The Clean Paper

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

Bir yapay zekâ ajanı hasta vakalarını baştan sona kendi başına yürüttü — ama simülatörde, geçmiş kayıtlar üzerinde

MIRA yeni bir tıbbi yapay zekâ türü: tek bir soruyu yanıtlamak yerine simüle edilmiş bir hastane kaydında tüm vakayı yürütüyor — öykü alıyor, test isteyip sonuçlarını okuyor, tanıya ulaşıyor ve istemleri yazıyor. Kamuya açık bir veritabanından alınan, önceden seçilmiş sekiz tanıyı kapsayan 574 geriye dönük vakada yazarlar, tanı doğruluğunda hekimlerden daha iyi sonuç verdiğini ve kararlarının büyük ölçüde kılavuzlarla uyumlu ve ilaç açısından güvenli olduğunu bildiriyor. Bu cümlenin her nitelemesi önemli: sistem geçmiş kayıtlarda, yalnızca metinle çalışan bir sandbox'ta test edildi; avantajının büyük kısmı net test sonuçları olan hastalıklardan geldi; doktorların yaklaşık iki katı kadar kan testi kullandı; bazı çıktılar da orijinal dosyada kaydedilenlere göre puanlandı. Gerçek ilerleme, bütün iş akışı boyunca eylem alan bir ajan — makinenin artık doktordan daha iyi tanı koyduğunun kanıtı değil. Yazarlar da bunu söylüyor: genellenebilirlik, güvenlik ve yönetişim için hâlâ ileriye dönük gerçek dünya çalışmalarına ihtiyaç var.

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

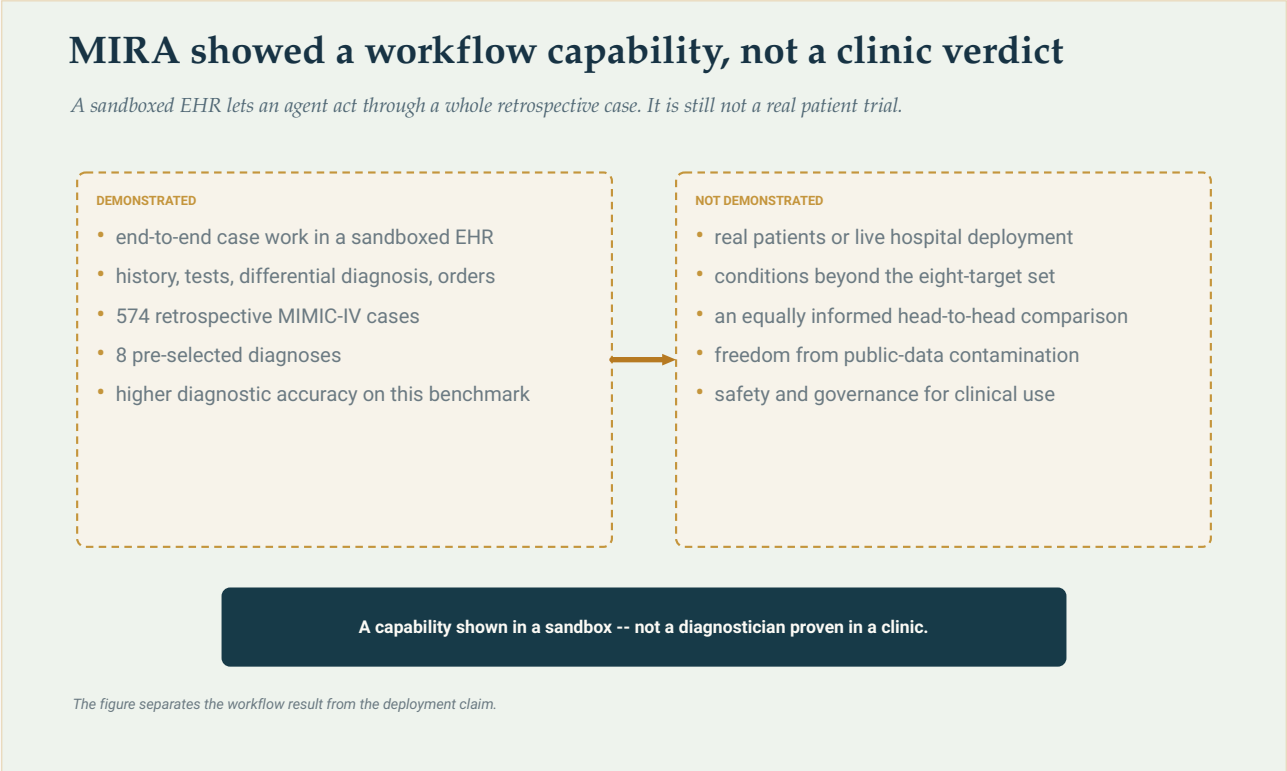
Bir soruyu yanıtlamak, vakayı yönetmekle aynı şey değildir

