MACHINE LEARNING

Performance of a large language model on the reasoning tasks of a physician

Peter G. Brodeur[1]†, Thomas A. Buckley[2]†, Zahir Kanjee[1], Ethan Goh[3,4], Evelyn Bin Ling[5], Priyank Jain[6], Stephanie Cabral[1,7], Raja-Elie Abdulnour[8], Adrian D. Haimovich[9], Jason A. Freed[10], Andrew Olson[11], Daniel J. Morgan[12,13], Jason Hom[5], Robert Gallo[14,15], Liam G. McCoy[1,16,17], Haadi Mombini[18], Christopher Lucas[1], Misha Fotoohi[1], Matthew Gwiazdon[1], Daniele Restifo[1], Daniel Restrepo[19], Eric Horvitz[20,21], Jonathan Chen[3,4,22]‡,Arjun K. Manrai[2]‡*, Adam Rodman[1]‡*

More than 65 years ago, complex clinical diagnostic reasoning cases were introduced as the gold standard for the evaluation of expert medical computing systems, a standard that has held ever since. In this study, we report the results of a physician evaluation of a large language model (LLM) on challenging clinical cases across five experiments with a baseline of hundreds of physicians. We then report a real-world study comparing human expert and artificial intelligence (AI) second opinions in randomly selected patients in the emergency room of a major tertiary academic medical center. In all experiments, the LLM outperformed physician baselines and displayed continued improvement from prior generations of AI clinical decision support. Our study suggests that LLMs have eclipsed most benchmarks of clinical reasoning, motivating the urgent need for prospective trials.

Artificial intelligence (AI) diagnostic support tools have been studied since the 1950s, after a landmark paper was published in *Science* by Ledley and Lusted (*1*), who advocated for case-based benchmarks as an evaluation standard, a standard that has held for more than the past half century (*2*–*8*). In particular, the *New England Journal of Medicine* (*NEJM*) clinicopathological case conference series has been seen as an aspirational goal post, tested by every differential diagnosis generator spanning primitive Bayesian systems, symbolic rules–based systems, and natural-language symptom checkers. Recently, large language models (LLMs) have consistently outperformed older models on these challenging cases, mirroring their performance in professional licensing exams, mathematics questions, software engineering, and engineering problems (*9*–*12*).

However, recent studies of LLMs in medicine have focused on narrow diagnostic tasks or on curated and constrained clinical vignettes (*7*, *13*, *14*). More importantly, most studies of LLMs for diagnosis and management to date have lacked human physician baselines. This was justifiable in previous generations of technology because of the overall poor performance of prior computational models on benchmarks. Given rapid improvement in LLMs and increasing "benchmark saturation," it is necessary to establish human baselines and study clinically grounded tasks. Here, we comprehensively evaluated the diagnostic and management reasoning capabilities of an advanced LLM (OpenAI o1 series) across several diagnostic and management reasoning tasks with baseline performance from hundreds of physicians. We further studied LLM second opinions in a blinded fashion against an expert physician baseline on randomly selected patients in a major academic tertiary care emergency department in Boston, Massachusetts.

Results

Quality of differential diagnoses and testing plans on *NEJM* clinicopathological conferences

We first evaluated o1-preview using the clinicopathologic conferences (CPCs) published by the *NEJM*, a standard for the evaluation of differential generators since the 1950s (*1*). There was substantial agreement between the two physicians evaluating the quality of o1-preview's differential diagnosis [agreement on 120/143 cases (84%), inter-rater reliability (κ) = 0.66]. o1-preview included the correct diagnosis in its differential in 78.3% of cases [95% confidence interval (CI), 70.7 to 84.8%] (Fig. 1). The first diagnosis suggested was the correct diagnosis in 52% of cases (95% CI, 44 to 61%). When expanding to also include potentially helpful or very close diagnoses, o1-preview was accurate on 97.9% (95% CI, 94.0 to 99.6%) of cases (Fig. 2A). We did not find evidence of a significant difference in performance before and after the pretraining cutoff date for o1-preview (79.8% accuracy before, 73.5% accuracy after, $P = 0.59$; table S1). In a subset of 101 cases from a prior study (*8*), o1-preview outperformed a human physician baseline in both top-1 and top-10 accuracy (table S2). On 70 cases used to evaluate GPT-4 in a prior study (*7*), o1-preview produced a response with the exact or a very close diagnosis in 88.6% of cases, compared with 72.9% of cases by GPT-4 ($P = 0.015$; Fig. 2B). Overall, o1-preview and GPT-4 performed identically on 48/70 (68.6%) of cases, o1-preview outperformed GPT-4 on 17/70 (24.3%) of cases, and GPT-4 outperformed o1-preview on 5/70 (7.1%) of cases (fig. S1).

