The Clean Paper

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

AI agent ทำเคสผู้ป่วยทั้งเคสได้เอง — แต่ในเครื่องจำลองและบนเวชระเบียนย้อนหลัง

MIRA เป็น AI ทางการแพทย์รูปแบบใหม่: แทนที่จะตอบคำถามเดียว มันทำทั้งเคสในเวชระเบียนโรงพยาบาลจำลอง — ซักประวัติ สั่งและอ่านผลตรวจ วินิจฉัย และเขียนคำสั่งรักษา บนเคสย้อนหลัง 574 เคสจากฐานข้อมูลสาธารณะ ครอบคลุมการวินิจฉัยที่เลือกไว้ล่วงหน้า 8 โรค ผู้เขียนรายงานว่ามันมีความแม่นยำวินิจฉัยสูงกว่าแพทย์ และการตัดสินใจส่วนใหญ่สอดคล้อง guideline และปลอดภัยด้านยา แต่ทุกข้อขยายในประโยคนั้นมีน้ำหนัก: มันทำงานใน sandbox บนข้อมูลเก่า ใช้ข้อความเท่านั้น; ความได้เปรียบจำนวนมากมาจากโรคที่มีผลตรวจชี้ชัด; มันส่งตรวจเลือดประมาณสองเท่าของแพทย์; และ outcome หลายตัวให้คะแนนเทียบกับสิ่งที่ chart เดิมบันทึก ความก้าวหน้าจริงคือ agent ที่ลงมือทำตลอด workflow — ไม่ใช่หลักฐานว่าเครื่องจักรวินิจฉัยดีกว่าแพทย์แล้ว ผู้เขียนเองบอกว่า generalization ความปลอดภัย และ governance ยังต้องการการศึกษาล่วงหน้าในโลกจริง

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

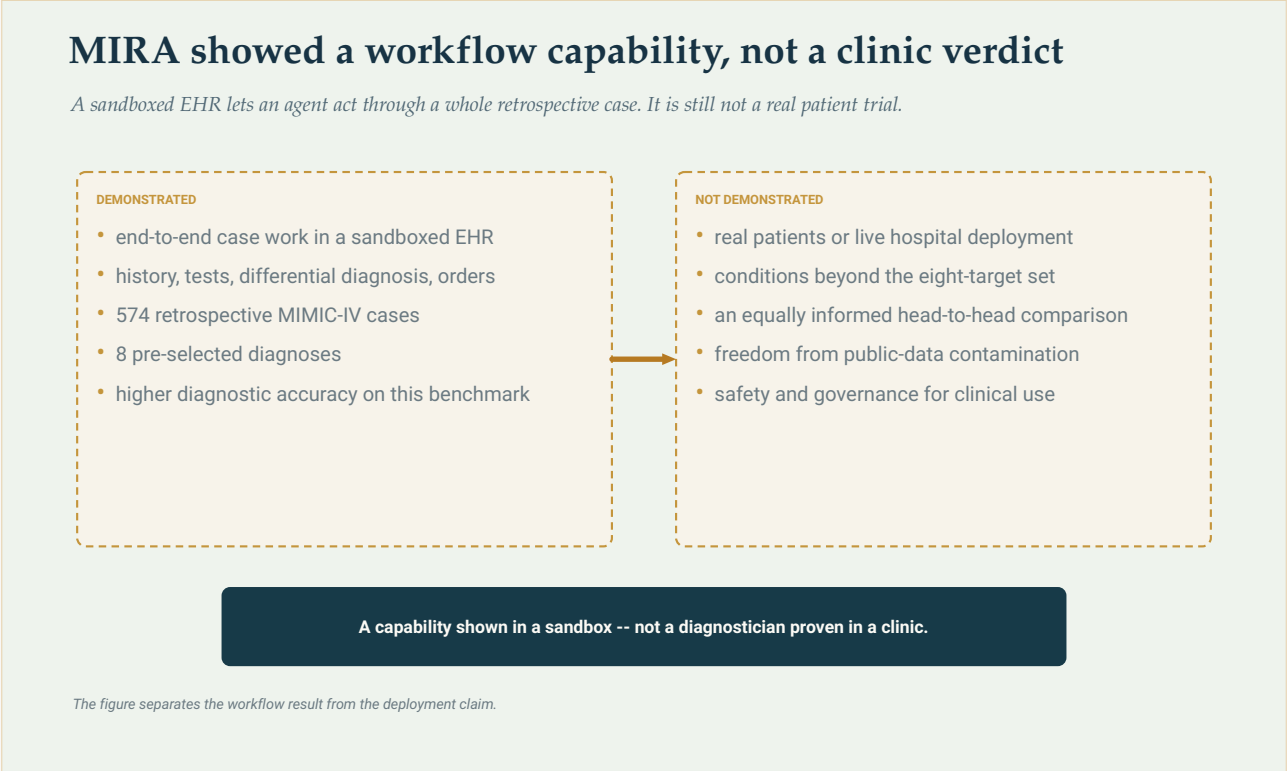
ตอบคำถามเก่ง ไม่เหมือนกับจัดการผู้ป่วยทั้งเคสได้

