



The CLEAN PAPER

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

AI Agent ניהל מקרי מטופלים שלמים בכוחות עצמו — בסימולטור, על רשומות עבר

MIRA היא סוג חדש של AI רפואי. במקום לענות על שאלה יחידה, היא עובדת מקרה שלם בתוך רשומת בית חולים מדומה — לוקחת היסטוריה, מזמינה וקוראת בדיקות, מגיעה לאבחנה וכותבת הוראות. על 574 מקרים רטרוספקטיביים ממאגר ציבורי, על פני שמונה אבחנות שנבחרו מראש, המחברים מדווחים שהיא עלתה על רופאים בדיוק האבחנות וקיבלה החלטות ברובן תואמות הנחיות ובטוחות מבחינת תרופות. כל מסייג במשפט הזה חשוב. היא רצה ב-sandbox על רשומות עבר, בטקסט בלבד, חלק גדול מן היתרון הגיע במחלות עם תוצאות בדיקה חד-משמעיות, היא השתמשה בערך פי שניים יותר בדיקות דם מן הרופאים, וכמה תוצאות קיבלו ציון מול מה שהתיק המקורי תיעד. ההתקדמות האמיתית היא agent שפועל לאורך workflow שלם — לא הוכחה שמכונה מאבחנת עכשיו טוב יותר מרופא. המחברים אומרים זאת. הכללה, בטיחות וממשל עדיין דורשים מחקרים פרוספקטיביים בעולם האמיתי.

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

לענות על שאלה זה לא אותו דבר כמו לנהל את המקרה

רוב סיפורי ה-AI הרפואי עוסקים במודל ש-עונה. נותנים לו ויניטה או שאלת בחינה, הוא מחזיר אבחנה או פסקת עצה ומקבל ציון טוב. שימושי, אבל צר: יועץ חכם שאף פעם לא נוגע בתיק הרפואי.

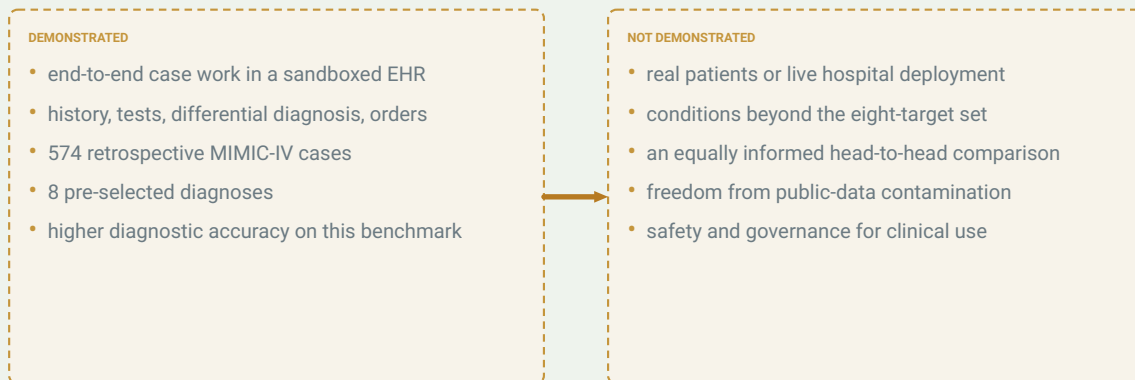
MIRA נבנתה לעשות משהו אחר, וההבדל הזה הוא החדשות האמיתיות. במקום לענות על שאלה אחת, היא עובדת על מקרה שלם: קוראת את רשומת המטופל, מחליטה איזו אנמנזה עדיין חסרה, מזמינה בדיקות מעבדה, הדמיה ומיקרוביולוגיה, קוראת את התוצאות, מצמצמת אבחנה מבדלת, ואז כותבת את ההוראות הבאות — מרשמים, אשפוז, הפניה לניתוח. היא עושה זאת באמצעות פעולות בתוך רשומה רפואית אלקטרונית, כפי שקלינאי עושה, במקום לייצר טקסט חופשי שאדם צריך להעתיק.

