The Clean Paper

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

AI agent kan ṣiṣẹ́ gbogbo patient case fúnra rẹ̀ — nínú simulator, lórí past records

MIRA jẹ́ irú medical AI tuntun: dípò kó dá ìbéèrè kan lóhùn, ó ṣiṣẹ́ gbogbo case nínú simulated hospital record — take history, order àti ka tests, dé diagnosis, kọ orders. Lórí 574 retrospective cases láti public database, kọjá eight pre-selected diagnoses, àwọn authors report pé ó ju physicians lọ lórí diagnostic accuracy, decisions rẹ̀ sì largely guideline-concordant àti medication-safe. Ṣùgbọ́n gbogbo qualifier ṣe pàtàkì: ó run nínú sandbox lórí past records, text-only; púpọ̀ edge rẹ̀ wá lórí conditions tí test results wọn clear; ó lo blood tests tó pọ̀ ju doctors lọ; òpọ̀ outcomes sì jẹ́ scored sí ohun tí original chart record. Genuine advance ni agent tó act across whole workflow — kì í ṣe proof pé machine diagnose dára ju doctor lọ. Àwọn authors fúnra wọn sọ pé generalization, safety àti governance ṣì nílò 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

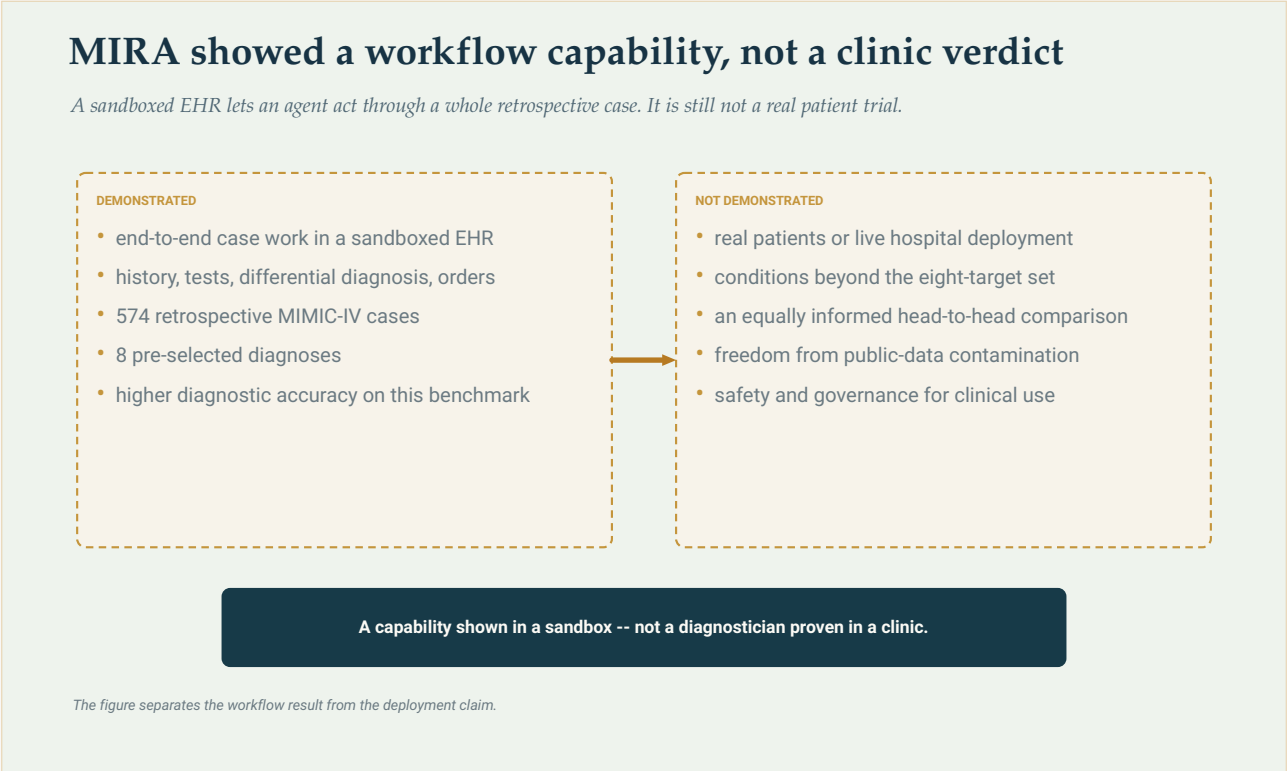
Dídáhùn ìbéèrè kì í ṣe ohun kan náà pẹ̀lú ṣíṣe gbogbo ọ̀ràn

