



The CLEAN PAPER

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

AI agent kan ɕiɕe gbogbo patient case funra re — ninu simulator, lori past records

MIRA je iru medical AI tuntun: dipo ko da ibere kan lohun, o siɕe gbogbo case ninu simulated hospital record — take history, order ati ka tests, de diagnosis, ko orders. Lori 574 retrospective cases lati public database, koda eight pre-selected diagnoses, awon authors report pe o ju physicians lo lori diagnostic accuracy, decisions re si largely guideline-concordant ati medication-safe. Sugbon gbogbo qualifier se pataki: o run ninu sandbox lori past records, text-only; pupo edge re wa lori conditions ti test results won clear; o lo blood tests to po ju doctors lo; opo outcomes si je scored si ohun ti original chart record. Genuine advance ni agent to act across whole workflow — ki i se proof pe machine diagnose dara ju doctor lo. Awon authors funra won so pe generalization, safety ati governance si nilo 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

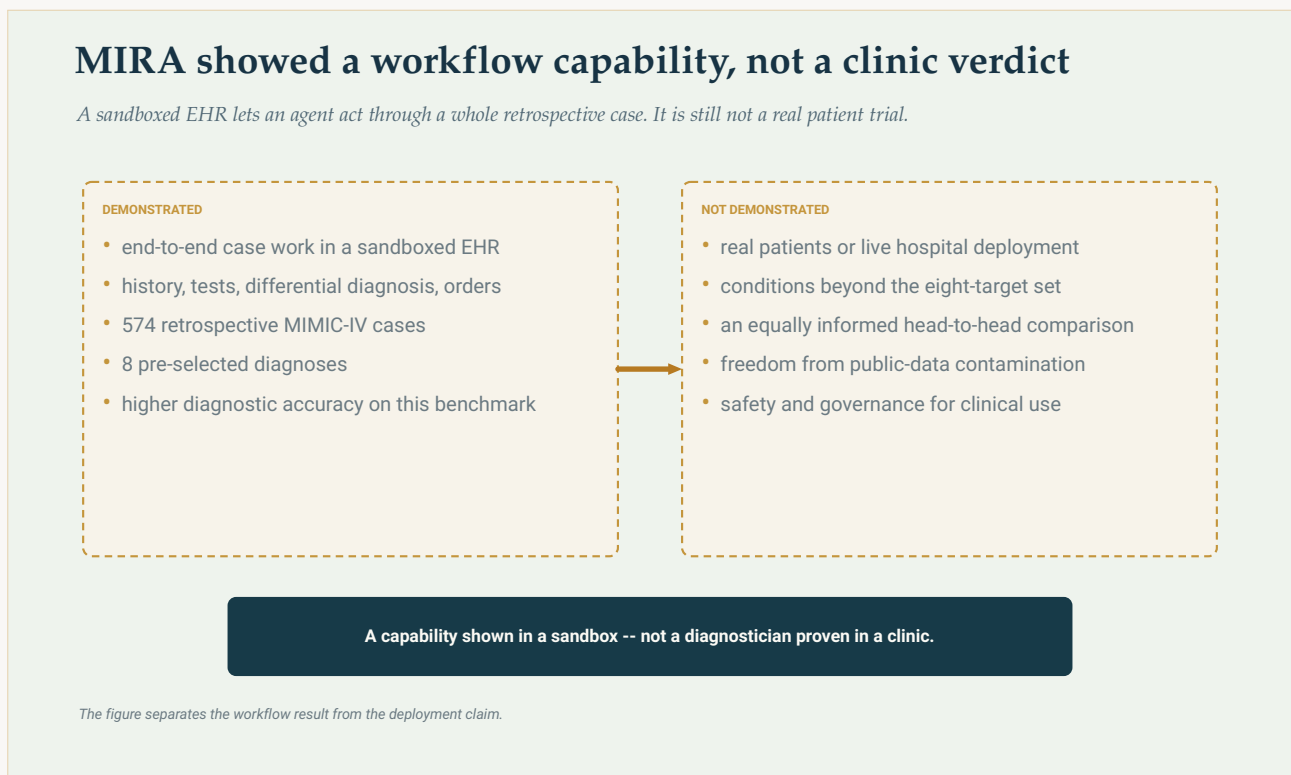
Dídáhùn ibéèrè kì í ẹ̀ ọ̀hun kan náà pẹ̀lú ẹ̀sẹ̀ gbogbo ọ̀ràn

Ọ̀pọ̀ ìtàn nípa ẹ̀sẹ̀gùn AI máa ní jẹ́ nípa àwọ̀se kan tó *dáhùn*. O fún un ní vignette tàbí exam ibéèrè, ó fún ọ ní idánimọ̀ àrùn tàbí paragraph imọ̀ràn, ó sì score dáadáa. Ó wúlò, ẹ̀sẹ̀gùn ó dín kù: bí consultant ọ̀lọ̀gbọ̀n kan tí kò fi ọ̀wọ̀ kan chart rí.