เรื่องราวเกี่ยวกับ AI ทางการแพทย์ส่วนใหญ่พูดถึงโมเดลที่ **ตอบคำถาม** เราให้กรณีตัวอย่างหรือข้อสอบหนึ่งข้อ มันตอบกลับด้วยการวินิจฉัยหรือคำแนะนำหนึ่งย่อหน้า แล้วได้คะแนนดี นั่นมีประโยชน์ แต่ขอบเขตยังแคบ: คล้ายที่ปรึกษาฉลาด ๆ ที่ไม่เคยลงมือแตะเวชระเบียนจริง

MIRA ถูกสร้างมาเพื่อทำสิ่งที่ต่างออกไป และความต่างนี้ต่างหากคือประเด็นสำคัญ แทนที่จะตอบคำถามเดียว มันพยายามจัดการ **ทั้งเคส** ตั้งแต่อ่านประวัติผู้ป่วย ตัดสินใจว่ายังต้องถามอะไรเพิ่ม สั่งตรวจเลือด ภาพถ่ายทางการแพทย์ และจุลชีววิทยา อ่านผล ตรวจสอบการวินิจฉัยแยกโรค แล้วเขียนคำสั่งรักษาต่อ — ตั้งแต่สั่งยา รับไว้รักษาในโรงพยาบาล ไปจนถึงส่งต่อเพื่อผ่าตัด มันทำสิ่งเหล่านี้ด้วยการออกคำสั่ง **ภายในเวชระเบียนอิเล็กทรอนิกส์** คล้ายกับขั้นตอนการทำงานของผู้ให้การรักษา ไม่ใช่เพียงสร้างข้อความให้มนุษย์นำไปคัดลอกต่อ

นี่เป็นก้าวจริง จากระบบที่ให้คำแนะนำไปสู่ระบบที่ลงมือทำงานหลายขั้นตอน แต่ก็เป็นก้าวประเภทเดียวกับที่พาดหัวข่าวชอบย่อเหลือเพียง “AI ทำได้ดีกว่าหมอ” ดังนั้นจึงต้องแยกให้ชัดว่าทดลองอะไร และทดลองที่ไหน

สรุปในประโยคเดียว: MIRA จัดการเคสตั้งแต่ต้นจนจบได้ด้วยตัวเอง และในชุดทดสอบนี้มีความแม่นยำในการวินิจฉัยสูงกว่าแพทย์ แต่ทั้งหมดเกิดใน **สภาพแวดล้อมจำลองที่แยกจากระบบจริง** ใช้เวชระเบียนย้อนหลัง ครอบคลุมโรคเพียงแปดกลุ่มที่เลือกไว้ล่วงหน้า สื่อสารผ่านข้อความเท่านั้น และข้อได้เปรียบส่วนใหญ่เกิดในโรคที่มีผลตรวจค่อนข้างชี้ชัด ความสามารถใหม่มีอยู่จริง แต่ตารางคะแนนไม่ใช่คลินิก



MIRA แสดงให้เห็นว่าสามารถทำงานตั้งแต่ต้นจนจบภายในเวชระเบียนอิเล็กทรอนิกส์จำลองได้ แต่ยังไม่ได้แสดงว่าสามารถใช้จริงในคลินิก ครอบคลุมการวินิจฉัยได้กว้าง หรือเอาชนะแพทย์ในการเปรียบเทียบที่ทั้งสองฝ่ายได้รับข้อมูลเท่าเทียมกัน

ผู้วิจัยทำอะไร

MIRA ย่อมาจาก Medical Intelligence for Reasoning and Action เป็นเอเจนต์อัตโนมัติที่สร้างบนโมเดลของ OpenAI ส่วนที่สนทนาและลงมือทำใช้ **GPT-4o** ส่วนขั้นตอนวางแผนแยกต่างหากใช้โมเดลให้เหตุผล **o1** ระบบทั้งหมดทำงานในกรอบที่รองรับมาตรฐาน HL7 FHIR ซึ่งเป็นมาตรฐานข้อมูลที่โรงพยาบาลจริงใช้แลกเปลี่ยนเวชระเบียน และมีชุดการกระทำทางคลินิกที่เป็นไปได้มากกว่าแปดหมื่นรายการ