Ọ̀pọ̀ ìtàn nípa ìṣègùn AI máa ń jẹ́ nípa àwòṣe kan tó *dáhùn*. O fún un ní vignette tàbí exam ìbéèrè, ó fún ọ ní ìdánimọ̀ àrùn tàbí paragraph ìmọ̀ràn, ó sì score dáadáa. Ó wúlò, ṣùgbọ́n ó dín kù: bí consultant ọlọ́gbọ́n kan tí kò fi ọwọ́ kan chart rí.

MIRA jẹ́ kí a ṣe nǹkan tó yàtọ̀, ìyàtọ̀ yẹn gan-an sì ni news náà. Dípò kó dáhùn ìbéèrè kan ṣoṣo, ó ṣiṣẹ́ gbogbo ọ̀ràn: ó ka aláìsàn àkọsílẹ̀, ó pinnu history wo ló ṣì nílò, ó order yàrá ìdánwò, imaging àti microbiology ìdánwò, ó ka èsì, ó dín differential ìdánimọ̀ àrùn kù, léyìn náà ó kọ orders tó tẹ̀lé — prescriptions, admission, referral fún surgery. Ó ṣe èyí nípa fífi actions ṣe *nínú* electronic health àkọsílẹ̀, bí clinician ṣe máa ń ṣe, dípò kó kan dá free text sílẹ̀ kí ènìyàn tún transcribe rẹ̀.

Ìyẹn jẹ́ ìgbésẹ̀ gidi: láti ètò tó ń fún ní ìmọ̀ràn sí ètò tó lè act káàkiri workflow. Ó tún jẹ́ irú ìgbésẹ̀ tí coverage máa ń flatten sí “AI outperforms doctors.” Nítorí náà ó tọ́ kí a sọ gangan ohun tí a ìdánwò, àti ibi tí a ìdánwò rẹ̀.

Ẹ̀dà gbolohun kan: MIRA ṣiṣẹ́ gbogbo ọ̀ràn fúnra rẹ̀, ó sì ju physicians lọ lórí diagnostic ìpéye nínú benchmark yìí — ṣùgbọ́n ó ṣe é nínú sandbox, lórí retrospective àkọsílẹ̀, kọjá ìdánimọ̀ àrùn mẹ́jọ tí a yan tẹ́lẹ̀, pẹ̀lú text nìkan, ó sì gba púpọ̀ nínú edge rẹ̀ lórí conditions tí ìdánwò èsì wọn mọ́ kedere jù. Advance náà jẹ́ gidi. Scoreboard náà kì í ṣe clinic.



MIRA fi hàn pé ó lè ṣiṣẹ́ workflow end-to-end nínú sandboxed EHR. Kò fi hàn pé a ti deploy rẹ̀ ní ayé gidi clinic, pé diagnostic coverage rẹ̀ gbooro, tàbí pé ìfiwéra rẹ̀ pẹ̀lú doctors jẹ́ fair head-to-head tí gbogbo ẹgbẹ́ ní information tó dọ́gba.