Tıbbi yapay zekâ haberlerinin çoğu, *yanıt veren* bir model hakkındadır. Modele bir vaka özeti ya da sınav sorusu verirsiniz; o da size bir tanı veya bir paragraf öneri verir ve yüksek puan alır. Yararlı, ama dar kapsamlıdır: dosyaya hiç dokunmayan zeki bir danışman gibi.

MIRA farklı bir şey yapmak üzere tasarlandı ve asıl haber de bu fark. Tek bir soruyu yanıtlamak yerine bütün bir vakayı yürütüyor: hastanın kaydını okuyor, hangi öykü bilgilerine hâlâ ihtiyaç duyduğuna karar veriyor, laboratuvar, görüntüleme ve mikrobiyoloji testleri istiyor, sonuçları okuyor, ayırıcı tanıyı daraltıyor ve ardından gerekli istemleri yazıyor — reçete, yatış, cerrahi sevk. Bunu, bir insanın daha sonra kopyalayacağı serbest metin üretmek yerine, bir klinisyenin yaptığı gibi elektronik sağlık kaydı *içinde* işlemler gerçekleştirerek yapıyor.

Bu gerçek bir adım: tavsiye veren bir sistemden, iş akışının tamamında eylem alan bir sisteme geçiş. Aynı zamanda haberlerde kolayca "yapay zekâ doktorları geçti" cümlesine indirgenen türden bir adım. Bu yüzden neyin, nerede test edildiğini dikkatle ayırmak gerekiyor.

Tek cümlelik özet: MIRA vakaları baştan sona kendi başına yürüttü ve bu benchmark'ta tanı doğruluğunda hekimlerden daha iyi sonuç verdi — ama bunu bir sandbox içinde, geriye dönük kayıtlarda, önceden seçilmiş sekiz tanı üzerinde, yalnızca metinle iletişim kurarak yaptı; avantajının büyük kısmı da test sonuçları en net olan hastalıklardan geldi. İlerleme gerçek. Skor tablosu ise klinik değil.



MIRA, sandbox içindeki bir EHR'de uçtan uca iş akışı yürütebildiğini gösterdi. Gerçek dünyada klinik kullanıma hazır olduğunu, geniş tanı kapsamını veya doktorlarla eşit bilgiye dayanan adil bir başa baş karşılaştırmayı göstermedi.

Yazarlar ne yaptı?

MIRA — Medical Intelligence for Reasoning and Action — OpenAI modelleri üzerine kurulmuş özerk bir ajandır: konuşan ve eylem alan bölüm **GPT-4o** ile çalışır; ayrı bir planlama adımı ise OpenAI'ın **o1** akıl yürütme modelini kullanır. Bunlar, gerçek hastanelerin kayıt aktarımında kullandığı veri standardı HL7 FHIR üzerine kurulmuş, standartlarla uyumlu özel bir çerçevenin içinde ve seksen binden fazla olası klinik eylemden oluşan bir araç setiyle birlikte çalışır. Bu sandbox içinde MIRA hastanın öyküsünü çekebilir, laboratuvar, görüntüleme ve mikrobiyoloji testleri isteyip yorumlayabilir, ayırıcı tanı oluşturabilir ve tedavi planı hazırlayabilir — ilaç yazabilir, cerrahi işlemler planlayabilir, yatış düzenleyebilir. Baştan belirtilmesi gereken bir ayrıntı da şu: MIRA'nın yanıtlarını puanlayan otomatik "hakemler" de GPT-4o idi — yani test edilen model ailesinin aynısı. Bir hakem bu endişeyi dile getirdikten sonra yazarlar puanlamayı uzmanlık belgesine sahip bir hekimin incelemesiyle destekledi.