MIRA jẹ́ kí a ẹ̀ nńkan tó yàtò, ìyàtò yẹn gan-an sì ni news náà. Dípò kó dáhùn ibéèrè kan soṣo, ó ẹ̀sẹ̀ gbogbo ọ̀ràn: ó ka aláìsàn àkọ̀sílẹ̀, ó pinnu history wo ló sì nílò, ó order yàrá idánwò, imaging àti microbiology idánwò, ó ka ẹ̀sì, ó dín differential idánimọ̀ àrùn kù, lẹ́yìn náà ó kọ̀ orders tó tẹ̀lẹ̀ — prescriptions, admission, referral fún surgery. Ó ẹ̀ ẹ̀yí nípa fífi actions ẹ̀ *nínú* electronic health àkọ̀sílẹ̀, bí clinician ẹ̀ máa ní ẹ̀, dípò kó kan dá free text sílẹ̀ kí ẹ̀nìyàn tún transcribe rẹ̀.

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

Èdà gbolohun kan: MIRA ẹ̀sẹ̀ gbogbo ọ̀ràn fúnra rẹ̀, ó sì ju physicians lọ lóri diagnostic ipéye nínú benchmark yíí — ẹ̀sẹ̀gùn ó ẹ̀ é nínú sandbox, lóri retrospective àkọ̀sílẹ̀, kọ́já idánimọ̀ àrùn méjọ́ tí a yan tẹ̀lẹ̀, pẹ̀lú text nńkan, ó sì gba púpọ̀ nínú edge rẹ̀ lóri conditions tí idánwò ẹ̀sì wọn mó kedere jù. Advance náà jẹ́ gidi. Scoreboard náà kì í ẹ̀ clinic.



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

Ohun tí àwọn olùkòwé ẹ

MIRA — Medical Intelligence fún Reasoning àti Action — jẹ autonomous agent tí a kọ lórí OpenAI àwòṣe: apá tó ní converse tí ó sì ní act n ẹ́ṣẹ́ lórí **GPT-4o**, step planning mǐràn sì lo OpenAI **o1** reasoning àwòṣe. Gbogbo wọn wà nínú custom àgbékalẹ́ tó tẹ́lẹ́ standards — a kọ ọ́ lórí HL7 FHIR, dátà standard tí àwọn hospitals gidi ní ló láti gbe àkọ̀sílẹ́ kírí — pẹ̀lú toolbox tó ní ẹ̀ṣẹ̀gùn actions tó ju ọ̀kẹ́ méjọ lọ. Nínú sandbox yẹn MIRA lè fa aláìsàn history jáde, order àti interpret labs, imaging àti microbiology, ẹ́ differential ìdánimọ̀ àrùn, kí ó sì formulate ìtoju plans — prescribe medicines, schedule surgery, plan admissions. Detail kan tó yẹ ká flag láti ìbèrẹ̀: automated “judges” tó score answers MIRA jẹ́ GPT-4o fúnra wọn — àwòṣe family kan náà tí a ní ìdánwò — lẹ́yìn reviewer kan raised concern, àwọn ò̀nkòwé fí review láti board-certified physician kún un.

ìdánwò set wá láti **MIMIC-IV**, public de-identified EHR database nílá. Nínú tó fẹ̀rẹ́ jẹ́ 300,000 aláìsàn tí Beth Israel Deaconess Medical Center toju láàárín 2008 sí 2019, àwọn ò̀nkòwé yàn **574 aláìsàn ọ̀ràn** kojá **eight target ìdánimọ̀ àrùn** — abdominal pathologies àti internal-medicine emergencies, pẹ̀lú appendicitis, pancreatitis, pneumonia àti urinary-tract infection. Fún ọ̀ràn kọ̀kan, MIRA bèrẹ́ pẹ̀lú picture tó lopin, ó sì ní láti pinnu step by step ohun tó yẹ kó ẹ́ lẹ́yìn náà — shape kan náà bí gidi workup, ẹ̀ṣẹ̀bọ̀n a ní ẹ́ṣẹ́ rẹ́ lórí àkọ̀sílẹ́ tí gidi ending rẹ́ tí wá lórí file.