ในสภาพแวดล้อมจำลอง MIRA สามารถดึงประวัติผู้ป่วย สั่งและตีความผลตรวจเลือด ภาพถ่ายทางการแพทย์ และจุลชีววิทยา สร้างรายการวินิจฉัยแยกโรค แล้วจัดทำแผนรักษา เช่น สั่งยา วางแผนผ่าตัด หรือรับผู้ป่วยเข้าโรงพยาบาล

มีรายละเอียดหนึ่งที่ควรระบุไว้ตั้งแต่ต้น: “ผู้ตัดสิน” อัตโนมัติที่ใช้ให้คะแนนคำตอบของ MIRA เองก็เป็น **GPT-4o** ซึ่งอยู่ในตระกูลโมเดลเดียวกับระบบที่กำลังถูกทดสอบ หลังผู้ประเมินบทความตั้งข้อกังวล ผู้เขียนจึงเพิ่มการตรวจทานโดยแพทย์ที่ได้รับวุฒิบัตรเฉพาะทางด้วย ชุดทดสอบมาจาก **MIMIC-IV** ฐานข้อมูลเวชระเบียนอิเล็กทรอนิกส์ขนาดใหญ่ที่เปิดให้สาธารณะใช้และลบข้อมูลระบุตัวบุคคลแล้ว จากผู้ป่วยราว 300,000 คนที่รักษาที่ Beth Israel Deaconess Medical Center ระหว่างปี 2008–2019 ผู้เขียนเลือก **574 เคส** ครอบคลุม **การวินิจฉัยเป้าหมาย 8 กลุ่ม** ทั้งโรคช่องท้องและภาวะฉุกเฉินทางอายุรกรรม เช่น ไส้ติ่งอักเสบ ตับอ่อนอักเสบ ปอดบวม และการติดเชื้อทางเดินปัสสาวะ แต่ละเคสเริ่มจากข้อมูลจำกัด MIRA ต้องตัดสินใจทีละขั้นว่าจะทำอะไรต่อ เหมือนโครงสร้างของการวินิจฉัยในชีวิตจริง เพียงแต่ระบบทำงานกับเคสย้อนหลังซึ่งผลลัพธ์จริงถูกบันทึกไว้แล้ว จากนั้นจึงนำผลงานไปเปรียบเทียบกับแพทย์ในเคสเดียวกัน

> **“เอเจนต์อัตโนมัติในเวชระเบียนจำลอง” หมายความว่าอย่างไร**
>
> คำสามคำนี้มีความหมายมาก จึงควรแยกทีละคำ
> **เอเจนต์ (agent)** หมายถึงระบบไม่ใช่แชตบอตที่รับคำถามแล้วตอบครั้งเดียว แต่ทำงานเป็นวงรอบ: ดูสถานะปัจจุบัน เลือกการกระทำ เช่น สั่งตรวจหรือถามประวัติเพิ่ม รับผล แล้วเลือกการกระทำถัดไป จนได้การวินิจฉัยและแผนการรักษา ความสามารถด้านภาษาและเหตุผลมาจากโมเดลภาษา ส่วนความเป็น “เอเจนต์” มาจากโครงสร้างรอบโมเดลที่แปลงข้อความเป็นคำสั่งที่ระบบเวชระเบียนอนุญาต และส่งผลกลับให้โมเดลตัดสินใจต่อ
> **เวชระเบียนแบบแซนด์บ็อกซ์ (sandboxed EHR)** หมายถึงสำเนาจำลองที่แยกจากโรงพยาบาลจริง MIRA สามารถ “สั่ง” การตรวจแล้วได้รับผลที่ผู้ป่วยคนนั้นเคยได้รับจริง เพราะเคสเป็นข้อมูลย้อนหลังและผลถูกบันทึกไว้แล้ว ไม่มีการกระทำใดไปแตะผู้ป่วยหรือระบบคลินิกจริง
> **อัตโนมัติ (autonomous)** หมายถึงระบบทำเคสตั้งแต่ต้นจนจบโดยไม่มีมนุษย์คอยตัดสินใจระหว่างทาง **ภายในแซนด์บ็อกซ์** ไม่ได้หมายความว่าปล่อยให้ระบบดูแลผู้ป่วยจริงโดยไม่มีการกำกับ สองข้ออ้างนี้ต่างกันมาก และงานนี้ทดสอบเฉพาะข้อแรก
> มีรายละเอียดอีกอย่างที่ช่วยให้เห็นข้อจำกัดชัดขึ้น MIRA จะสั่งตรวจอะไรก็ได้ แต่ระบบคืนผลได้เฉพาะการตรวจที่ผู้ป่วยคนนั้น **เคยได้รับจริง** ในชีวิตจริง หากขอตรวจเลือดที่มีอยู่ในเวชระเบียน ระบบก็คืนค่าจริงให้ แต่ถ้าขอภาพสแกนที่ไม่เคยมี ระบบจะตอบว่าไม่สามารถทำได้ ไม่สร้างตัวเลขขึ้นมาเอง เช่นเดียวกับการถามประวัติ หากไม่มีข้อมูลในเวชระเบียน ผู้ป่วยจำลองก็ตอบเพียงว่าไม่ทราบ ดังนั้น MIRA ไม่สามารถเสก “การตรวจเด็ดขาด” ที่โลกจริงไม่เคยทำได้ มันต้องทำงานกับร่องรอยที่เวชระเบียนจริงมีอยู่
> ความแตกต่างนี้สำคัญ เพราะคำว่า “อัตโนมัติ” เป็นคำที่ถูกอ่านเกินความหมายได้ง่ายที่สุด

