

Trial finds GPT-4 access did not improve doctors' diagnoses, but GPT-4 alone scored higher

Source copy kept by wheresthe.ai for the story "Trial finds GPT-4 access did not improve doctors' diagnoses, but GPT-4 alone scored higher" (<https://wheresthe.ai/d/us-jama-llm-diagnostic-reasoning-rct-2024-10/>).

Source: <https://doi.org/10.1001/jamanetworkopen.2024.40969>

Text saved 21 Oct 2026 (the abstract from PubMed (NCBI), because the journal page refuses automated reading).

SHA-256 of the saved text: 5c02b04c8f394daf613abc8fb9685df4c8a393489521c6ee8a2703a9572fe45a

1. JAMA Netw Open. 2024 Oct 1;7(10):e2440969. doi: 10.1001/jamanetworkopen.2024.40969.

Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial.

Goh E(1)(2), Gallo R(3), Hom J(4), Strong E(4), Weng Y(5), Kerman H(6)(7), Cool JA(6)(7), Kanjee Z(6)(7), Parsons AS(8), Ahuja N(4), Horvitz E(9)(10), Yang D(11), Milstein A(2), Olson APJ(12), Rodman A(6)(7), Chen JH(1)(2)(13).

Author information: (1)Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California. (2)Stanford Clinical Excellence Research Center, Stanford University, Stanford, California. (3)Center for Innovation to Implementation, VA Palo Alto Health Care System, Palo Alto, California. (4)Department of Hospital Medicine, Stanford University School of Medicine, Stanford, California. (5)Quantitative Sciences Unit, Stanford University School of Medicine, Stanford, California. (6)Department of Hospital Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts. (7)Department of Hospital Medicine, Harvard Medical School, Boston, Massachusetts. (8)Department of Hospital Medicine, School of Medicine, University of Virginia, Charlottesville. (9)Microsoft Corp, Redmond, Washington. (10)Stanford Institute for Human-Centered Artificial Intelligence, Stanford, California. (11)Department of Hospital Medicine, Kaiser Permanente, Oakland, California. (12)Department of Hospital Medicine, University of Minnesota Medical School, Minneapolis. (13)Division of Hospital Medicine, Stanford University, Stanford, California.

Update of medRxiv. 2024 Mar 14:2024.03.12.24303785. doi: 10.1101/2024.03.12.24303785. medRxiv. 2024 Aug 07:2024.08.05.24311485. doi: 10.1101/2024.08.05.24311485.

IMPORTANCE: Large language models (LLMs) have shown promise in their performance on both multiple-choice and open-ended medical reasoning examinations, but it remains unknown whether the use of such tools improves physician diagnostic reasoning. **OBJECTIVE:** To assess the effect of an LLM on physicians' diagnostic reasoning compared with conventional resources. **DESIGN, SETTING, AND PARTICIPANTS:** A single-blind randomized clinical trial was conducted from November 29 to December 29, 2023. Using remote video conferencing and in-person participation across multiple academic medical institutions, physicians with training in family medicine, internal medicine, or emergency medicine were recruited. **INTERVENTION:** Participants were randomized to either access the LLM in addition to conventional diagnostic resources or conventional resources only, stratified by career stage. Participants were allocated 60 minutes to review up to 6 clinical vignettes. **MAIN OUTCOMES AND MEASURES:** The primary outcome was performance on a standardized rubric of diagnostic performance based on differential diagnosis accuracy, appropriateness of supporting and opposing factors, and next diagnostic evaluation steps, validated and graded via blinded expert consensus. Secondary outcomes included time spent per case (in seconds) and final diagnosis accuracy. All analyses followed the intention-to-treat principle. A secondary exploratory analysis evaluated the standalone performance of the LLM by comparing the primary outcomes between the LLM alone group and the conventional resource group. **RESULTS:** Fifty physicians (26 attendings, 24 residents; median years in practice, 3 [IQR, 2-8]) participated virtually as well as at 1 in-person site. The median diagnostic reasoning score per case was 76% (IQR, 66%-87%) for the LLM group and 74% (IQR, 63%-84%) for the conventional resources-only group, with an adjusted difference of 2 percentage points (95% CI, -4 to 8 percentage points; $P = .60$). The median time spent per case for the LLM group was 519 (IQR, 371-668) seconds, compared with 565 (IQR, 456-788) seconds for the conventional resources group, with a time difference of -82 (95% CI, -195 to 31; $P = .20$) seconds. The LLM alone scored 16 percentage points (95% CI, 2-30 percentage points; $P = .03$) higher than the conventional resources group. **CONCLUSIONS AND RELEVANCE:** In this trial, the availability of an LLM to physicians as a diagnostic aid did not significantly improve clinical reasoning compared with conventional resources. The LLM alone demonstrated higher performance than both physician groups, indicating the need for technology and workforce development to realize the potential of physician-artificial

intelligence collaboration in clinical practice. TRIAL REGISTRATION: ClinicalTrials.gov Identifier: NCT06157944.

DOI: 10.1001/jamanetworkopen.2024.40969 PMCID: PMC11519755 PMID: 39466245 [Indexed for MEDLINE]

Conflict of interest statement: Conflict of Interest Disclosures: Dr Kanjee reported book royalties from and paid membership on the Wolters Kluwer advisory board for medical education products, and personal fees from Oakstone Publishing for continuing medical education lectures on evidence-based medicine outside the submitted work. Dr Parsons reported receiving grants from the American Medical Association and the Southern Group on Educational Affairs outside the submitted work. Dr Yang reported being an employee of the Gordon and Betty Moore Foundation during the conduct of the study. Dr Milstein reported receiving personal fees for advisory board membership from the Peterson Center of Healthcare; holding stock options in Emsana Health, Amino Health, FNF Advisors, JRSL LLC, Embold, EZPT/Somatic Health, and Prealize outside the submitted work; and membership on the Leapfrog Group Board Intermountain Healthcare Board. Dr Olson reported receiving grants from 3M and the Agency for Healthcare Quality and Research outside the submitted work. Dr Chen reported receiving grants from the National Institutes of Health National Institute on (NIH) National Institute of Allergy and Infectious Diseases, NIH National Institute on Drug Abuse Clinical Trials Network, and the American Heart Association; nonfinancial support from Reaction Explorer LLC; personal fees from multiple legal offices as a medicolegal expert witness; grants from Google Inc and Stanford University; and personal fees from ISHI Health Consulting outside the submitted work. No other disclosures were reported.