appendicitis na pancreatitis, ambako picha au thamani ya maabara inaweza kufunga swali. Kwa pneumonia na maambukizi ya njia ya mkojo—sababu mbili za kawaida sana za kufika idara ya dharura—AI na madaktari wote walifanya vibaya

zaidi, na tofauti kati yao ilikuwa ndogo.

Pia kulikuwa na kutolingana kwa kiasi cha taarifa zilizotumiwa. MIRA ilitegemea maabara zaidi: ilitumia karibu **51%** ya vipengele vya maabara vilivyopatikana katika huduma ya kawaida, dhidi ya karibu **28%** kwa madaktari waliothibitishwa na bodi — wastani wa vipimo saba zaidi vya damu kwa kila kesi. Waandishi wanaeleza kuwa kiwango hicho bado kilikuwa chini ya matumizi ya kawaida yaliyorekodiwa katika hifadhidata na hawakuona ongezeko la kimfumo la upigaji picha ghali wa CT/MRI. Hata hivyo, taarifa nyingi zaidi peke yake zinaweza kuongeza usahihi wa utambuzi. Kwa hiyo huu si ulinganisho safi wa wahusika wenye taarifa sawa, jambo ambalo Xing alibainisha moja kwa moja.

Kwa nini "kuwashinda madaktari" si sawa na "kuwa daktari bora"

Ni rahisi kutafsiri ushindi katika kipimo linganishi kupita kiasi. Kuna hatua tatu kati ya "MIRA ilipata alama ya juu" na "MIRA ni bora."

Kwanza, **ilipimwa dhidi ya nini.** Profesa Julie Jacko (University of Edinburgh) alibainisha kuwa baadhi ya matokeo ya MIRA yalipimwa kwa kulinganisha na kile kilichoandikwa katika rekodi ya awali. Hivyo mfumo unaweza kupata alama nzuri kwa **kurudia tabia ya kliniki iliyorekodiwa**, hata kama tabia hiyo si huduma bora zaidi inayowezekana. Rekodi ya zamani inakuwa ufunguo wa majibu. Hiyo ni njia inayokubalika ya kujenga kipimo linganishi, lakini hupima kufanana na yaliyofanyika, si usahihi kwa maana kamili. Tatizo hili linaathiri zaidi vipimo vya kulinganisha mipango ya matibabu; usahihi mkuu wa utambuzi ulihesabiwa dhidi ya utambuzi wa wakati wa kutoka hospitalini, ambao uko karibu zaidi na matokeo halisi.

Pili, kulikuwa na **kutolingana kwa taarifa**: MIRA ilitumia vipimo vingi zaidi vya damu — karibu 51% ya vipengele vilivyopatikana dhidi ya 28% kwa madaktari — na hivyo ilikuwa na ushahidi mwingi zaidi. Sehemu ya tofauti ya usahihi inaweza kutokana na taarifa hizo za ziada, si uwezo wa hoja pekee.

Tatu, **ilipimwa wapi.** Hizi zilikuwa kesi za zamani katika mazingira ya majaribio yaliyotengwa, zenye aina nane za utambuzi zilizochaguliwa mapema na mawasiliano ya maandishi tu. Tathmini halisi ya kliniki, kama Dk Dominic Olivieri alivyosema, hutegemea si tu kile mgonjwa anachosema bali pia jinsi anavyosema, pamoja na uchunguzi wa mwili, tabia inayoonekana na lugha ya mwili — vitu ambavyo wakala wa maandishi pekee havioni.

Kuna pia wasiwasi mwingine ambao wakaguzi waliibua: **kuchanganyika kwa data ya mafunzo**. MIMIC-IV ni hifadhidata ya umma na imejadiliwa sana, kwa hiyo modeli ya lugha iliyofundishwa kwenye intaneti inaweza kuwa imeona makala, mijadala ya kesi au hata sehemu za data yenyewe. Dk Midhun Parakkal Unni (University of Sheffield) alibainisha kwamba ikiwa baadhi ya majibu yalikuwa tayari katika data ya mafunzo, sehemu ya utendaji inaweza kuwa kumbukumbu badala ya hoja za kliniki; ni kurudia kwa utafiti kwa njia huru pekee kunakoweza kutenganisha hayo. Waandishi hawapuuzi tatizo hili: wanaandika kwamba matokeo yanaweza kutafsiriwa kwa tahadhari kama "possible upper bound" na kwamba yanaweza kuzidisha uwezo wa kujumlisha kwa kesi nyingine za

umma. Kwa maneno mengine, makala yenyewe inaweka kikomo kwenye kichwa chake cha habari.