พบอะไร

ในชุดทดสอบ **MIRA มีความแม่นยำในการวินิจฉัยสูงกว่าแพทย์** แต่ตัวเลขพาดหัวจริง ๆ มีสามชุดและไม่ควรนำมาปนกัน เมื่อดู MIRA เพียงลำพังใน **574 เคสทั้งหมด** มันวินิจฉัยถูก **88.9%** โดยเทียบกับการวินิจฉัยตอนจำหน่ายที่บันทึกจริงใน

MIMIC-IV ส่วนการเปรียบเทียบโดยตรงกับมนุษย์ แพทย์ไม่ได้ทำครบ 574 เคส แต่ทำชุดย่อยร่วมกัน **311 เคส** ด้วยข้อมูลและเครื่องมือแบบเดียวกับที่ MIRA ได้รับ

ใน 311 เคสนั้น MIRA ได้ **87.8%** และถูกเทียบกับแพทย์สองกลุ่ม กลุ่มแรกเป็น **แพทย์อาวุโส** สี่คนที่ได้รับวุฒิบัตรเฉพาะทางและมีประสบการณ์ 7–11 ปี ได้ **78.1%** กลุ่มที่สองเป็น **กลุ่มประสบการณ์ผสม** ซึ่งส่วนใหญ่เป็นแพทย์ประจำบ้านรุ่นต้นร่วมกับผู้เชี่ยวชาญสองคน ได้ **71.1%** ลำดับจึงเป็น MIRA ก่อน แพทย์อาวุโสตาม และกลุ่มที่มีแพทย์รุ่นต้นมากกว่าตามหลัง ผู้เขียนใช้ถ้อยคำระมัดระวังกว่านั้น โดยระบุว่า MIRA “ทำได้อย่างน้อยเทียบเท่า และหลายครั้งดีกว่า” แพทย์ในโรคทั้งแปดกลุ่ม

การตัดสินใจต่อจากการวินิจฉัยก็ดูดีพอสมควร แผนรักษาส่วนใหญ่สอดคล้องแนวทาง ยาปลอดภัย และการตัดสินใจรับเข้าโรงพยาบาลเหมาะสม ผู้เขียนรายงานว่าไม่พบข้อผิดพลาดด้านยาระดับรุนแรงในห้าหมวดความปลอดภัย ได้แก่ ปฏิกิริยาระหว่างยา การปรับขนาดยาตามไต ภูมิแพ้ ความเสี่ยง QT และการสั่ง opioid และพบใบสั่งยาถูกต้อง 467 จาก 468 เคส ขณะเดียวกันก็ระบุชัดว่าระบบ “ยังไม่ได้ความน่าเชื่อถือ 100%”

หากมองเพียงตัวเลข นี่เป็นผลลัพธ์ที่โดดเด่น: เอเจนต์ทำทั้งเคสและยังมีคะแนนวินิจฉัยสูงกว่าแพทย์ แต่ **รูปแบบของชัยชนะ** สำคัญพอ ๆ กับตัวชัยชนะเอง

ผู้เชี่ยวชาญอิสระที่อ่านงานชี้ว่า MIRA ได้เปรียบมากที่สุดในโรคที่มีผลตรวจค่อนข้างชี้ชัด เช่น ไส้ติ่งอักเสบและตับอ่อนอักเสบ ซึ่งภาพสแกนหรือค่าตรวจเฉพาะสามารถตัดสินคำตอบได้ค่อนข้างตรง ขณะที่ปอดบวมและการติดเชื้อทางเดินปัสสาวะ — สองเหตุผลที่พบบ่อยมากในการมาห้องฉุกเฉิน — ทั้ง AI และแพทย์ทำได้แย่ลง และช่องว่างระหว่างกันเล็กลง ดูความเห็นผู้เชี่ยวชาญจาก Science Media Centre