Lẹ́yìn náà wón fí performance rẹ́ wé physicians lórí ọ̀ràn kan náà.

Kí ní “autonomous agent nínú sandboxed EHR” túmọ̀ sí gangan?

Ọ̀rọ̀ méta yíí ní ẹ́ ẹ́ṣẹ́ púpọ̀, torí náà ó dára ká tú wọn sílẹ̀.

Agent túmọ̀ sí pé ètò náà kí í ẹ́ chatbot tó kan ní dá prompt lóhùn. Ó ní run loop kan: wo current ipò, yan action (order ìdánwò yíí, bèèrẹ́ history yẹn), wo èsì, yan next action — tí tí tó fí 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 rẹ́.

Sandboxed EHR túmọ̀ sí simulated, walled-off copy tí medical-record ètò, kí í ẹ́ live hospital ètò. MIRA lè “order” ìdánwò kan, ó sì gba èsì tí aláìsàn yẹn gba gidi, nítorí ọ̀ràn náà jẹ́ historical, answer sì tí wá nínú àkọ̀sílẹ́. Kò sí ohun tí MIRA ẹ́ tó kan gidi aláìsàn tàbí gidi clinic.

Autonomous túmọ̀ sí pé ó parí ọ̀ràn láti ìbèrẹ́ dé ọ̀pin láìsí ènìyàn in the loop — *nínú sandbox*. Kò túmọ̀ sí pé wón dá a sílẹ́ kí ó máa manage aláìsàn láìsí supervision. Claims méjì yẹn yàtò gan-an, àkọ̀kọ̀ nikan ní a ìdánwò.

Ohun kan síí nípa sandbox, torí ó ọ̀rùn láti picture rẹ́ lónà tí kò tọ́: MIRA lè *order* ìdánwò èyíkéyíí tó bá fẹ́, ẹ̀ṣẹ̀bọ̀n ètò lè fún un ní èsì nikan bí ìdánwò yẹn bá **tí ẹ́lẹ́ 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 bẹ́: bí àkọ̀sílẹ́ kò bá ní ohun tí MIRA bèèrẹ́, simulated aláìsàn kan máa sọ pé kò mò. Nítorí náà MIRA kò lè conjure decisive ìdánwò kan tí a kò ẹ́ rí — ó lè ẹ́ṣẹ́ pẹ̀lú ohun tí gidi workup ní nikan.

Ìyàtò yíí ẹ́ pàtàkì torí àkọ̀lé word náà — “autonomous” — ló ọ̀rùn jù láti ka bí claim kejì.

Ohun tí wọn rí

Lórí benchmark nàà, **MIRA ju physicians lọ lórí diagnostic ìpéye** — ùgbón àkọlẹ nàà bo numbers mèta yàtò, ó sì rọrùn láti mix wọn; torí nàà ká yà wọn sọtò.

Lórí gbogbo **574 ọ̀ràn**, MIRA fúnra rẹ̀ pẹ̀ ìdánimò àrùn tó tó **88.9%** ìgbà — wọn score rẹ̀ sí discharge ìdánimò àrùn tí a àkọsílẹ̀ gidi fún aláìsàn kọ̀ọkan nínú MIMIC-IV. Fún head-to-head pẹ̀lú ènìyàn, doctors kò sìşẹ̀ gbogbo 574; wọn sìşẹ̀ shared **311-ọ̀ràn subset**, pẹ̀lú àkọsílẹ̀ àti irinşẹ̀ kan nàà tí MIRA ní. Lórí 311 ọ̀ràn wònyẹn, MIRA score **87.8%**, wọn sì fi wé egbé méjì tí a recruit lóttò. Àkọkọ̀ jẹ́ **senior egbé**: board-certified physicians méré̀n tí wọn ní experience ọ̀dún 7 sí 11, tí wọn score **78.1%**. Èkejì jẹ́ **mixed-seniority egbé**: púpò jù lọ junior residents pẹ̀lú specialists méjì, tó sún mọ́ bí gidi emergency department ẹ̀ staff, tí wọn score **71.1%**. Cases kan nàà, irinşẹ̀ kan nàà — order sì jẹ́ AI àkọkọ̀, seasoned doctors kejì, junior-heavy egbé kẹta. ìwé ìwádíí fúnra rẹ̀ fi careful phrasing sọ pé MIRA jẹ́ “consistently equivalent to, àti often exceeded” physicians kọ́já gbogbo diseases méjọ.