זהו צעד אמיתי: ממערכת שמייצעת למערכת שפועלת לאורך תהליך עבודה שלם. וזה בדיוק סוג הצעד שהסיקור נוטה לשטח ל"AI עולה על רופאים". לכן כדאי לדייק במה נבדק — ואיפה.

גרסת המשפט האחד: MIRA טיפלה במקרים שלמים בכוחות עצמה ובמבחן ההשוואה הזה עלתה על רופאים בדיוק האבחנתי — אבל עשתה זאת בסביבת הדמיה מבודדת, על רשומות רטרוספקטיביות, על פני שמונה אבחנות שנבחרו מראש, בתקשורת טקסט בלבד, וחלק גדול מן היתרון הגיע במחלות עם תוצאות הבדיקה הברורות ביותר. ההתקדמות אמיתית. לוח התוצאות אינו המרפאה.

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.



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

The figure separates the workflow result from the deployment claim.

MIRA הדגימה יכולת תהליך עבודה מקצה לקצה בתוך EHR בסביבת הדמיה מבודדת. היא לא הדגימה פריסה קלינית בעולם האמיתי, כיסוי אבחנתי רחב או השוואה הוגנת שבה היא והרופאים קיבלו אותו מידע.

מה עשו המחקרים

MIRA — Action and Reasoning for Intelligence Medical — היא מערכת אוטונומית המבוססת על מודלים של OpenAI: החלק שמשוחח ופועל רץ על GPT-4o, ושלב תכנון נפרד משתמש במודל ההנמקה o1 של OpenAI. אלה יושבים בתוך מסגרת מותאמת ותואמת תקנים — המבוססת על HL7, FHIR תקן הנתונים שבו בתי חולים אמיתיים משתמשים להעברת רשומות — עם ארגז כלים של יותר משמונים אלף פעולות קליניות אפשריות. בתוך סביבת ההדמיה המבודדת MIRA יכולה לשלוף היסטוריה של מטופל, להזמין ולפרש מעבדות, הדמיה ומיקרוביולוגיה, לבנות אבחנה מבדלת ולנסח תכניות טיפול — לרשום תרופות, לתזמן פרוצדורות ניתוחיות ולתכנן אשפוזים. פרט אחד שכדאי לסמן מראש: ה"שופטים" האוטומטיים שציינו את תשובות MIRA היו בעצמם GPT-4o — אותה משפחת מודלים שנבדקה — והמחקרים גיבו זאת בבדיקה של רופא מוסמך לאחר שמבקר עמיתים העלה את החשש.

סט הבדיקה הגיע מ-MIMIC-IV, מאגר EHR גדול, ציבורי ומנוטרל-זהויות. מתוך כ-300,000 מטופלים שטופלו ב-Beth Center Medical Deaconess Israel בין 2008 ל-2019, המחקרים בחרו 574 מקרי מטופלים על פני שמונה אבחנות יעד — פתולוגיות בטניות ומצבי חירום ברפואה פנימית, ובהם, pneumonia, appendicitis, וזיהום בדרכי השתן. בכל מקרה MIRA התחילה מתמונה מוגבלת והייתה צריכה להחליט, שלב אחר שלב, מה לעשות הלאה — אותה צורה כמו בירור אמיתי, אך מול רשומה שסופה האמיתי כבר קיים בתיק.

לאחר מכן הביצועים שלה הושושו לרופאים על אותם מקרים.

מה באמת פירוש "סוכן אוטונומי בתוך EHR בסביבת הדמיה מבודדת"