ยังมีความไม่สมมาตรด้านข้อมูลที่สำคัญ MIRA ใช้ค่าตรวจเลือดประมาณ **51%** ของรายการที่มีอยู่ในการดูแลตามปกติ เทียบกับประมาณ **28%** สำหรับแพทย์ที่ได้รับวุฒิบัตรเฉพาะทาง คิดเป็นค่าตรวจเลือดเพิ่มขึ้นมัธยฐานราวเจ็ดรายการต่อเคส ผู้เขียนเน้นว่าตัวเลขนี้ยังต่ำกว่าระดับการใช้ข้อมูลจริงในฐานข้อมูลและไม่ได้มีการเพิ่มการตรวจภาพราคาแพงอย่างเป็นระบบ แต่ข้อมูลมากขึ้นย่อมช่วยเพิ่มความแม่นยำในการวินิจฉัยได้ด้วยตัวมันเอง ดังนั้นนี่ไม่ใช่การแข่งแบบสองฝ่ายได้รับข้อมูลเท่ากันทุกประการ

ทำไม “ชนะหมอ” ไม่เท่ากับ “เป็นหมอที่ดีกว่า”

ชัยชนะบนชุดทดสอบถูกอ่านเกินจริงได้ง่าย มีอย่างน้อยสามเรื่องคั่นอยู่ระหว่าง “MIRA ได้คะแนนสูงกว่า” กับ “MIRA ดีกว่าแพทย์”

อย่างแรกคือ **เกณฑ์ที่ใช้ให้คะแนน** ผลลัพธ์สำคัญหลายตัวถูกนิยามเทียบกับสิ่งที่ถูกบันทึกไว้ในฐานข้อมูลเดิม ดังนั้นระบบบางส่วนจึงได้รับคะแนนจากการ **ทำซ้ำพฤติกรรมทางคลินิกที่ถูกบันทึกในอดีต** ไม่ใช่จากการแสดงว่าการดูแลนั้นดีที่สุดในความหมายสัมบูรณ์ เวชระเบียนย้อนหลังจึงทำหน้าที่เป็นเฉลย นี่เป็นวิธีสร้างชุดทดสอบที่สมเหตุสมผล แต่ต้องเข้าใจว่ามันวัดความสอดคล้องกับสิ่งที่เคยเกิดขึ้น ไม่ใช่ความถูกต้องสูงสุดในทุกกรณี ข้อจำกัดนี้กระทบตัวชี้วัดด้านการรักษามากที่สุด ส่วนคะแนนวินิจฉัยหลักเทียบกับการวินิจฉัยตอนจำหน่าย ซึ่งใกล้กับผลลัพธ์จริงมากกว่า

อย่างที่สองคือ **ความไม่สมมาตรของข้อมูล** ที่กล่าวไปแล้ว MIRA ใช้ผลตรวจเลือดมากกว่าอย่างชัดเจน ข้อมูลมากขึ้นย่อมชื่อความแม่นยำบางส่วนได้ ไม่ใช่ทุกส่วนของช่องว่างจะต้องมาจากการให้เหตุผลที่เหนือกว่า

อย่างที่สามคือ **สภาพแวดล้อมที่ใช้ทดสอบ** นี่เป็นแซนด์บ็อกซ์ ใช้เคสย้อนหลัง ครอบคลุมโรคแปดกลุ่มที่เลือกไว้ล่วงหน้า และสื่อสารผ่านข้อความเท่านั้น ในชีวิตจริง การประเมินผู้ป่วยไม่ได้ขึ้นอยู่กับสิ่งที่ผู้ป่วยพูดอย่างเดียว แต่รวมถึงการตรวจร่างกาย สีหน้า พฤติกรรม และภาษากาย ซึ่งเอเจนต์ที่อ่านข้อความจากเวชระเบียนย้อนหลังไม่มีโอกาสเห็น และหากปัญหาของผู้ป่วยไม่อยู่ในโรคแปดกลุ่มที่เลือก งานนี้ก็ไม่ได้บอกว่า MIRA จะทำอย่างไร

ยังมีข้อกังวลเงียบ ๆ อีกเรื่องคือ **การปนเปื้อนจากข้อมูลฝึก** MIMIC-IV เป็นฐานข้อมูลสาธารณะที่ถูกใช้และเขียนถึงอย่างแพร่