Test kümesi, kamuya açık ve kimlik bilgileri çıkarılmış büyük bir EHR veritabanı olan **MIMIC-IV**'ten geldi. Beth Israel Deaconess Medical Center'da 2008–2019 arasında tedavi edilen yaklaşık 300.000 hasta arasından yazarlar, apandisit, pankreatit, zatürre ve idrar yolu enfeksiyonu dahil olmak üzere karın patolojileri ve iç hastalıkları acillerini kapsayan **sekiz hedef tanıda 574 hasta vakası** seçti. Her vakada MIRA sınırlı bir başlangıç bilgisiyle yola çıktı ve gerçek bir klinik değerlendirmeye benzer biçimde adım adım ne yapacağına karar verdi — fakat bunu, gerçek sonucun zaten kayıtlarda bulunduğu tarihsel bir vaka üzerinde yaptı.

Ardından performansı aynı vakalarda hekimlerle karşılaştırıldı.

> **"Sandbox EHR içindeki özerk ajan" aslında ne demek?**
>
> Burada üç kelime çok fazla yük taşıyor; bu yüzden açmakta fayda var.
> **Ajan**, sistemin bir isteme yanıt veren sohbet botu olmadığı anlamına gelir. Bir döngü yürütür: mevcut duruma bakar, bir eylem seçer (şu testi iste, şu öykü bilgisini sor), sonucu görür, bir sonraki eylemi seçer — tanıya ve plana ulaşana kadar. "Zekâ" kısmı dil modelidir; *ajanlık* ise modelin metnini izin verilen EHR işlemlerine çeviren ve sonuçları yeniden modele besleyen etrafındaki yazılım altyapısıdır.
> **Sandbox EHR**, canlı bir hastane sistemi değil, duvarlarla çevrilmiş simüle edilmiş bir tıbbi kayıt kopyası demektir. MIRA bir testi "isteyebilir" ve hastanın gerçekten aldığı sonucu görebilir, çünkü vaka tarihseldir ve sonuç zaten kayıtta vardır. Yaptığı hiçbir işlem gerçek bir hastaya veya kliniğe dokunmaz.
> **Özerk**, vakayı insan döngüde olmadan baştan sona tamamladığı anlamına gelir — *sandbox içinde*. Hastaları gözetimsiz yönetmek üzere serbest bırakıldığı anlamına **gelmez**. Bunlar çok farklı iddialardır ve yalnızca ilki test edildi.
> Sandbox hakkında bir ayrıntı daha önemli, çünkü bunu yanlış hayal etmek kolaydır: MIRA istediği herhangi bir testi *isteyebilir*, ancak sistem yalnızca o test gerçek hayatta bu hastaya

> **gerçekten yapılmışsa** bir sonuç verebilir. Hastanın yaptırdığı bir kan panelini isterse gerçek değerleri alır; kimsenin istemediği bir taramayı isterse uydurma bir sayı değil, "N/A — gerçekleştirilemedi" yanıtını alır. Öykü de aynı şekilde işler: kayıt MIRA'nın sorduğu bilgiyi içermiyorsa simüle hasta yalnızca bilmediğini söyler. Dolayısıyla MIRA gerçekte hiç yapılmamış tek bir belirleyici testi sihirli biçimde ortaya çıkaramaz; yalnızca gerçek değerlendirmede mevcut olanlarla çalışabilir.
> Bu ayrım önemli, çünkü başlıktaki en kolay yanlış okunacak kelime "özerk"tir.

Ne buldular?

Benchmark'ta **MIRA tanı doğruluğunda hekimlerden daha iyi performans gösterdi** — ama bu başlık üç farklı sayıyı saklıyor ve bunları birbirine karıştırmak kolay; o yüzden ayırmak gerekiyor.