אותן. לפרק כדאי ולכן עבודה, הרבה כאן עושות מילים שלוש Agent פירושו שהמערכת אינה צ'טבוט שעונה ל-prompt. היא מריצה לולאה: מסתכלת על המצב הנוכחי, בוחרת פעולה (להזמין בדיקה זו, לשאול על פרט היסטוריה אחר), רואה את התוצאה, בוחרת פעולה נוספת — עד שהיא מגיעה לאבחנה ולתכנית. ה"אינטליגנציה" היא מודל שפה; ה-agency היא התשתית שסביבו, שהופכת את הטקסט של המודל לפעולות EHR מותרות ומחזירה את התוצאות למודל. EHR Sandboxed פירושו עותק מדומה ומבודד של מערכת רשומות רפואיות, לא מערכת חיה של בית חולים. MIRA יכולה "להזמין" בדיקה ולקבל את התוצאה שהמטופל באמת קיבל, משום שהמקרה היסטורי והתשובה כבר רשומה. שום פעולה שלה אינה נוגעת במטופל אמיתי או במרפאה אמיתית. Autonomous פירושו שהיא משלימה את המקרה מקצה לקצה בלי אדם בלולאה — בתוך סביבת ההדמיה המבודדת. זה לא אומר ששחררו אותה לנהל מטופלים בלי פיקוח. אלה טענות שונות מאוד, ורק הראשונה נבדקה.

עוד דבר על סביבת ההדמיה המבודדת, כי קל לדמיין אותו לא נכון: MIRA יכולה להזמין כל בדיקה שהיא רוצה, אבל המערכת יכולה להחזיר תוצאה רק אם הבדיקה באמת בוצעה אצל אותו מטופל בחיים האמיתיים. אם היא מבקשת פאנל דם שהמטופל קיבל, היא מקבלת את הערכים האמיתיים; אם היא מבקשת סריקה שאיש לא הזמין, היא מקבלת "N/A" — "be not could", לעולם לא מספר מומצא. ההיסטוריה פועלת באותה דרך: אם הרשומה אינה מכילה את מה ש-MIRA שואלת, המטופל המדומה פשוט אומר שאינו יודע. לכן MIRA אינה יכולה לזמן את הבדיקה המכריעה היחידה שמעולם לא נעשתה — היא יכולה לעבוד רק עם מה שהבירור האמיתי במקרה כלל בפועל. ההבחנה חשובה משום שמילת הכותרת — "אוטונומי" — היא זו שהכי קל לקרוא כאילו מדובר במשמעות השנייה.

מה הם מצאו

במבחן ההשוואה, MIRA עלתה על רופאים בדיוק האבחנתי — אבל הכותרת מסתירה שלושה מספרים שונים שקל לערבב, ולכן כדאי להפריד.

לבדה, על פני כל 574 המקרים, MIRA נתנה את האבחנה הנכונה ב-88.9% מן המקרים — לפי אבחנת השחרור שנרשמה בפועל לכל מטופל ב-MIMIC-IV. בהשוואה הישירה לבני אדם, הרופאים לא עבדו על כל 574 המקרים; הם עבדו על תת-קבוצה משותפת של 311 מקרים, עם אותן רשומות ואותם כלים שהיו ל-MIRA. באותם 311 מקרים MIRA קיבלה 87.8%, והשווה לשתי קבוצות שגויסו בנפרד. הראשונה הייתה קבוצה בכירה: ארבעה רופאים מוסמכים עם 7–11 שנות ניסיון, שקיבלו 78.1%. השנייה הייתה קבוצה ברמות ותק מעורבות: בעיקר מתמחים צעירים ועוד שני מומחים, קרוב יותר לאופן שבו מחלקת חירום אמיתית מאוישת, שקיבלו 71.1%. אותם מקרים, אותם כלים — והסדר יצא AI ראשון, הרופאים המנוסים שניים, והצוות הצעיר יותר שלישי. הניסוח הזהיר של המאמר עצמו הוא ש-MIRA הייתה "שקולה באופן עקבי, ולעיתים קרובות עלתה" על הרופאים בכל שמונה המחלות.

גם ההחלטות בהמשך שרשרת הטיפול החזיקו מעמד: ברובן תואמות הנחיות, בטוחות מבחינת תרופות ומתאימות מבחינת אשפוז. המחברים מדווחים על אפס טעויות תרופה בחומרה גבוהה בחמש קטגוריות בטיחות (אינטראקציות בין תרופות, מינון כלייתי, אלרגיות, סיכון QT ומתן אופיואידים) ועל מרשמים נכונים ב-467 מתוך 468 מקרים — ובאותה נשימה מציינים שהמערכת "לא הגיעה לאמינות של 100%". בקריאה כפשוטה, זו תוצאה מרשימה: סוכן שמנהל את המקרה כולו ויוצא לפני רופאים באבחנה.