Ohun tí àwọn olùkọ̀wé ṣe

MIRA — Medical Intelligence fún Reasoning àti Action — jẹ́ autonomous agent tí a kọ lórí OpenAI àwòṣe: apá tó ń converse tí ó sì ń act n ṣiṣẹ́ lórí **GPT-4o**, step planning mìíràn sì lo OpenAI **o1** reasoning àwòṣe. Gbogbo wọn wà nínú custom àgbékalẹ̀ tó tẹ̀lé standards — a kọ ọ́ lórí HL7 FHIR, dátà standard tí àwọn hospitals gidi ń lò láti gbe àkọsílẹ̀ kiri — pẹ̀lú toolbox tó ní ìṣègùn actions tó ju ọ̀kẹ́ méjọ lọ. Nínú sandbox yẹn MIRA lè fa aláìsàn history jáde, order àti interpret labs, imaging àti microbiology, ṣe differential ìdánimọ̀ àrùn, kí ó sì formulate ìtọju plans — prescribe medicines, schedule surgery, plan admissions. Detail kan tó yẹ ká flag láti ìbẹ̀rẹ̀: automated "judges" tó score answers MIRA jẹ́ GPT-4o fúnra wọn — àwòṣe family kan náà tí a ń ìdánwò — léyìn reviewer kan raised concern, àwọn òǹkọ̀wé fi review láti board-certified physician kún un.

ìdánwò set wá láti **MIMIC-IV**, public de-identified EHR database ńlá. Nínú tó fẹ́rẹ̀ẹ́ jẹ́ 300,000 aláìsàn tí Beth Israel Deaconess Medical Center tọju láàárín 2008 sí 2019, àwọn òǹkọ̀wé yan **574 aláìsàn ọ̀ràn** kọjá **eight target ìdánimọ̀ àrùn** — abdominal pathologies àti internal-medicine emergencies, pẹ̀lú appendicitis, pancreatitis, pneumonia àti urinary-tract infection. Fún ọ̀ràn kọọkan, MIRA bẹ̀rẹ̀ pẹ̀lú picture tó lopin, ó sì ní láti pinnu step by step ohun tó yẹ kó ṣe léyìn náà — shape kan náà bí gidi workup, ṣùgbọ́n a ń ṣiṣẹ́ rẹ̀ lórí àkọsílẹ̀ tí gidi ending rẹ̀ ti wà lórí file.

Léyìn náà wọ́n fi performance rẹ̀ wé physicians lórí ọ̀ràn kan náà.

Kí ni "autonomous agent nínú sandboxed EHR" túmọ̀ sí gangan?

Ọ̀rọ̀ méta yìí ń ṣe iṣẹ́ púpọ̀, torí náà ó dára ká tú wọn sílẹ̀.

Agent túmọ̀ sí pé ètò náà kì í ṣe chatbot tó kan ń dá prompt lóhùn. Ó ń run loop kan: wo current ipò, yan action (order ìdánwò yìí, béèrè history yẹn), wo èsì, yan next action — títí tó fi dé ìdánimọ̀ àrùn àti plan. "Intelligence" náà ni èdè àwòṣe; *agency* náà ni scaffolding tó yí text àwòṣe padà sí EHR operations tí a gba láàyè, tó sì feed èsì padà sínú rẹ̀.

Sandboxed EHR túmọ̀ sí simulated, walled-off copy ti medical-record ètò, kì í ṣe live hospital ètò. MIRA lè "order" ìdánwò kan, ó sì gba èsì tí aláìsàn yẹn gba gidi, nítorí ọ̀ràn náà jẹ́ historical, answer sì ti wà nínú àkọsílẹ̀. Kò sí ohun tí MIRA ṣe tó kan gidi aláìsàn tàbí gidi clinic.

Autonomous túmọ̀ sí pé ó parí ọ̀ràn láti ìbẹ̀rẹ̀ dé òpin láìsí ènìyàn in the loop — *nínú sandbox*. Kò túmọ̀ sí pé wọ́n dá a sílẹ̀ kí ó máa manage aláìsàn láìsí supervision. Claims méjì yẹn yàtọ̀ gan-an, àkọ́kọ́ nìkan ni a ìdánwò.