MIRA, **574 vakanın tamamında** tek başına çalışırken doğru tanıyı vakaların **%88,9**'unda verdi — puanlama, her hasta için MIMIC-IV'e kaydedilmiş gerçek taburculuk tanısına göre yapıldı. İnsanlarla doğrudan karşılaştırmada doktorlar 574 vakanın tümünü çalışmadı; MIRA'nın eriştiği aynı kayıt ve araçlarla ortak bir **311 vakalık alt küme** üzerinde çalıştı. Bu 311 vakada MIRA **%87,8** aldı ve iki ayrı hekim grubuyla karşılaştırıldı. İlki, 7–11 yıllık deneyime sahip dört uzman hekimden oluşan **kıdemli grup**tu ve **%78,1** aldı. İkincisi, çoğunluğu genç asistanlar ve iki uzmandan oluşan, gerçek bir acil servisin kadro yapısına daha yakın **karma kıdem grubu**ydu ve **%71,1** aldı. Aynı vakalar, aynı araçlar — sıralama yapay zekâ birinci, deneyimli doktorlar ikinci, genç ağırlıklı ekip üçüncü oldu. Makalenin kendi dikkatli ifadesiyle MIRA, sekiz hastalığın tamamında hekimlere "tutarlı biçimde eşdeğerdi ve çoğu zaman onları aştı."

Sonraki kararları da genel olarak iyi dayandı: büyük ölçüde kılavuzlarla uyumlu, ilaç güvenliği açısından uygun ve yatış kararlarında makuldü. Yazarlar beş güvenlik kategorisinde (ilaç etkileşimleri, böbrek dozlaması, alerjiler, QT riski ve opioid reçetelemesi) yüksek şiddette ilaç hatası olmadığını ve 468 vakanın 467'sinde reçetenin doğru olduğunu bildiriyor — ama sistemin "%100 güvenilirliğe ulaşmadığını" da belirtiyorlar. İlk bakışta çarpıcı bir sonuç: tüm vakayı yöneten bir ajan ve tanıda doktorların önüne geçiyor.

Ama galibiyetin *şekli*, galibiyet kadar önemli. Makaleyi inceleyen bağımsız uzmanlar avantajın nereden geldiğine dikkat çekti. Science Media Centre'ın uzman panelinde Dr Wei Xing (University of Sheffield), yapay zekânın tanı doğruluğunda doktorları geçtiği manşet sonucunun **büyük ölçüde apandisit ve pankreatit gibi net test sonuçları olan hastalıklardan** kaynaklandığını belirtti; burada belirleyici bir tarama veya laboratuvar değeri sorunu çözüyor. İnsanların acil servise en sık başvurduğu nedenlerden ikisi olan zatürre ve idrar yolu enfeksiyonunda ise *hem* yapay zekâ hem doktorlar en kötü sonuçlarını aldı ve aralarındaki fark en küçüktü.

İki tarafın oyunu oynama biçiminde de makalenin kendi sayılarında görülen bir asimetri vardı. MIRA laboratuvara daha çok yaslandı: rutin bakımda mevcut laboratuvar analitlerinin yaklaşık **%51**'ini kullandı; uzman hekimlerde bu oran yaklaşık **%28** idi — vaka başına medyan yedi ek kan parametresi. Yazarlar bunun veri setindeki rutin bakım taban çizgisinin *altında* olduğunu, yani "her şeyi

iste" stratejisi olmadığını dikkatle belirtiyor ve daha pahalı kesitsel görüntülemede sistematik artış bulmadıklarını bildiriyor. Yine de daha fazla bilgi, tek başına daha yüksek tanı doğruluğu sağlayabilir — dolayısıyla bu, eşit derecede bilgilendirilmiş iki taraf arasında tam anlamıyla benzer koşullarda bir yarış değildir; Xing de bu farkı doğrudan vurguladı. Maliyet, rahatsızlık ya da gecikmeyi tartmadan her ucuz kan testini istemekte serbest bir klinisyen de kâğıt üzerinde daha isabetli görünürdü.

"Doktorları geçmek" neden "daha iyi doktor olmak" değildir?

Bir benchmark galibiyetini fazla yorumlamak kolaydır. "MIRA daha yüksek puan aldı" ile "MIRA daha iyidir" arasında üç şey duruyor.