We next evaluated the ability of o1-preview to select the next diagnostic test in the *NEJM* CPCs for a subset of 136 cases. Two physicians scored the suggested test plan produced by o1-preview [agreement on 113/130 cases (87%), $\kappa = 0.26$] with respect to the actual management of the patient described in the CPC. The proportion of agreements was high, but the κ was low as a result of severe class imbalance. In 87.5% of cases, o1-preview selected the correct test to order; in another 11% of cases, the chosen testing plan was judged by the two physicians to be helpful; and in 1.5% of cases, it would have been unhelpful (Fig. 2C). Inference costs and examples of model outputs are shown in tables S3 to S5.

Presentation of reasoning in *NEJM* Healer diagnostic cases

We used 20 clinical reasoning cases from the *NEJM* Healer curriculum (*15*) that were also evaluated in a prior study using GPT-4 (*16*). *NEJM* Healer cases are virtual patient encounters designed for the assessment of clinical reasoning (*15*). There was substantial agreement of Revised-IDEA (R-IDEA) scores— a validated 10-point scale for evaluating four core domains of documenting clinical reasoning (*17*)—between the two physicians [agreement on 79/80 (99%) cases, $\kappa = 0.89$]. For 78/80 of the cases, o1-preview achieved a perfect R-IDEA score, significantly

[1]Department of Internal Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [3]Stanford Division of Computational Medicine, Stanford University, Stanford, CA, USA. [4]Stanford Clinical Excellence Research Center, Stanford University, Stanford, CA, USA. [5]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [6]Department of Internal Medicine, Cambridge Health Alliance, Cambridge, MA, USA. [7]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA, USA. [8]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA. [9]Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. [10]Department of Hematology-Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [11]Department of Hospital Medicine, University of Minnesota Medical School, Minneapolis, MN, USA. [12]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA. [13]Veterans Affairs Maryland Healthcare System, Baltimore, MD, USA. [14]Center for Innovation to Implementation, VA Palo Alto Health Care System, Palo Alto, CA USA. [15]Division of Hospital Medicine at Zuckerberg San Francisco General Hospital, University of California, San Francisco, San Francisco, CA, USA. [16]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [17]Division of Neurology, University of Alberta, Edmonton, AB, Canada. [18]Technology and Innovation Group, Beth Israel Lahey Health, Boston, MA, USA. [19]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [20]Microsoft, Redmond, WA, USA. [21]Stanford Institute for Human-Centered Artificial Intelligence, Stanford, CA, USA. [22]Division of Hospital Medicine, Department of Medicine, Stanford University, Stanford, CA, USA. ***Corresponding author. Email: arjun_manrai@hms.harvard.edu (A.K.M.); arodman@bidmc.harvard.edu (A.R.)** †These authors contributed equally to this work. ‡These authors contributed equally to this work.

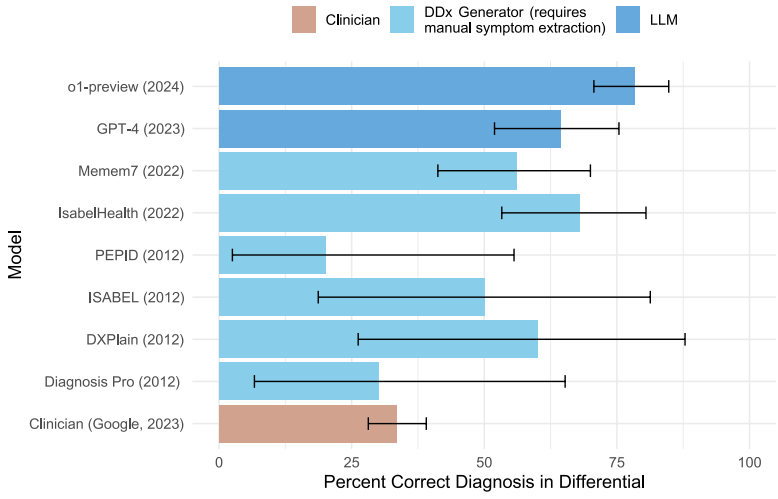

Fig. 1. Performance of differential diagnosis generators and LLMs on *NEJM* clinicopathologic conferences (CPCs), 2012 to 2024. Bar plot showing the accuracy of including the correct diagnosis in the differential for differential diagnosis (DDx) generators and LLMs on the *NEJM* CPCs, sorted by year. Data for other LLMs or DDx generators were obtained from the literature (materials and methods). The 95% CIs were computed with a one-sample binomial test.