Ohun kan síi nípa sandbox, torí ó rọrùn láti picture rẹ̀ lọ́nà tí kò tọ́: MIRA lè *order* ìdánwò èyíkéyìí tó bá fẹ́, ṣùgbọ́n ètò lè fún un ní èsì nìkan bí ìdánwò yẹn bá **ti ṣẹlẹ̀ gidi** fún aláìsàn yẹn ní gidi life. Bí o bá order blood panel tí aláìsàn gba, o máa gba gidi values; bí o bá order scan tí nobody ordered, ètò máa dá "N/A — kò lè be performed" padà, kò ní invent number. History náà rí bẹ́ẹ̀: bí àkọsílẹ̀ kò bá ní ohun tí MIRA béèrè, simulated aláìsàn kan máa sọ pé kò mọ̀. Nítorí náà MIRA kò lè conjure decisive ìdánwò kan tí a kò ṣe rí — ó lè ṣiṣẹ́ pẹ̀lú ohun tí gidi workup ní nìkan.

Ìyàtọ̀ yìí ṣe pàtàkì torí àkọlé word náà — "autonomous" — ló rọrùn jù láti ka bí claim kejì.

Ohun tí wọ́n rí

Lórí benchmark náà, **MIRA ju physicians lọ lórí diagnostic ìpéye** — ṣùgbọ́n àkọlé náà bo numbers mẹ́ta yàtọ̀, ó sì rọrùn láti mix wọn; torí náà ká yà wọn sọ́tọ̀.

Lórí gbogbo **574 ọ̀ràn**, MIRA fúnra rẹ̀ pè ìdánimọ̀ àrùn tó tọ́ **88.9%** ìgbà — wọ́n score rẹ̀ sí discharge ìdánimọ̀ àrùn tí a àkọsílẹ̀ gidi fún aláìsàn kọọkan nínú MIMIC-IV. Fún head-to-head pẹ̀lú ènìyàn, doctors kò ṣiṣẹ́ gbogbo 574; wọ́n ṣiṣẹ́ shared **311-ọ̀ràn subset**, pẹ̀lú àkọsílẹ̀ àti irinṣẹ́ kan náà tí MIRA ní. Lórí 311 ọ̀ràn wọ̀nyẹn, MIRA score **87.8%**, wọ́n sì fi wé ẹgbẹ́ méjì tí a recruit lọ́tọ̀. Àkọ́kọ́ jẹ́ **senior ẹgbẹ́**: board-certified physicians mẹ́rin tí wọ́n ní experience ọdún 7 sí 11, tí wọ́n score **78.1%**. Èkejì jẹ́ **mixed-seniority ẹgbẹ́**: púpọ̀ jù lọ junior residents pẹ̀lú specialists méjì, tó sún mọ́ bí gidi emergency department ṣe staff, tí wọ́n score **71.1%**. Cases kan náà, irinṣẹ́ kan náà — order sì jẹ́ AI àkọ́kọ́, seasoned doctors kejì, junior-heavy ẹgbẹ́ kẹta. ìwé ìwádìí fúnra rẹ̀ fi careful phrasing sọ pé MIRA jẹ́ “consistently equivalent to, àti often exceeded” physicians kọjá gbogbo diseases méjọ.

Downstream decisions rẹ̀ tún dáa: púpọ̀ wọn tẹ̀lé guidelines, medication-safe, wọ́n sì appropriate lórí admission. Àwọn òǹkọ̀wé report pé kò sí high-severity medication àṣìṣe nínú àábò categories márùn-ún (drug interactions, renal dosing, allergies, QT-risk àti opioid prescribing), prescriptions sì tọ́ ní 467 nínú 468 ọ̀ràn — pẹ̀lú note pé ètò náà “did not achieve 100% reliability.” Bí a bá gba numbers wọ̀nyẹn gẹ́gẹ́ bí wọ́n ṣe wà, èsì náà striking: agent kan ń run gbogbo ọ̀ràn, ó sì wá lókè doctors lórí ìdánimọ̀ àrùn.