Birincisi, **neye göre puanlandığı.** Profesör Julie Jacko (University of Edinburgh), MIRA'nın bazı temel çıktılarının alttaki veri setinde belgelenenlere göre tanımlandığını belirtti — yani sistem, mutlaka en iyi bakımı göstermek yerine **kaydedilmiş klinik davranışı yeniden üretmekten** ödüllendiriliyor. Tarihsel dosya cevap anahtarına dönüşüyor. Bu, benchmark oluşturmak için makul bir yöntem; ama mutlak anlamda doğruluğu değil, yapılanla uyumu ölçüyor. Makaleye hakkını vermek gerekirse bu sorun en çok *tedavi uyumu* metriklerini etkiliyor — MIRA'nın istemleri kayıttakilerle eşleşti mi? — oysa başlıktaki tanı doğruluğu taburculuk tanısına göre puanlanıyor; bu, yalnızca dokümantasyonu taklit etmekten gerçek sonuca daha yakın.

İkincisi, yukarıda sözü edilen **bilgi asimetrisi**: çok daha fazla kan testi (mevcut analitlerin yaklaşık %51'i, karşısında %28) daha fazla kanıt demektir. Doğruluk farkının bir kısmı muhtemelen akıl yürütmeyle değil, daha fazla bilgi "satın alınarak" elde edildi.

Üçüncüsü, **nerede çalıştığı.** Bu, geriye dönük vakalarda, önceden seçilmiş sekiz tanıda ve yalnızca metinle çalışan bir sandbox'tı. Dr Dominic Oliver'ın (University of Oxford) belirttiği gibi gerçek klinik değerlendirme yalnızca hastaların ne söylediğine değil, *nasıl* söylediğine de dayanır — fizik muayene, gözlenen davranış ve beden dili de buna dahildir; bitmiş bir kaydı okuyan metin tabanlı bir ajan bunların hiçbirini görmez. Ve sorunu seçilmiş sekiz tanıdan birine girmeyen bir hasta, bu çalışmanın söyleyebileceği alanın tamamen dışındadır.

Hakemlerin dile getirdiği daha sessiz bir kaygı da var: **veri kontaminasyonu.** MIMIC-IV kamuya açık ve hakkında çok şey yazılmış bir veri seti; dolayısıyla açık internet üzerinde eğitilmiş bir dil modeli makaleleri, vaka tartışmalarını veya verinin kendisini daha önce görmüş olabilir. Dr Midhun Parakkal Unni (University of Sheffield) bunu doğrudan vurguladı — yanıtların bir kısmı eğitim verisindeyse performansın bir bölümü klinik akıl yürütme değil, bellektir; ikisini ancak bağımsız tekrar ayırabilir. Dikkat çekici biçimde yazarlar bu noktayı geçiştirmiyor: sonuçlarının "ihtiyatla olası bir üst sınır olarak yorumlanabileceğini" ve "diğer kamuya açık vakalara genellemeyi olduğundan fazla gösterebileceğini" yazıyorlar — kendi manşetine kendisi tavan koyan bir makale.

Bunların hiçbiri MIRA'yı daha az ilginç yapmıyor. Doğru okumanın ölçülü olması gerektiğini gösteriyor: geriye dönük bir benchmark'ta tüm kayıt boyunca eylem alabilen bir ajan, testleri net olan tanılarda hekimleri geçti — kısmen daha fazla test isteyerek, kısmen de kaydedilmiş dosyayla uyum sağlayarak. Bu gerçek bir yetenek gösterimidir; makinelerin artık doktorlardan daha iyi tanı koyduğuna dair bir hüküm değildir.

Bunun kanıtlamadığı şeyler

- MIRA'nın **gerçek hastalarda çalıştığını göstermez**. Her vaka geriye dönük ve simüleydi; hiçbir hasta MIRA tarafından yönetilmedi.
- Sekiz tanının **ötesinde çalıştığını göstermez**. Önceden seçilmiş kümenin dışındaki hastalıklar — tıbbın dağınık, ayrıştırılmamış çoğunluğu — test edilmedi.
- **Kullanıma sokmanın güvenli olduğunu göstermez**. Yazarlar da genellenebilirlik, güvenlik ve yönetişimin ileriye dönük gerçek dünya çalışmalarına ihtiyaç duyduğunu yazıyor.
- Doktorlarla **adil bir başa baş karşılaştırma kurmaz**. Çok daha fazla kan testinden yararlandı (mevcut analitlerin yaklaşık %51'i, karşısında %28) ve bazı tedavi çıktıları "en iyi bakım" için bağımsız bir gerçekliğe değil, kısmen kaydedilmiş dosyaya göre puanlandı.
- Sonucun **ezberden tamamen arınmış olduğunu göstermez**. Kamuya açık MIMIC-IV verileri modelin eğitim verisiyle çakışmış olabilir; bunu dışlamak için bağımsız tekrar gerekir.
- "**Yapay zekâ doktorların yerini alır**" anlamına gelmez. Bu, kayıt boyunca *eylem alabilen* bir karar destek sistemidir; uzmanların ortak görüşü, gerçek kullanımın klinisyenlerle ortaklık içinde olacağı, yetkinin onlarda kalacağı ve metin kaydının dışarıda bıraktığı her şeyi onların sağlayacağı yönünde.