หลาย โมเดลภาษาที่ฝึกจากอินเทอร์เน็ตอาจเคยเห็นบทความ การอภิปรายเคส หรือข้อมูลบางส่วนมาก่อน หากคำตอบบางส่วนเคยอยู่ในข้อมูลฝึก ความสามารถบางส่วนอาจมาจากความจำมากกว่าการให้เหตุผลทางคลินิก ต้องอาศัยการทำซ้ำกับข้อมูลอิสระจึงจะแยกสองอย่างออกจากกันได้ ที่น่าสนใจคือผู้เขียนไม่ได้มองข้ามประเด็นนี้ แต่เขียนว่าผลลัพธ์อาจถือเป็น **ขอบเขตบนโดยประมาณ** และอาจประเมินความสามารถในการนำไปใช้กับเคสสาธารณะอื่นสูงเกินจริง

สิ่งเหล่านี้ไม่ได้ทำให้ MIRA ไม่น่าสนใจ แต่ทำให้การอ่านที่ถูกต้องต้องได้สัดส่วน: ในชุดทดสอบย้อนหลัง เอเจนต์ที่สามารถลงมือทำงานตลอดเวชระเบียนมีคะแนนวินิจฉัยสูงกว่าแพทย์ โดยเฉพาะในโรคที่มีผลตรวจชี้ชัด ส่วนหนึ่งเพราะใช้ข้อมูลตรวจมากกว่า และตัวชี้วัดบางส่วนอิงสิ่งที่ถูกบันทึกไว้ในเวชระเบียน นี่เป็นการสาธิตความสามารถจริง ไม่ใช่คำตัดสินว่าเครื่องวินิจฉัยดีกว่าหมอแล้ว

สิ่งที่ผลนี้ *ยังพิสูจน์ไม่ได้*

- **ไม่ได้แสดงว่า MIRA ใช้ได้กับผู้ป่วยจริง** ทุกเคสเป็นข้อมูลย้อนหลังและการจำลอง ไม่มีผู้ป่วยคนใดได้รับการดูแลจริงจากระบบ
- **ไม่ได้แสดงว่าใช้ได้เกินโรคแปดกลุ่มที่เลือกไว้** ความเจ็บป่วยส่วนใหญ่ในชีวิตจริงซับซ้อนและยังไม่ถูกจำแนกตั้งแต่ต้น งานนี้ไม่ได้ทดสอบสถานการณ์นั้น
- **ไม่ได้แสดงว่าปลอดภัยพอสำหรับการใช้งานจริง** ผู้เขียนระบุเองว่าความสามารถในการนำไปใช้ทั่วไป ความปลอดภัย และธรรมาภิบาลยังต้องทดสอบแบบไปข้างหน้าในโลกจริง
- **ไม่ได้เป็นการเปรียบเทียบกับแพทย์ที่ข้อมูลเท่าเทียมกันอย่างสมบูรณ์** MIRA ใช้ผลตรวจเลือดมากกว่ามาก และตัวชี้วัดด้านการรักษาบางส่วนให้คะแนนเทียบกับเวชระเบียนย้อนหลัง
- **ไม่ได้พิสูจน์ว่าผลปราศจากการจำข้อมูล** MIMIC-IV เป็นข้อมูลสาธารณะที่อาจทับซ้อนกับข้อมูลฝึกของโมเดล ต้องทดสอบซ้ำกับข้อมูลอิสระเพื่อแยกความจำออกจากการให้เหตุผล
- **ไม่ได้หมายความว่า "AI แทนหมอ"** ความสามารถของระบบคือการช่วยตัดสินใจและลงมือทำภายในเวชระเบียน แต่การใช้จริงยังต้องอาศัยผู้ให้การรักษาที่มีอำนาจและเติมข้อมูลที่ข้อความในเวชระเบียนไม่มี

หลักฐานน่าเชื่อถือเพียงใด

ควรแยกข้ออ้างออกเป็นสองส่วน เพราะความแข็งแรงของหลักฐานต่างกันมาก

ในฐานะการสาธิตความสามารถ — ว่าเอเจนต์ที่ใช้โมเดลภาษาสามารถจัดการเคสทั้งเคสภายในเวชระเบียนจำลอง เชื่อมประวัติ การตรวจ การวินิจฉัย และคำสั่งรักษาตั้งแต่ต้นจนจบ — งานนี้มีความใหม่จริงและค่อนข้างน่าเชื่อ นี่คือส่วนที่ควรให้ความสนใจมากที่สุด