Ṣùgbọ́n *shape* ti win náà ṣe pàtàkì gẹ́gẹ́ bí win fúnra rẹ̀. Independent specialists tó ka ìwé ìwádìí tọ́ka sí ibi tí edge náà ti wá. Lórí Science Media Centre expert panel, Dr Wei Xing (University of Sheffield) sọ pé àkọlé figure — AI beating doctors on diagnostic ìpéye — jẹ́ **mostly driven by conditions pẹ̀lú clear ìdánwò èsì**, bí appendicitis àti pancreatitis, níbi tí decisive scan tàbí lab value lè settle ìbéèrè. Fún pneumonia àti urinary-tract infections — méjì nínú commonest reasons tí people fi lọ emergency department — *AI àti doctors méjèèjì* ṣe poorest, gap tó wà láàárín wọn sì kere jù.

Asymmetry kan tún wà nínú bí ẹgbẹ́ méjèèjì ṣe play, ó sì wà nínú numbers ìwé ìwádìí fúnra rẹ̀. MIRA rely lórí lab ju: ó lo tó **51%** ti yàrá ìdánwò analytes tó available nínú routine care, sí roughly **28%** fún board-certified physicians — median blood parameters méje síi per ọ̀ràn. Àwọn òǹkọ̀wé ṣọ́ra láti frame èyí gẹ́gẹ́ bí *below* routine-care baseline dataset náà, kì í ṣe order-everything strategy, wọ́n sì report pé kò sí systematic mú pọ̀ sí i nínú higher-cost cross-sectional imaging. Síbẹ̀, information tó pọ̀ síi fúnra rẹ̀ lè fúnni ní diagnostic ìpéye tó ga — torí náà èyí kì í ṣe like-for-like contest gangan láàárín players tí info wọn dọ́gba, gap tí Xing flag taara. Clinician tí a bá jẹ́ kó order gbogbo cheap blood ìdánwò láì ronú cost, discomfort tàbí delay náà lè look sharper lórí ìwé ìwádìí.

Kí ló dé tí “beat doctors” kì í ṣe “better doctor”?

Ó rọrùn láti over-read benchmark win. Ohun mẹ́ta wà láàárín “MIRA score ga jù” àti “MIRA dára jù.”

Àkọ́kọ́, **ohun tí a fi score rẹ̀ wé.** Professor Julie Jacko (University of Edinburgh) sọ pé ọ̀pọ̀ key èsì MIRA ni a define relative sí ohun tí a document nínú underlying dataset — èyí túmọ̀ sí ètò náà gba reward fún **reproducing recorded ìṣègùn ìhùwàsí**, kì í ṣe dandan fún demonstrating optimal care. Historical chart di answer key. Ìyẹn reasonable fún benchmark, ṣùgbọ́n ohun tó ń measure ni agreement pẹ̀lú ohun tí a ṣe, kì í ṣe correctness ní absolute sense. Ní fairness sí ìwé ìwádìí, issue yìí kan *treatment-alignment* metrics jù — ṣé orders MIRA match chart? —

nígbà tí àkọlé diagnostic ìpéye ni a score sí discharge ìdánimọ̀ àrùn, tó sún mọ́ gidi èsì ju documentation echo lọ.

Èkejì, **information asymmetry** tí a ti mẹ́nuba: far sí i blood ìdánwò (tó 51% ti analytes available, sí 28%) túmọ̀ sí ẹ̀rí tó pọ̀ síi. Ó ṣeé ṣe pé apá kan ti ìpéye gap náà ni a *ra* pẹ̀lú info, kì í ṣe reasoning nìkan.