Kanıt ne kadar güçlü?

İddiayı ikiye ayırmak gerekiyor, çünkü iki yarının kanıtı çok farklı.

Bir yetenek gösterimi olarak — bir dil modeli ajanının bir EHR sandbox'ı içinde öykü, test, tanı ve istemleri zincirleyerek bütün bir klinik vakayı baştan sona yürütebildiği — çalışma gerçekten yeni ve makul ölçüde ikna edici. Yeni olan kısım budur ve dikkat edilmeye değer taraf da budur.

Bir üstünlük iddiası olarak — MIRA'nın *hekimlerden daha iyi* olduğu — kanıt sınırlıdır ve dikkatle okunmalıdır. Belirli bir geriye dönük benchmark'ta, sekiz tanıda, bilgi asimetrisiyle (daha fazla test), tarihsel dosyadan türetilmiş bir cevap anahtarıyla ve gerçek bir eğitim-verisi kontaminasyonu olasılığıyla geçerlidir. "Ajan bu benchmark'ta etkileyici performans gösterdi" demek için yeterlidir. "Ajan bir doktordan daha iyi tanı koyucudur" demek için yeterli değildir ve yazarlar da ikinci iddiayı ileri sürmüyor.

En yararlı tutum ne "AI doktorları yendi" ne de "sadece oyuncak" demektir. Daha doğrusu: kayıtta yalnızca soruları yanıtlamak yerine *eylem alan* yeni tür bir tıbbi yapay zekâ, zor bir geriye dönük benchmark'ta iyi sonuç verdi; şimdi ise henüz denenmediği yerde kendini kanıtlaması gerekiyor: gerçek, önceden ayrıştırılmamış hastalarda, ileriye dönük olarak ve yönetişim altında.

Neden önemli?

Birkaç yıldır tıbbi yapay zekâdaki ilginç soru sessizce değişiyor. Eskiden "bir model tanıyı doğru koyabilir mi?" idi — modeller de giderek daha temiz test kümelerinde evet demeye devam etti. MIRA daha zor bir soruya geçişi temsil ediyor: "bir model *işi yapabilir mi* — bilgi toplayabilir, test isteyebilir, yorumlayabilir, karar verebilir ve tüm iş akışı boyunca eyleme geçebilir mi?" Bu daha yararlı ve daha dürüst bir soru, çünkü hem zorluk hem risk eylem kısmında yaşıyor.

Bu yüzden çerçeveleme önemli. Manşet refleksi — *AI doktorları geçiyor* — sonucun en az yeni ve en az desteklenen bölümüne işaret ediyor. Gerçekten yeni olan daha küçük ama daha sonuç doğurucu: bütün kayıt boyunca ilerleyebilen bir ajan. Bu yetenek dayanırsa, kimin işin başında olduğunu değiştirmeden çok önce **iş akışını** değiştirir. Doktor ortadan kalkmaz; istem verme, yorumlama, sonuçların peşinden gitme — yani iş akışı — böyle bir aracın ilk dokunacağı yerdir.

Ve ispat yükünü doğru yöne çevirir. Geriye dönük vakalarda bir liderlik tablosu galibiyeti, ileriye dönük çalışma yapmak için nedendir; onun yerine geçmez. Yazarlar da bunu söylüyor. MIRA'nın dürüst okuması bir sonraki çalışmaya davettir — alanın zaten bildiği standartla: benchmark'ta değil, hastalarda ölçmek.