ในฐานะข้ออ้างว่าเหนือกว่าแพทย์ หลักฐานมีขอบเขตชัดเจน ผลนี้เป็นจริงบนชุดทดสอบย้อนหลังเฉพาะหนึ่งชุด ครอบคลุมโรคแปดกลุ่ม มีความไม่สมมาตรของข้อมูลจากการใช้ผลตรวจมากกว่า และใช้เวชระเบียนย้อนหลังเป็นเกณฑ์ให้คะแนนบางส่วน รวมทั้งยังมีความเป็นไปได้ที่ข้อมูลสาธารณะจะเคยปรากฏในข้อมูลฝึก หลักฐานเพียงพอที่จะพูดว่า "เอเจนต์ทำผลงานได้น่าประทับใจบนชุดทดสอบนี้" แต่ยังไม่พอที่จะพูดว่า "เอเจนต์เป็นผู้วินิจฉัยที่ดีกว่าแพทย์" และผู้เขียนเองก็ไม่ได้อ้างอย่างหลัง

ท่าทีที่เหมาะจึงไม่ใช่ "AI ชนะหมอแล้ว" และไม่ใช่ "เป็นแค่ของเล่น" แต่คือ: AI ทางการแพทย์ชนิดใหม่ที่ **ลงมือทำงานตลอดเวชระเบียน** แทนที่จะตอบคำถามแยก ๆ ทำผลงานได้ดีบนชุดทดสอบย้อนหลังที่ค่อนข้างยาก และขั้นต่อไปคือการพิสูจน์สิ่งที่ยังไม่เคยทดสอบ — ผู้ป่วยจริงที่ไม่ได้ถูกคัดเป็นโรคกลุ่มใดล่วงหน้า การทดลองแบบไปข้างหน้า และการกำกับดูแลที่เหมาะสม

ทำไมจึงสำคัญ

ในช่วงไม่กี่ปีที่ผ่านมา คำถามที่น่าสนใจเกี่ยวกับ AI ทางการแพทย์ค่อย ๆ เปลี่ยนไป เดิมคือ “โมเดลวินิจฉัยถูกหรือไม่?” และโมเดลก็ทำคะแนนดีขึ้นเรื่อย ๆ บนชุดทดสอบที่สะอาดขึ้นเรื่อย ๆ MIRA ย้ายคำถามไปสู่สิ่งที่ยากกว่า: **โมเดลทำงานจริงทั้งกระบวนการได้หรือไม่ — เก็บข้อมูล สังตรวจ อ่านผล ตัดสินใจ และลงมือทำ?** นี่เป็นคำถามที่มีประโยชน์และซื่อตรงกว่า เพราะจุดที่ระบบต้อง “ลงมือ” คือจุดที่ทั้งความยากและความเสี่ยงเริ่มจริงจัง

จึงยิ่งต้องระวังการตีกรอบ พาดหัวแบบ “AI ทำได้ดีกว่าหมอ” ชี้ไปยังส่วนที่ใหม่ที่สุดน้อยกว่าและมีข้อจำกัดมากกว่า สิ่งใหม่ที่แท้จริงเล็กกว่าแต่มีผลลึกกว่า: **เอเจนต์ที่เดินผ่านเวชระเบียนทั้งระบบได้** หากความสามารถนี้ยืนยันได้ในโลกจริง สิ่งแรกที่มันเปลี่ยนน่าจะเป็นขั้นตอนการทำงาน — การสังตรวจ การติดตามผล การรวบรวมข้อมูล — มากกว่าการเปลี่ยนว่าใครเป็นผู้รับผิดชอบสูงสุด

และงานนี้ยังกำหนดภาระการพิสูจน์ในทิศทางที่ถูกต้อง ชัยชนะบนชุดทดสอบย้อนหลังเป็นเหตุผลให้ทำการทดลองแบบไปข้างหน้า ไม่ใช่สิ่งทดแทนการทดลองนั้น ผู้เขียนเองก็พูดเช่นนั้น การอ่าน MIRA อย่างตรงไปตรงมาคือการมองว่าเป็นคำเชิญไปสู่งานวิจัยขั้นต่อไป ซึ่งต้องวัดกับผู้ป่วย ไม่ใช่กับตารางคะแนน