Ẹ̀kẹta, **ibi tí ó ti run.** Èyí jẹ́ sandbox, lórí retrospective ọ̀ràn, kọjá ìdánimọ̀ àrùn mẹ́jọ tí a yan tẹ́lẹ̀, ní text nìkan. Real ìṣègùn assessment, gẹ́gẹ́ bí Dr Dominic Oliver (University of Oxford) ṣe sọ, kì í dale lórí ohun tí alâìsàn sọ nìkan, ṣùgbọ́n lórí bí wọ́n ṣe sọ ọ́ — pẹ̀lú physical examination, observed ìhùwàsí àti body èdè, gbogbo ohun tí text-only agent tó ń ka finished àkọsílẹ̀ kò rí. alâìsàn tí ìṣòro rẹ̀ kò wọ̀ nínú ìdánimọ̀ àrùn mẹ́jọ yẹn sì wà níta ohun tí ìwádìí yìí lè sọ nípa rẹ̀.

Quiet concern kan tún wà tí reviewers gbe: **dátà contamination.** MIMIC-IV jẹ́ public, wọ́n sì ti kọ́ púpọ̀ nípa rẹ̀, torí náà èdè àwòṣe tí a train lórí open internet lè ti rí ìwé ìwádìí, ọ̀ràn discussions, tàbí dátà náà fúnra rẹ̀. Dr Midhun Parakkal Unni (University of Sheffield) flag èyí taara — bí díẹ̀ nínú answers bá wà nínú training dátà, part of performance jẹ́ ìrántí, kì í ṣe ìṣègùn reasoning, independent replication nìkan ló lè yà méjèèjì. Ó ṣe pàtàkì pé òǹkọ̀wé kò dismiss concern náà: wọ́n kọ pé èsì wọn “lè be cautiously interpreted as a ṣeé ṣe upper bound” àti pé wọ́n “lè overestimate generalization to other public ọ̀ràn” — ìwé ìwádìí kan tó fi ceiling sí àkọlé ara rẹ̀.

Kò sí nínú èyí tó mú MIRA dín interesting. Ó kan mú reading tó tọ́ jẹ́ calibrated: lórí retrospective benchmark, agent tó lè act kọjá gbogbo àkọsílẹ̀ ju physicians lọ lórí ìdánimọ̀ àrùn tí ìdánwò wọn clean — apá kan torí ó order ìdánwò púpọ̀ síi, apá kan torí ó agree pẹ̀lú recorded chart. Èyí jẹ́ gidi capability demonstration, kì í ṣe verdict pé machines ti diagnose dára ju doctors lọ.

Ohun tí èyí kò fi ẹ̀rí múlẹ̀

- Kò fi hàn pé MIRA ṣiṣẹ́ lórí gidi alâìsàn. Gbogbo ọ̀ràn jẹ́ retrospective àti simulated; MIRA kò manage alâìsàn gidi kankan.
- Kò fi hàn pé ó ṣiṣẹ́ kọjá ìdánimọ̀ àrùn mẹ́jọ. Conditions níta pre-selected set — messy, undifferentiated majority of medicine — a kò ìdánwò wọn.
- Kò fi hàn pé ó láìléwu láti deploy. Àwọn òǹkọ̀wé fúnra wọn kọ pé generalization, ààbò àti governance ṣì nílò prospective, ayé gidi ìwádìí.
- Kò establish fair head-to-head pẹ̀lú doctors. Ó lo blood ìdánwò púpọ̀ jù (tó 51% ti analytes available sí 28%), ọ̀pọ̀ ìtọju èsì sì jẹ́ scored partly sí recorded chart dípò ground-truth best care.
- Kò fi hàn pé èsì náà free of memorization. Public MIMIC-IV dátà lè overlap pẹ̀lú training àwòṣe; independent replication ni yóò nílò láti òfin èyí out.
- Kò túmọ̀ sí “AI replaces doctors.” Ó jẹ́ decision ṣe àtìlẹ́yìn fún tó lè *act* kọjá àkọsílẹ̀; consensus reviewers ni pé gidi lò yóò jẹ́ partnership pẹ̀lú clinicians, tí authority yóò wà lọ́wọ́ wọn, wọ́n sì máa supply ohun tí text àkọsílẹ̀ kò lè ní.