Downstream decisions rẹ̀ tún dára: púpò wọn tẹ̀lẹ̀ guidelines, medication-safe, wọn sì appropriate lórí admission. Àwọn ònkòwé report pé kò sí high-severity medication àşìşẹ̀ nínú ààbò categories márùn-ún (drug interactions, renal dosing, allergies, QT-risk àti opioid prescribing), prescriptions sì tó ní 467 nínú 468 ọ̀ràn — pẹ̀lú note pé ètò nàà “did not achieve 100% reliability.” Bí a bá gba numbers wònyẹn gégé bí wọn ẹ̀ wà, èsì nàà striking: agent kan ní run gbogbo ọ̀ràn, ó sì wá lókè doctors lórí ìdánimò àrùn.

Şùgbón *shape* ti win nàà ẹ̀ pàtàkì gégé bí win fúnra rẹ̀. Independent specialists tó ka ìwé ìwádíí tóka sí ibi tí edge nàà ti wá. Lórí Science Media Centre expert panel, Dr Wei Xing (University of Sheffield) sọ pé àkọlẹ̀ 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 ìbèèrè. Fún pneumonia àti urinary-tract infections — méjì nínú commonest reasons tí people fi lọ emergency department — *AI àti doctors méjèjèjì* ẹ̀ poorest, gap tó wà láàárín wọn sì kere jù.

Asymmetry kan tún wà nínú bí egbé méjèjèjì ẹ̀ play, ó sì wà nínú numbers ìwé ìwádíí 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 ònkòwé şóra láti frame èyí gégé bí *below* routine-care baseline dataset nàà, kì í ẹ̀ order-everything strategy, wọn sì report pé kò sí systematic mú pò sí i nínú higher-cost cross-sectional imaging. Síbẹ̀, information tó pò síi fúnra rẹ̀ lè fúnni ní diagnostic ìpéye tó ga — torí nàà èyí kì í ẹ̀ 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 nàà lè look sharper lórí ìwé ìwádíí.

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

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

Àkọkọ̀, **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 tùmò sí ètò nàà gba reward fún **reproducing recorded ìşègùn ihùwàsì**, kì í ẹ̀ dandan fún demonstrating optimal care. Historical chart di answer key. Ìyẹn reasonable fún benchmark, ùgbón ohun tó ní measure ni agreement pẹ̀lú ohun tí a ẹ̀, kì í ẹ̀ correctness ní absolute sense. Ní fairness sí ìwé ìwádíí, issue yíi kan *treatment-alignment* metrics jù — ẹ̀ orders MIRA match chart? —

nígba tí àkọlẹ diagnostic ipéye ni a score sí discharge idánimò àrùn, tó sún mó gidi èsì ju documentation echo lọ.

Èkejì, **information asymmetry** tí a ti mẹnuba: far sí i blood idánwò (tó 51% ti analytes available, sí 28%) túmò sí ẹrì tó pò síi. Ó ẹẹ ẹ pé apá kan ti ipéye gap nàà ni a *ra* pèlú info, kì í ẹ reasoning nìkan.

Èkẹta, **ibi tí ó ti run**. Èyí jẹ sandbox, lórí retrospective ọràn, kojá idánimò àrùn méjọ tí a yan tẹlẹ, ní text nìkan. Real isègùn assessment, gégé bí Dr Dominic Oliver (University of Oxford) ẹ sọ, kì í dale lórí ohun tí aláìsàn sọ nìkan, sùgbón lórí bí wọn ẹ sọ ọ — pèlú physical examination, observed ihùwàsí àti body èdè, gbogbo ohun tí text-only agent tó ní ka finished àkọsílẹ kò rí. aláìsàn tí isẹro rẹ kò wò nínú idánimò àrùn méjọ yẹn sì wà níta ohun tí iwádíi yíi lè sọ nípa rẹ.