Hakuna jambo kati ya hayo linalofanya MIRA isiwe ya kuvutia. Linaweka tu tafsiri sahihi katika kiwango kinachofaa: katika kipimo cha kesi za zamani, wakala aliyekuwa na uwezo wa kutenda katika rekodi nzima aliwashinda madaktari hasa kwenye utambuzi wenye vipimo vya wazi, kwa sehemu kwa kuagiza vipimo vingi zaidi na kwa sehemu kwa kulingana na kile kilichoandikwa kwenye rekodi. Huo ni ushahidi wa uwezo muhimu; si ushahidi kwamba tayari ni daktari bora katika kliniki halisi.

Hili halithibitishi nini

- **Halionyeshi MIRA inafanya kazi kwa wagonjwa halisi.** Kila case ilikuwa retrospective na simulated; hakuna mgonjwa aliyewahi kusimamiwa nayo.
- **Halionyeshi inafanya kazi nje ya utambuzi nane.** Conditions nje ya pre-selected set — messy, undifferentiated majority ya medicine — hazikujaribiwa.
- **Halionyeshi iko safe kutumia katika mazingira halisi.** Waandishi wenyewe wanasema generalization, usalama na governance bado zinahitaji prospective katika ulimwengu halisi tafiti.
- **Halijengi fair ulinganisho wa moja kwa moja na madaktari.** Ilitumia blood tests nyingi zaidi (karibu 51% ya available analytes dhidi ya 28%), na matokeo ya matibabu kadhaa zilipimwa partly dhidi ya recorded chart badala ya ground-truth best huduma.
- **Halionyeshi matokeo haina memorization.** Public MIMIC-IV data inaweza overlap na modeli training, jambo ambalo huru replication linahitaji kuondoa.
- **Haimaanishi "AI replaces madaktari."** Ni msaada wa maamuzi inayoweza *kutenda* across record; reviewers wanakubaliana real use itakuwa partnership na clinicians, ambao wanabaki na authority na kutoa yote ambayo text record haioni.

Ushahidi una nguvu kiasi gani?

Gawa dai mbili, kwa sababu ushahidi ni tofauti sana.

Kama onyesho la uwezo, kazi hii ni mpya na inashawishi kiasi: inaonyesha kwamba wakala unaotumia modeli ya lugha unaweza kuendesha kesi nzima ndani ya EHR ya majaribio, akiunganisha historia, vipimo, utambuzi na maagizo kutoka mwanzo hadi mwisho. Hii ndiyo sehemu mpya zaidi ya kazi, na ndiyo inayostahili umakini mkubwa.

Kama dai la kuwa bora kuliko madaktari, ushahidi una mipaka na unahitaji tahadhari. Unatokana na kipimo kimoja cha kesi za zamani, aina nane za utambuzi, kutolingana kwa taarifa kwa sababu MIRA ilitumia vipimo vingi zaidi, ufunguo wa majibu unaotokana kwa sehemu na rekodi ya zamani, na uwezekano wa data ya mafunzo kuwa imechanganyika na MIMIC-IV. Hayo yanatosha kusema kwamba wakala alifanya vizuri sana katika kipimo hiki. Hayatoshi kusema kwamba wakala ni mtaalamu bora wa utambuzi kuliko daktari, na waandishi hawadai hivyo.

Msimamo wenye manufaa si "AI beats madaktari" wala "just a toy." Ni huu: aina mpya ya AI ya

kitabibu, inayotenda katika rekodi nzima badala ya kujibu maswali pekee, ilifanya vizuri katika kipimo kigumu cha kesi za zamani. Sasa inahitaji kujithibitisha kwa wagonjwa halisi wasiochaguliwa mapema, katika tafiti za mbele, chini ya taratibu za usalama na usimamizi.

Kwa nini ni muhimu

Kwa miaka michache, swali la interesting katika AI ya kitabibu limebadilika kimya. Zamani lilikuwa "can a modeli get the utambuzi right?" — na modeli ziliendelea kusema ndiyo kwenye cleaner and cleaner seti za majaribio. MIRA inaonyesha shift kwenda swali gumu zaidi: "can a modeli *do the job* — gather, order, interpret, decide, act — across whole workflow?" Hilo ni swali muhimu zaidi na honest zaidi, kwa sababu acting ndiyo mahali difficulty na hatari zote zinaishi.

Ndiyo maana namna ya kuwasilisha matokeo ni muhimu. Kichwa cha habari *AI outperforms madaktari* kinaelekeza kwenye sehemu isiyo mpya sana na yenye ushahidi wenye mipaka zaidi. Kitu kipya zaidi ni uwezo wa wakala kusafiri katika rekodi nzima na kutenda hatua kwa hatua. Uwezo huo, ukithibitika, unaweza kubadilisha **mtiririko wa kazi** kabla haujabadilisha nani anaongoza huduma. Daktari habadilishwi; kazi za kuagiza vipimo, kutafsiri matokeo na kufuatilia hatua zinazofuata ndizo eneo la kwanza ambalo zana kama hii inaweza kubadilisha.