Báwo ni ẹ̀rí ṣe lágbára tó?

Yà claim náà sí méjì, torí ẹ̀rí fún apá méjèèjì yàtọ̀ gan-an.

Gẹ́gẹ́ bí capability demonstration — pé language-model agent lè run gbogbo ìṣègùn ọ̀ràn nínú EHR sandbox, chaining history, ìdánwò, ìdánimọ̀ àrùn àti orders end to end — iṣẹ́ náà novel gidi, ó sì reasonably convincing. Èyí ni apá tuntun, èyí sì ni apá tó yẹ ká fiyè sí.

Gẹ́gẹ́ bí superiority claim — pé MIRA *dára ju physicians lọ* — ẹ̀rí náà bounded, ó sì yẹ ká ka a pẹ̀lú caution. Ó hold lórí specific retrospective benchmark, ìdánimọ̀ àrùn méjọ, pẹ̀lú information asymmetry (ìdánwò púpọ̀ síi), answer key tí a fa láti historical chart, àti live possibility of training-data contamination. Ìyẹn tó láti sọ pé "agent performed impressively on this benchmark." Kò tó láti sọ pé "agent is a better diagnostician than a doctor," àwọn òǹkọ̀wé náà kò claim apá kejì.

Stance tó wúlò jù kì í ṣe "AI beats doctors" tàbí "just a toy." Ó jẹ́ pé: irú ìṣègùn AI tuntun kan — tó *act* kọjá àkọsílẹ̀ dípò kó kan dá ìbéèrè lóhùn — ṣe dáadáa lórí difficult retrospective benchmark, báyìí ó sì ní láti fi ẹ̀rí múlẹ̀ ara rẹ̀ níbi tí a kò tíì try rẹ̀: lórí gidi, undifferentiated aláìsàn, prospectively, lábẹ́ governance.

Kí nìdí tí ó fi ṣe pàtàkì?

Fún ọdún díẹ̀, interesting ìbéèrè nínú ìṣègùn AI ti yí quietly. Ó máa ń jẹ́ "lè a àwòṣe get the ìdánimọ̀ àrùn right?" — àwòṣe sì ń dá "yes" lóhùn lórí cleaner àti cleaner ìdánwò sets. MIRA mark shift sí harder ìbéèrè: "lè a àwòṣe *do the job* — gather, order, interpret, decide, act — across a whole workflow?" Ìbéèrè yìí wúlò ju, ó sì honest ju, torí acting ni difficulty àti ewu méjèèjì ti ń gbé.

Ìdí nìyẹn tí framing fi ṣe pàtàkì. àkọlé reflex — *AI outperforms doctors* — ń tọ́ka sí least novel àti least supported part of èsì. Ohun tuntun gidi kere síi ṣùgbọ́n ó consequential ju: agent kan tó lè move nípasẹ̀ entire àkọsílẹ̀. Bí capability yẹn bá hold up, ó máa change **workflow** pẹ́ kí ó tó change ẹni tó wà in charge. Doctor kò disappear; ordering, interpreting, chasing èsì — workflow náà — ni ibi tí irinṣẹ́ bí èyí yóò kọ́kọ́ land.

Ó tún reset burden of proof sí ìtọ́ṣọ́nà tó tọ́. Leaderboard win lórí retrospective ọ̀ràn jẹ́ reason láti run prospective ìdánwò, kì í ṣe substitute fún un. Àwọn òǹkọ̀wé sọ bẹ́ẹ̀. Honest reading MIRA jẹ́ invitation sí next ìwádìí yẹn — pẹ̀lú standard tí field ti mọ̀: measure lórí aláìsàn, kì í ṣe benchmarks nìkan.