Quiet concern kan tún wà tí reviewers gbe: **dátà contamination**. MIMIC-IV jẹ public, wọn sì ti kọ púpò nípa rẹ, torí nàà èdè àwòṣe tí a train lórí open internet lè ti rí iwé iwádíi, ọràn discussions, tàbí dátà nàà fúnra rẹ. Dr Midhun Parakkal Unni (University of Sheffield) flag èyí taara — bí diẹ nínú answers bá wà nínú training dátà, part of performance jẹ irántí, kì í ẹ isègùn reasoning, independent replication nìkan ló lè yà méjèèjì. Ó ẹ pàtàkì pé ònkòwé kò dismiss concern nàà: wọn kọ pé èsì wọn “lè be cautiously interpreted as a ẹẹ ẹ upper bound” àti pé wọn “lè overestimate generalization to other public ọràn” — iwé iwádíi kan tó fi ceiling sí àkọlẹ ara rẹ.

Kò sí nínú èyí tó mú MIRA dín interesting. Ó kan mú reading tó tọ jẹ calibrated: lórí retrospective benchmark, agent tó lè act kojá gbogbo àkọsílẹ ju physicians lọ lórí idánimò àrùn tí idánwò wọn clean — apá kan torí ó order idánwò púpò síi, apá kan torí ó agree pèlú recorded chart. Èyí jẹ gidi capability demonstration, kì í ẹ verdict pé machines ti diagnose dára ju doctors lọ.

Ohun tí èyí kò fi ẹrì múlẹ

- Kò fi hàn pé MIRA ẹṣẹ lórí gidi aláìsàn. Gbogbo ọràn jẹ retrospective àti simulated; MIRA kò manage aláìsàn gidi kankan.
- Kò fi hàn pé ó ẹṣẹ kojá idánimò àrùn méjọ. Conditions níta pre-selected set — messy, undifferentiated majority of medicine — a kò idánwò wọn.
- Kò fi hàn pé ó láilẹwu láti deploy. Àwọn ònkòwé fúnra wọn kọ pé generalization, ààbò àti governance sì nílò prospective, ayé gidi iwádíi.
- Kò establish fair head-to-head pèlú doctors. Ó lo blood idánwò púpò jù (tó 51% ti analytes available sí 28%), ọpọ itọju èsì sì jẹ scored partly sí recorded chart dípò ground-truth best care.
- Kò fi hàn pé èsì nàà free of memorization. Public MIMIC-IV dátà lè overlap pèlú training àwòṣe; independent replication ni yóò nílò láti òfin èyí out.
- Kò túmò sí “AI replaces doctors.” Ó jẹ decision ẹ àtílẹyìn fún tó lè *act* kojá àkọsílẹ; consensus reviewers ni pé gidi lò yóò jẹ partnership pèlú clinicians, tí authority yóò wà lẹwọ wọn, wọn sì máa supply ohun tí text àkọsílẹ kò lè ní.

Báwo ni ẹrì ẹ se lágbára tó?

Yà claim nàà sí méjì, torí ẹrì fún apá méjèèjì yàtò gan-an.

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

Gégé bí superiority claim — pé MIRA *dára ju physicians lọ* — èrí nàà bounded, ó sì yẹ ká ka a pèlú caution. Ó hold lórí specific retrospective benchmark, ìdánimò àrùn méjọ, pèlú information asymmetry (ìdánwò púpò sí), 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 ònkòwé nàà kò claim apá kejì.

Stance tó wúlò jù kì í ṣe “AI beats doctors” tàbí “just a toy.” Ó jẹ pé: irú ìṣẹ̀gùn AI tuntun kan — tó *act* kojá àkọsílẹ̀ dípò kó kan dá ibèèrè lóhùn — ṣe dáadàa lórí difficult retrospective benchmark, báyii ó sì ní láti fi èrí múlẹ̀ ara rẹ níbi tí a kò tì try rẹ: lórí gidi, undifferentiated aláìsàn, prospectively, lábẹ̀ governance.

Kí nìdí tí ó fi ṣe pàtàkì?

Fún ọ̀dún diẹ, interesting ibèèrè nínú ìṣẹ̀gùn AI ti yí quietly. Ó máa n jẹ “lè a àwòṣe get the ìdánimò àrùn right?” — àwòṣe sì n dá “yes” lóhùn lórí cleaner àti cleaner ìdánwò sets. MIRA mark shift sí harder ibèèrè: “lè a àwòṣe *do the job* — gather, order, interpret, decide, act — across a whole workflow?” Ibèèrè yí wúlò ju, ó sì honest ju, torí acting ni difficulty àti ewu méjèèjì ti n gbé.