Na hilo linaweka mzigo wa uthibitisho mahali panapofaa. Kushinda kwenye jedwali la alama la kesi za zamani ni sababu ya kufanya jaribio la mbele, si mbadala wake. Waandishi wanasema hivyo wenyewe. Usomaji mwaminifu wa MIRA ni mwaliko wa utafiti unaofuata, kwa kiwango ambacho tiba tayari inakijua: kupimwa kwa wagonjwa, si kwa vipimo vya maabara vya AI pekee.

Muhtasari safi

MIRA ni mfumo wakala wa AI unaojiendesha ndani ya EHR ya majaribio iliyotengwa. Unaweza kuchukua historia ya mgonjwa, kuagiza na kutafsiri vipimo vya maabara, picha na microbiolojia, kufikia utambuzi na kuandika mpango wa matibabu. Ulijaribiwa kwenye kesi 574 za zamani kutoka MIMIC-IV, katika aina nane za utambuzi zilizochaguliwa mapema, na ulipata usahihi wa utambuzi wa 88.9%; katika ulinganisho wa moja kwa moja wa kesi 311 ulipata 87.8% dhidi ya 78.1% kwa madaktari waliothibitishwa na bodi. Maamuzi mengi pia yaliendana na miongozo na yalikuwa salama kwa dawa. Lakini tathmini ilikuwa ya kuigiza kwa rekodi za zamani na kwa maandishi tu; faida kubwa ilionekana katika hali zenye vipimo vinavyotoa jibu wazi; MIRA ilitumia vipimo vingi zaidi vya damu; baadhi ya matokeo ya matibabu yalipimwa dhidi ya kile kilichorekodiwa kwenye rekodi ya awali; na hifadhidata ya umma inaleta uwezekano wa kuchanganyika na data ya mafunzo. Maendeleo halisi ni uwezo wa wakala kutenda katika mtiririko mzima wa kazi badala ya kujibu maswali yaliyotengwa. Waandishi wanasema wazi kuwa uwezo wa kujumlisha, usalama na usimamizi bado vinahitaji tafiti za mbele katika mazingira halisi. Hivyo huu ni ushahidi wa uwezo wa kuvutia, si uthibitisho kwamba AI hutambua magonjwa vizuri kuliko daktari katika kliniki halisi.

Ukaguzi bila kupamba

Makala inaonyesha nini: Wakala unaotumia modeli ya lugha unaweza kuendesha kesi nzima kutoka mwanzo hadi mwisho ndani ya EHR ya majaribio — historia, vipimo, utambuzi na maagizo — na katika kipimo cha kesi 574 za zamani katika aina nane za utambuzi, MIRA ilipata usahihi mkubwa kuliko madaktari (88.9% kwa jumla; 87.8% dhidi ya 78.1% kwa madaktari waliothibitishwa na bodi), bila makosa makubwa ya dawa yaliyoripotiwa.

Kinachowezekana lakini hakijathibitishwa: Kwamba uwezo huu utaleta faida kwa wagonjwa halisi ambao hawajachaguliwa mapema; kwamba tofauti ya usahihi inatokana na hoja bora badala ya vipimo vingi zaidi, kufanana na rekodi ya zamani au data ya umma ambayo huenda ilionekana wakati wa mafunzo.

Isichoonyeshwa: Kwamba MIRA inafanya kazi nje ya utambuzi nane; kwamba ni safe kutumia katika mazingira halisi; kwamba inawashinda madaktari katika fair equally-informed ulinganisho; kwamba inachukua nafasi ya clinicians; kwamba matokeo hazina training-data contamination.

Vikwazo vikuu: Tathmini ya kesi za zamani, ya kuigiza na ya maandishi tu; aina nane za utambuzi zilizochaguliwa mapema; kutolingana kwa taarifa kwa sababu MIRA ilitumia vipimo vingi zaidi vya damu; baadhi ya matokeo ya matibabu yalipimwa kwa sehemu dhidi ya rekodi ya zamani; na MIMIC-IV ni kipimo cha umma ambacho modeli ya lugha inaweza kuwa imeona wakati wa mafunzo, jambo ambalo waandishi wenyewe wanasema linaweza kufanya alama kuwa kikomo cha juu cha utendaji halisi.

Msomaji wa kawaida anapaswa kuwa na uhakika kiasi gani? Uhakika mkubwa kwamba MIRA ni onyesho halisi na jipya la uwezo: wakala anayefanya kazi katika rekodi nzima, si chatbot tu. Uhakika mdogo kwamba ni mtaalamu bora wa utambuzi kuliko daktari: dai hilo linategemea kipimo kimoja cha kesi za zamani na limechanganywa na tofauti katika uagizaji wa vipimo, namna ya kutoa alama na uwezekano wa kuchanganyika kwa data. Hitimisho la waandishi wenyewe ndilo salama zaidi — tafiti za mbele katika mazingira halisi zinahitajika kabla ya kusema lina maana gani kwa huduma ya wagonjwa.

Vyanzo

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