אבל צורת הניצחון חשובה לא פחות מן הניצחון. מומחים עצמאיים שקראו את המאמר הצביעו על המקום שממנו הגיע היתרון. ב-פאנל המומחים של Centre Media Science, ד"ר Xing Wei מאוניברסיטת Sheffield ציין שהמספר הראשי — AI שמנצח רופאים בדיוק אבחנתי — הונע בעיקר בידי מחלות עם תוצאות בדיקה חד-משמעיות, כמו pancreatitis-i appendicitis, שבהן סריקה או ערך מעבדה מכריע פותרים את השאלה. בדלקת ריאות ובזיהומים בדרכי השתן — שתיים מן הסיבות השכיחות ביותר שבגללן אנשים מגיעים לחדר מיון — גם ה-AI וגם הרופאים היו הכי חלשים, והפער ביניהם היה הקטן ביותר.

הייתה גם א-סימטריה באופן שבו שני הצדדים שיחקו, והיא מופיעה במספרי המאמר עצמו. MIRA נשענה יותר על המעבדה: היא השתמשה בכ-51% מן האנליטים המעבדתיים הזמינים בטיפול שגרתי, לעומת כ-28% אצל הרופאים המוסמכים — חציון של שבעה פרמטרי דם נוספים לכל מקרה. המחברים מקפידים למסגר זאת כ-פחות מקו הבסיס של הטיפול השגרתי במאגר עצמו, לא כאסטרטגיית "להזמין הכול", ומדווחים שאין עלייה שיטתית בהדמיה חתכית יקרה יותר. ובכל זאת, יותר מידע יכול לבדו לייצר דיוק אבחנתי גבוה יותר — ולכן זו אינה ממש תחרות זהה בין שחקנים בעלי אותו מידע, פער ש-Xing ציין במפורש. גם קלינאי שחופשי להזמין כל בדיקת דם זולה בלי לשקלל עלות, אי-נוחות או עיכוב היה נראה חד יותר על הנייר.

למה "ניצחה רופאים" אינו "רופא טוב יותר"

קל לקרוא יותר מדי לתוך יתרון במבחן השוואה. שלושה דברים מפרידים בין MIRA "קיבלה ציון גבוה יותר" לבין MIRA "טובה יותר".

ראשית, מול מה היא קיבלה ציון. פרופ' Jacko Julie מאוניברסיטת Edinburgh ציינה שכמה מן התוצאות המרכזיות של MIRA מוגדרות ביחס למה שתועד במאגר המקורי — כלומר המערכת מתוגמלת על שחזור ההתנהגות הקלינית שתועדה, לא בהכרח על הדגמת הטיפול האופטימלי. התיק ההיסטורי נעשה מפתח התשובות. זו דרך סבירה לבנות מבחן השוואה, אבל היא מודדת התאמה למה שנעשה ולא נכונות במובן מוחלט. לזכות המאמר, הבעיה הזאת פוגעת

במיוחד במדדי התאמת הטיפול — האם ההוראות של MIRA תאמו את התיק? — בעוד שהדיוק האבחנתי הראשי נמדד מול אבחנת השחרור, שקרובה יותר לתוצאה אמיתית מאשר הד של תיעוד.

שנית, א-סימטריית המידע שכבר הוזכרה: הרבה יותר בדיקות דם (כ-51% מן האנליטים הזמינים לעומת 28%) פירושן יותר ראיות. סביר שחלק מפער הדיוק בקנה במידע, לא רק נובע מהנמקה טובה יותר.