Ìdí nìyẹn tí framing fi ṣe pàtàkì. àkólé reflex — *AI outperforms doctors* — n tóka sí least novel àti least supported part of èsì. Ohun tuntun gidi kere sí ùgbọ̀n ó consequential ju: agent kan tó lè move nípasẹ̀ entire àkọsílẹ̀. Bí capability yẹn bá hold up, ó máa change **workflow** pé kí ó tó change ẹ̀ni tó wà in charge. Doctor kò disappear; ordering, interpreting, chasing èsì — workflow nàà — ni ibi tí irinṣẹ̀ bí èyí yóò kọ́kọ́ land.

Ó tún reset burden of proof sí ìtọ́sọ̀nà tó tọ. Leaderboard win lórí retrospective ọ̀ràn jẹ reason láti run prospective ìdánwò, kì í ṣe substitute fún un. Àwọn ònkòwé sọ bèẹ̀. Honest reading MIRA jẹ invitation sí next iwádíí yẹn — pèlú standard tí field ti mò: measure lórí aláìsàn, kì í ṣe benchmarks nìkan.

Àkótán kedere

MIRA jẹ autonomous AI agent tó n operate sandboxed electronic health àkọsílẹ̀: ó lè take history, order àti interpret labs, imaging àti microbiology, reach ìdánimò àrùn, kí ó sì write ìtoju plans. Ní ìdánwò lórí 574 retrospective ọ̀ràn láti public MIMIC-IV database, kojá eight pre-selected ìdánimò àrùn, ó ju physicians lọ lórí diagnostic ipéye (88.9% overall; 87.8% sí 78.1% head-to-head), decisions rẹ̀ sì largely guideline-concordant àti medication-safe. Ùgbọ̀n evaluation nàà jẹ simulation lórí past àkọsílẹ̀, text-only; púpò 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 ìtoju èsì ni a score sí ohun tí original chart àkọsílẹ̀; public dataset nàà sì fa gidi ewu of training-data contamination — tí ònkòwé fúnra wọn pe ní ṣe ẹ̀ ṣe upper bound lórí numbers wọn. Genuine advance nàà ni agent tó act across whole workflow dípò isolated ibèèrè. Àwọn ònkòwé sọ kedere pé generalization, ààbò àti governance sì nílò prospective, ayé gidi iwádíí — ìyẹn ni reading tó tọ: impressive capability demonstration, kì í ṣe proof pé AI diagnose *dára ju* doctors lọ, kì í

sì ẹ̀ ẹ̀tò tó ready fún gidi clinic.

Àyèwò láísí àṣejù

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

Ohun tó ẹ́ẹ̀ gbà sùgbọ̀n tí a kò fi ẹ̀rí múlẹ̀: Pé capability yíí máa translate sí ànfààní fún gidi, undifferentiated aláìsàn; pé ìpéye edge nàà wá láti reasoning tó dára ju dípò ìdánwò púpò síí, agreement pèlú recorded chart, tàbí public dátà tí àwòṣe rántí.

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

Main limitations: Retrospective, simulated, text-only evaluation; eight pre-selected ìdánimò àrùn; information asymmetry (far sí i blood ìdánwò — tó 51% ti available analytes sí 28%, nínú low-cost bloods, kì í ẹ̀ ẹ̀ imaging); ìtoju ẹ̀sì partly scored sí historical chart; àti public benchmark (MIMIC-IV) tí ẹ̀dè àwòṣe lè ti rí nínú training — tí ònkòwé flag gégé bí ẹ́ẹ̀ ẹ̀ upper bound on performance.

Confidence wo ni gbogbogbò reader yẹ kí ó ní? High pé MIRA jẹ gidi, novel capability demonstration — agent tó *act* kojá àkọ̀sílẹ̀, kì í ẹ̀ chatbot nikan. Low pé ó jẹ *better diagnostician than a physician*: claim yẹn bounded sí retrospective benchmark kan, ó sì confounded by test-ordering, scoring against chart, àti ẹ́ẹ̀ ẹ̀ contamination. Bottom line ònkòwé fúnra wọn ni tó tó — a nílò prospective, ayé gidi iwádíí kí èyí tó tùmò 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>