สรุป

MIRA เป็นเอเจนต์ AI อัตโนมัติที่ทำงานในเวชระเบียนอิเล็กทรอนิกส์จำลอง สามารถซักประวัติ สั่งและตีความผลตรวจเลือด ภาพถ่ายทางการแพทย์และจุลชีววิทยา สรุปการวินิจฉัย และเขียนแผนการรักษาได้ เมื่อทดสอบกับเคสย้อนหลัง 574 เคสจากฐานข้อมูลสาธารณะ MIMIC-IV ครอบคลุมโรคแปดกลุ่ม MIRA มีความแม่นยำในการวินิจฉัยสูงกว่าแพทย์ (88.9% ในทุกเคส; 87.8% เทียบกับ 78.1% ในการเปรียบเทียบโดยตรงกับแพทย์อาวุโส) และการตัดสินใจด้านการรักษาส่วนใหญ่สอดคล้องแนวทางและปลอดภัยด้านยา แต่การประเมินทั้งหมดเป็นการจำลองบนเวชระเบียนย้อนหลัง ใช้ข้อความอย่างเดียว ข้อได้เปรียบส่วนใหญ่เกิดในโรคที่มีผลตรวจชี้ชัด MIRA ใช้ผลตรวจเลือดมากกว่าแพทย์มาก (ประมาณ 51% ของรายการที่มี เทียบกับ 28%) ตัวชี้วัดด้านการรักษาบางส่วนให้คะแนนเทียบกับเวชระเบียนเดิม และฐานข้อมูลสาธารณะทำให้มีความเสี่ยงจริงเรื่องข้อมูลฝึกทับซ้อน ผู้เขียนเองมองว่าตัวเลขอาจเป็นขอบเขตบนของความสามารถ ความก้าวหน้าที่แท้จริงคือ **เอเจนต์ที่ทำงานตลอดขั้นตอนการดูแลได้** ไม่ใช่หลักฐานว่า AI วินิจฉัยดีกว่าแพทย์ และยังไม่ใช่ระบบที่พร้อมใช้กับผู้ป่วยจริง

ประเมินแบบไม่เกินจริง

งานวิจัยแสดงอะไร: เอเจนต์ที่ใช้โมเดลภาษาอย่าง MIRA สามารถจัดการเคสทางคลินิกตั้งแต่ต้นจนจบภายในเวชระเบียนจำลอง — ตั้งแต่ประวัติ การตรวจ การวินิจฉัย ไปจนถึงคำสั่งรักษา — และบนชุดทดสอบย้อนหลัง 574 เคสในโรคแปดกลุ่มมีความแม่นยำในการวินิจฉัยสูงกว่าแพทย์ โดยไม่พบข้อผิดพลาดด้านยาระดับรุนแรง

อะไรที่เป็นไปได้แต่ยังไม่พิสูจน์: ความสามารถนี้อาจแปลเป็นประโยชน์ต่อผู้ป่วยจริงที่ไม่ได้ถูกคัดโรคล่วงหน้า และข้อได้เปรียบด้านความแม่นยำอาจสะท้อนการให้เหตุผลที่ดีกว่า ไม่ใช่เพียงการใช้ผลตรวจมากกว่า การทำตามสิ่งที่เวชระเบียนเคยบันทึก หรือการจำข้อมูลสาธารณะ

มันไม่ได้แสดงอะไร: ไม่ได้แสดงว่า MIRA ใช้ได้เกินโรคแปดกลุ่ม ปลอดภัยพอสำหรับการใช้งานจริง เอาชนะแพทย์ในการเปรียบเทียบที่ข้อมูลเท่าเทียมกัน แทนผู้ให้การรักษาได้ หรือปราศจากการปนเปื้อนจากข้อมูลฝึก

ข้อจำกัดหลัก: การประเมินย้อนหลังและจำลอง ใช้ข้อความอย่างเดียว โรคแปดกลุ่มที่เลือกไว้ล่วงหน้า มีความไม่สมมาตรด้าน

ข้อมูลจากการใช้ผลตรวจเลือดมากกว่า ตัวชี้วัดการรักษาบางส่วนให้คะแนนเทียบกับเวชระเบียนเดิม และ MIMIC-IV เป็นชุดข้อมูลสาธารณะที่โมเดลภาษาอาจเคยเห็นระหว่างการฝึก ผู้เขียนเองจึงระบุว่าผลอาจเป็นขอบเขตบนของประสิทธิภาพ

ผู้อ่านทั่วไปควรมั่นใจแค่ไหน? มั่นใจสูงว่านี่เป็นการสาธิตความสามารถรูปแบบใหม่จริง — เอเจนต์ที่ **ลงมือทำงานตลอดเวชระเบียน** ไม่ใช่แชตบอตตอบคำถาม แต่ควรมีความมั่นใจต่ำต่อข้ออ้างว่า MIRA เป็นผู้วินิจฉัยที่ดีกว่าแพทย์โดยทั่วไป ข้ออ้างนั้นยังจำกัดอยู่ในชุดทดสอบย้อนหลังหนึ่งชุดและมีปัจจัยกวนหลายอย่าง ขั้นต่อไปที่จำเป็นคือการศึกษาแบบไปข้างหน้าในโลกจริงกับผู้ป่วยจริง

แหล่งข้อมูล

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