שלישית, איפה המערכת רצה. זו הייתה סביבת הדמיה מבודדת, על מקרים רטרוספקטיביים, על פני שמונה אבחנות שנבחרו מראש, בטקסט בלבד. הערכה קלינית אמיתית, כפי שניסח ד"ר Oliver Dominic מאוניברסיטת Oxford, תלויה לא רק במה שהמטופל אומר אלא גם באיך הוא אומר זאת — לצד בדיקה גופנית, התנהגות נצפית ושפת גוף, שאף אחת מהן אינה נגישה לסוכן טקסטואלי שקורא רשומה גמורה. ומטופל שהבעיה שלו אינה שייכת לאחת משמונה האבחנות פשוט מחוץ למה שהמחקר הזה יכול לומר עליו.

יש גם דאגה שקטה יותר שמבקרים העלו: זיהום נתונים (data contamination). MIMIC-IV ציבורי ונכתב עליו הרבה, ולכן מודל שפה שאומן על האינטרנט הפתוח אולי כבר ראה מאמרים, דיוני מקרים או אפילו את הנתונים עצמם. ד"ר Unni Parakkal Midhun מאוניברסיטת Sheffield העלה זאת ישירות — אם חלק מן התשובות היו בנתוני האימון, חלק מן הביצועים הם זיכרון ולא reasoning קליני, ורק שחזור עצמאי יכול להבחין בין השניים. חשוב שהמחברים אינם מנפנים את הנקודה: הם כותבים שאפשר "לפרש בזהירות את התוצאות כגבול עליון אפשרי" ושייתכן שהן "מעריכות יתר על המידה הכללה למקרים ציבוריים אחרים" — מאמר שמציב תקרה על הכותרת של עצמו.

שום דבר מזה אינו הופך את MIRA לפחות מעניינת. הוא הופך את הקריאה הנכונה למכילת: במבחן ההשוואה הרטרוספקטיבי, סוכן שיכול לפעול לאורך כל הרשומה עלה על רופאים באבחנות עם בדיקות נקיות — חלקית מפני שהזמין יותר בדיקות, חלקית מפני שהתאים לתיק שתועד. זו הדגמת יכולת אמיתית, לא פסק דין שמכונות מאבחנות עכשיו טוב יותר מרופאים.

מה זה לא מוכיח

- הוא לא מראה ש-MIRA עובדת על מטופלים אמיתיים. כל מקרה היה רטרוספקטיבי ומדומה; אף מטופל לא נוהל על ידה.
- הוא לא מראה שהיא עובדת מעבר לשמונה אבחנות. מצבים מחוץ לסט שנבחר מראש — הרוב המבולגן והלא-ממויין של הרפואה — לא נבדקו.
- הוא לא מראה שהיא בטוחה לפריסה. המחברים עצמם כותבים שהכללה, בטיחות וממשל עדיין דורשים מחקרים פרוספקטיביים בעולם האמיתי.
- הוא לא מבסס השוואה הוגנת מול רופאים. הוא השתמש בהרבה יותר בדיקות דם (כ-51% מן האנליטים הזמינים לעומת 28%), וכמה תוצאות טיפול קיבלו ציון חלקית מול התיק שתועד, לא מול אמת-מידה של טיפול מיטבי.
- הוא לא מראה שהתוצאה חופשייה משינון. נתוני MIMIC-IV הציבוריים עשויים לחפוף עם אימון המודל, דבר ששחזור עצמאי יצטרך לשלול.
- הוא לא אומר "AI מחליף רופאים". זו תמיכה בהחלטות שיכולה לפעול בתוך רשומה; הקונצנזוס של המבקרים הוא ששימוש אמיתי יהיה בשותפות עם קלינאים, ששומרים על סמכות ומספקים את כל מה שרשומת טקסט משאירה בחוץ.

עד כמה הראיות חזקות?

כדאי לפצל את הטענה לשניים, משום שהראיות שונות מאוד בכל חצי.

כהדגמת יכולת — שסוכן המבוסס על מודל שפה יכול לנהל מקרה קליני שלם בתוך EHR בסביבת הדמיה מבודדת, ולשרשר היסטוריה, בדיקות, אבחנה והוראות מקצה לקצה — העבודה באמת חדשה ומשכנעת במידה סבירה. זה החלק החדש וזה החלק שכדאי לשים אליו לב.