Kısa özet

MIRA, sandbox içindeki bir elektronik sağlık kaydında çalışan özerk bir yapay zekâ ajanıdır: öykü alabilir, laboratuvar, görüntüleme ve mikrobiyoloji testleri isteyip yorumlayabilir, tanıya ulaşabilir ve tedavi istemlerini yazabilir. Kamuya açık MIMIC-IV veritabanından alınan, önceden seçilmiş sekiz tanıyı kapsayan 574 geriye dönük vakada test edildiğinde, yazarlar tanı doğruluğunda hekimlerden daha iyi sonuç verdiğini (%88,9 genel; başa baş karşılaştırmada %87,8'e karşı %78,1) ve büyük ölçüde kılavuzlarla uyumlu, ilaç güvenliği açısından uygun kararlar aldığını bildiriyor. Ama değerlendirme geçmiş kayıtlarda, yalnızca metinle yapılan bir simülasyondu; avantajın büyük kısmı sonuçları net testlere sahip hastalıklardan geldi; MIRA doktorların yaklaşık iki katı kadar kan testi kullandı (mevcut analitlerin yaklaşık %51'ine karşı %28); bazı tedavi sonuçları orijinal dosyada ne yazdığına göre puanlandı; ve kamuya açık veri seti gerçek bir eğitim-verisi kontaminasyonu riski taşıyor — yazarlar da kendi sayılarının olası bir üst sınır olabileceğini söylüyor. Gerçek ilerleme, tek tek sorulara yanıt vermek yerine iş akışının tamamında eylem alan bir ajandır. Yazarlar genellenebilirlik, güvenlik ve yönetişimin hâlâ ileriye dönük gerçek dünya çalışmalarına ihtiyaç duyduğunu açıkça belirtiyor — doğru okuma da budur: etkileyici bir yetenek gösterimi; AI'ın doktorlardan daha iyi tanı koyduğunun kanıtı değil ve gerçek klinikte kullanıma hazır bir sistem değil.

Abartısız değerlendirme

Makale ne gösteriyor: Bir dil modeli ajanı (MIRA), sandbox içindeki bir EHR'de bir klinik vakayı baştan sona — öykü, testler, tanı, istemler — yürütebiliyor ve sekiz tanıyı kapsayan 574 vakalık geriye

dönük benchmark'ta tanı doğruluğunda hekimlerden daha iyi sonuç veriyor (%88,9 genel; uzman hekimlerle karşılaştırmada %87,8'e karşı %78,1); yüksek şiddette ilaç hatası da raporlanmadı.

Makul ama kanıtlanmamış olan: Bu yeteneğin gerçek, önceden ayrıştırılmamış hastalara fayda sağlayacağı; doğruluk avantajının daha fazla test istemekten, kaydedilmiş dosyayla uyumdan veya kamuya açık veriyi ezberlemekten ziyade daha iyi akıl yürütmeyi yansıttığı.

Göstermediği şey: MIRA'nın sekiz tanısının dışında çalıştığı; kullanıma sokulmasının güvenli olduğu; doktorları adil ve eşit bilgiye sahip bir karşılaştırmada geçtiği; klinisyenlerin yerini aldığı; sonuçların eğitim-verisi kontaminasyonundan arınmış olduğu.

Başlıca sınırlamalar: Geriye dönük, simüle edilmiş, yalnızca metin tabanlı değerlendirme; önceden seçilmiş sekiz tanı; bilgi asimetrisi (çok daha fazla kan testi — düşük maliyetli kan testlerinde mevcut analitlerin yaklaşık %51'ine karşı %28, görüntülemede değil); tedavi çıktılarının kısmen tarihsel dosyaya göre puanlanması; ve bir dil modelinin eğitim sırasında görmüş olabileceği kamuya açık bir benchmark (MIMIC-IV) — yazarların da performans için olası bir üst sınır olarak işaret ettiği nokta.

Genel okuyucu ne kadar güvenmeli? MIRA'nın gerçek ve yeni bir yetenek gösterimi olduğuna — yalnızca chatbot değil, kayıt boyunca *eylem alan* bir ajan — yüksek güven. Bir *hekimden daha iyi tanı koyucu* olduğuna düşük güven: bu iddia tek bir geriye dönük benchmark'la sınırlı ve test isteme, dosyaya göre puanlama ve olası kontaminasyonla karışmış durumda. Yazarların kendi sonucu doğru olanı: bunun hasta bakımı için anlam taşımadan önce ileriye dönük gerçek dünya çalışmalarında sınanması gerekiyor.

Kaynaklar

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