Àkótán kedere

MIRA jẹ́ autonomous AI agent tó ń operate sandboxed electronic health àkọsílẹ̀: ó lè take history, order àti interpret labs, imaging àti microbiology, reach ìdánimọ̀ àrùn, kí ó sì write ìtọju plans. Ní ìdánwò lórí 574 retrospective ọ̀ràn láti public MIMIC-IV database, kọjá eight pre-selected ìdánimọ̀ àrùn, ó ju physicians lọ lórí diagnostic ìpéye (88.9% overall; 87.8% sí 78.1% head-to-head), decisions rẹ̀ sì largely guideline-concordant àti medication-safe. Ṣùgbọ́n evaluation náà jẹ́ simulation lórí past àkọsílẹ̀, text-only; púpọ̀ edge rẹ̀ wá lórí conditions tí ìdánwò wọn clear-cut; MIRA lo far sí i blood ìdánwò ju doctors lọ (tó 51% ti analytes available sí 28%); ọ̀pọ̀ ìtọju èsì ni a score sí ohun tí original chart àkọsílẹ̀; public dataset náà sì fa gidi ewu of training-data contamination — tí òǹkọ̀wé fúnra wọn pe ní ṣeé ṣe upper bound lórí numbers wọn. Genuine advance náà ni agent tó act across whole workflow dípò isolated ìbéèrè. Àwọn òǹkọ̀wé sọ kedere pé generalization, ààbò àti governance ṣì nílò prospective, ayé gidi ìwádìí — ìyẹn ni reading tó tọ́: impressive capability demonstration, kì í ṣe proof pé AI diagnose dára ju doctors lọ, kì í

sì ṣe ètò tó ready fún gidi clinic.

Àyẹ̀wò láìsí àṣejù

Ohun tí ìwé ìwádìí fi hàn: Language-model agent kan (MIRA) lè run gbogbo ìṣègùn ọ̀ràn end to end nínú sandboxed EHR — history, ìdánwò, ìdánimọ̀ àrùn, orders — àti, lórí 574-ọ̀ràn retrospective benchmark kọjá ìdánimọ̀ àrùn méjọ, ó ju physicians lọ lórí diagnostic ìpéye (88.9% overall; 87.8% sí 78.1% lòdì sí board-certified physicians) láìsí high-severity medication àṣìṣe.

Ohun tó ṣeé gbà ṣùgbọ́n tí a kò fi ẹ̀rí múlẹ̀: Pé capability yìí máa translate sí àìnfàànì fún gidi, undifferentiated aláìsàn; pé ìpéye edge náà wá láti reasoning tó dára ju dípò ìdánwò púpọ̀ síi, agreement pẹ̀lú recorded chart, tàbí public dátà tí àwòṣe rántí.

Ohun tí kò fi hàn: Pé MIRA ṣiṣẹ́ níta ìdánimọ̀ àrùn méjọ; pé ó láìléwu láti deploy; pé ó beat doctors nínú fair, equally informed ìfiwéra; pé ó replace clinicians; tàbí pé èsì free of training-data contamination.

Main limitations: Retrospective, simulated, text-only evaluation; eight pre-selected ìdánimọ̀ àrùn; information asymmetry (far sí i blood ìdánwò — tó 51% ti available analytes sí 28%, nínú low-cost bloods, kì í ṣe imaging); ìtọju èsì partly scored sí historical chart; àti public benchmark (MIMIC-IV) tí èdè àwòṣe lè ti rí nínú training — tí òǹkọ̀wé flag gẹ́gẹ́ bí ṣeé ṣe upper bound on performance.

Confidence wo ni gbogbogbò reader yẹ kí ó ní? High pé MIRA jẹ́ gidi, novel capability demonstration — agent tó *act* kọjá àkọsílẹ̀, kì í ṣe chatbot nìkan. Low pé ó jẹ́ *better diagnostician than a physician*: claim yẹn bounded sí retrospective benchmark kan, ó sì confounded by test-ordering, scoring against chart, àti ṣeé ṣe contamination. Bottom line òǹkọ̀wé fúnra wọn ni tó tọ́ — a nílò prospective, ayé gidi ìwádìí kí èyí tó túmọ̀ sí ohunkóhun fún aláìsàn care.

Àwọn orísun

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