כטענת עליונות — ש-MIRA טובה מרופאים — הראיות מוגבלות וצריך לקרוא אותן בזהירות. הן מחזיקות במבחן השוואה רטרופקטיבי מסוים, בשמונה אבחנות, עם א-סימטריית מידע (יותר בדיקות) ומפתח תשובות שנלקח בחלקו מן התיק ההיסטורי, ועם אפשרות אמיתית של זיהום נתוני אימון. זה מספיק כדי לומר "הסוכן ביצע באופן מרשים במבחן ההשוואה הזה". זה לא מספיק כדי לומר "הסוכן מאבחן טוב יותר מרופא", והמחברים אינם טוענים את הדבר השני.

העמדה השימושית ביותר אינה AI "מנצח רופאים" וגם לא "רק צעצוע". היא: סוג חדש של AI רפואי — כזה שפועל לאורך הרשומה במקום רק לענות על שאלות — עשה עבודה טובה במבחן השוואה רטרופקטיבי קשה, ועכשיו צריך להוכיח את עצמו במקום שעדיין לא נוסה: על מטופלים אמיתיים ולא-ממוינים, פרוספקטיבית ותחת ממשל.

למה זה חשוב

בשנים האחרונות השאלה המעניינת ב-AI רפואי השתנתה בשקט. פעם היא הייתה "האם המודל יכול לתת את האבחנה הנכונה?" — ומודלים המשיכו לענות כן, על סטי בדיקה נקיים יותר ויותר. MIRA מסמנת מעבר לשאלה קשה יותר: "האם מודל יכול לעשות את העבודה — לאסוף, להזמין, לפרש, להחליט, לפעול — לאורך תהליך עבודה שלם?" זו שאלה שימושית יותר וגם כנה יותר, משום שהפעולה היא המקום שבו נמצאים גם הקושי וגם הסיכון.

לכן המסגור חשוב. רפלקס הכותרת — AI עולה על רופאים — מצביע על החלק הכי פחות חדש והכי פחות מבוסס בתוצאה. הדבר החדש באמת קטן יותר ובעל השלכות גדולות יותר: סוכן שיכול לנוע לאורך כל הרשומה. אם היכולת הזאת תחזיק מעמד, היא תשנה את תהליך העבודה הרבה לפני שתשנה מי אחראי עליו. הרופא אינו מוחלף; ההזמנה, הפרשנות, המעקב אחר תוצאות — תהליך העבודה — הם המקום שבו כלי כזה ינחת ראשון.

והוא מאפס את נטל ההוכחה לכיוון הנכון. מקום גבוה בטבלת דירוג על מקרים רטרופקטיביים הוא סיבה להריץ ניסוי פרוספקטיבי, לא תחליף לו. המחברים אומרים זאת בעצמם. הקריאה הכנה של MIRA היא הזמנה למחקר הבא — תחת הסטנדרט שהתחום כבר מכיר: נמדד על מטופלים, לא על מבחני השוואה.

בקצרה

MIRA היא מערכת AI אוטונומית שפועלת בתוך רשומה רפואית אלקטרונית בסביבת הדמיה מבודדת: היא יכולה לקחת היסטוריה, להזמין ולפרש מעבדות, הדמיה ומיקרוביולוגיה, להגיע לאבחנה ולכתוב תכנית טיפול. בבדיקה על 574 מקרים רטרופקטיביים מן המאגר הציבורי MIMIC-IV, על פני שמונה אבחנות שנבחרו מראש, היא עלתה על רופאים בדיוק האבחנות (88.9% בכלל המקרים; 87.8% לעומת 78.1% בהשוואה הישירה) וקיבלה החלטות ברובן תואמות הנחיות ובטוחות מבחינת תרופות. אבל ההערכה הייתה סימולציה על רשומות עבר, בטקסט בלבד; חלק גדול מן היתרון הגיע במחלות עם תוצאות בדיקה חד-משמעיות; MIRA השתמשה בהרבה יותר בדיקות דם מן הרופאים

(כ-51% מן האנליטים הזמינים לעומת 28%); כמה תוצאות טיפול קיבלו ציון מול מה שהרשומה המקורית תיעדה; והמאגר הציבורי מעלה סיכון אמיתי לזיהום נתוני אימון — שהמחברים עצמם מכנים גבול עליון אפשרי למספרים. ההתקדמות האמיתית היא סוכן שפועל לאורך תהליך עבודה שלם במקום לענות על שאלות מבודדות. המחברים מפורשים שהכללה, בטיחות וממשל עדיין דורשים מחקרים פרוספקטיביים בעולם האמיתי — וזו הקריאה הנכונה: הדגמת יכולת מרשימה, לא הוכחה ש-AI מאבחן טוב יותר מרופאים, ולא מערכת מוכנה למרפאה אמיתית.

בדיקה מפוכחת

מה המאמר מראה: סוכן המבוסס על מודל שפה (MIRA) יכול לנהל מקרה קליני שלם מקצה לקצה בתוך EHR בסביבת הדמיה מבודדת — היסטוריה, בדיקות, אבחנה, הוראות — ובמבחן השוואה רטרוספקטיבי של 574 מקרים על פני שמונה אבחנות, עלה על רופאים בדיוק אבחנותי (88.9% בכלל; 87.8% לעומת 78.1% מול רופאים מוסמכים), בלי טעויות תרופה בחומרה גבוהה.

מה סביר אך לא הוכח: שהיכולת הזאת מתורגמת לתועלת למטופלים אמיתיים ולא-ממוינים; שיתרון הדיוק משקף reasoning טוב יותר ולא יותר בדיקות שהוזמנו, התאמה לתיק שתועד או נתונים ציבוריים שנשנו.

מה הוא לא מראה: ש-MIRA עובדת מחוץ לשמונה האבחנות; שהיא בטוחה לפריסה; שהיא מנצחת רופאים בהשוואה הוגנת ובעלת אותו מידע; שהיא מחליפה קלינאים; או שהתוצאות נקיות מזיהום נתוני אימון.

המגבלות העיקריות: הערכה רטרוספקטיבית, מדומה ובטקסט בלבד; שמונה אבחנות שנבחרו מראש; א-סימטריית מידע (הרבה יותר בדיקות דם — כ-51% מן האנליטים הזמינים לעומת 28%, בבדיקות דם זולות ולא בהדמיה); תוצאות טיפול שנמדדו חלקית מול התיק ההיסטורי; ובמבחן השוואה ציבורי (MIMIC-IV) שמודל שפה אולי ראה באימון — דבר שהמחברים מסמנים כגבול עליון אפשרי לביצועים.

כמה ביטחון צריך להיות לקורא כללי? גבוה בכך ש-MIRA היא הדגמת יכולת אמיתית וחדשה — סוכן ש-פועל לאורך הרשומה, לא רק צ'טבוט. נמוך בכך שהיא מאבחנת טוב יותר מרופא: הטענה הזאת מוגבלת למבחן השוואה רטרוספקטיבי אחד ומעורפלת בידי הזמנת בדיקות, ניקוד מול התיק וזיהום אפשרי. השורה התחתונה של המחברים היא הנכונה — צריך מחקר פרוספקטיבי בעולם האמיתי לפני שיש לכך משמעות לטיפול במטופלים.

מקורות

(2026) Nature agents, intelligence artificial medical autonomous Towards al., et Ferber
<https://doi.org/10.1038/s41586-026-10675-5>

abstract peer-reviewed — 42310457 PubMed
<https://pubmed.ncbi.nlm.nih.gov/42310457/>

(2026) AMIE and MIRA to reaction expert — Centre Media Science
<https://www.sciencemediacentre.org/expert-reaction-to-presentation-of-two-new-medical-ai->

framing doctors' 'outperforms the illustrates — (2026) News-Medical
<https://www.news-medical.net/news/20260621/Autonomous-medical-AI-outperforms-doctors-in-s.aspx>