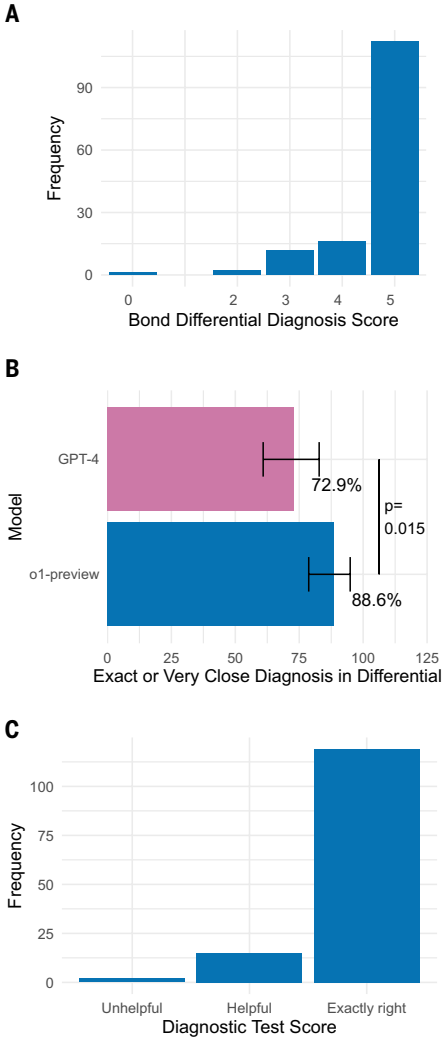

Fig. 2. Quality of differential diagnosis and diagnostic test selection in *NEJM* clinicopathologic conferences (CPCs). (**A**) Histogram of o1-preview performance as measured by the Bond score on the complete set of 143 cases from 2021 to 2024. (**B**) Comparison of o1-preview with a previous evaluation of GPT-4 in providing the exact or very close diagnosis (Bond scores 4 to 5) on the same 70 cases. Bars are annotated with the accuracy of each model; 95% CIs were computed with a one-sample binomial test. The *P* value was computed with McNemar's test. (**C**) Performance of o1-preview in predicting the next diagnostic tests that should be ordered. Performance was measured by two physicians using a Likert scale of "Unhelpful," "Helpful," and "Exactly right." We excluded seven cases from the total case set in which it did not make sense to ask for the next test from the total case set (supplementary text 1C).

outperforming GPT-4 (47/80, $P < 0.0001$), attending physicians (28/80, $P < 0.0001$), and resident physicians (16/72, $P < 0.0001$), as shown in Fig. 3A. We measured the proportion of "cannot-miss" diagnoses identified by o1-preview during the initial triage presentation (Fig. 3B). The median proportion of cannot-miss diagnoses included for o1-preview was 0.92 [interquartile range (IQR) 0.62 to 1.0], although this was not significantly higher than GPT-4, attending physicians, or residents.

Grey Matters management cases

We used five clinical vignettes based on real cases from a previous study developed and scored with consensus methods from 25 physician experts (*18*). Each clinical vignette was presented to the model and was followed by a series of questions regarding next steps in management. Two physicians scored responses by o1-preview for the five cases, with substantial agreement ($\kappa = 0.71$). The median score for the o1-preview per case was 89% (IQR 79 to 91%) (Fig. 4A), which compared favorably with GPT-4 (median 42%, IQR 33 to 52%), physicians with access to GPT-4 (median 41%, IQR 31 to 54%), and physicians with conventional resources (median 34%, IQR 23 to 48%). Using the mixed-effects model, o1-preview scored 41.0 percentage points higher than GPT-4 alone (95% CI, 28.7 to 53.4; $P < 0.001$), 41.9 percentage points higher than physicians with GPT-4 (95% CI, 31.8 to 52.0; $P < 0.001$), and 48.4 percentage points higher than physicians with conventional resources (95% CI, 38.3 to 58.5; $P < 0.001$).

Landmark diagnostic cases

We used six clinical vignettes from a previous study that compared GPT-4 to 50 generalist physicians (*19*). The cases derive from a landmark study of computer-based diagnostic systems, containing the history of present illness, past medical history, physical exam, and diagnostic studies (*20*). The cases have never been publicly released, specifically to protect evaluation validity against memorization. Two physicians scored responses by o1-preview to the six diagnostic reasoning cases, with moderate agreement for total score ($\kappa = 0.42$). The median score for the o1-preview model per case was 97% (IQR 95 to 100%) (Fig. 4B). This is compared with historical control data for which GPT-4 scored 92% (IQR 82 to 97%), physicians with access to GPT-4 scored 76% (IQR 66 to 87%), and physicians with conventional resources scored a median of 74% (IQR 63 to 84%). Using the mixed-effects model, o1-preview performed comparably to GPT-4 (4.4% higher, 95% CI, −19.0 to 27.7%; $P = 0.7$), physicians with GPT-4 (18.6% higher, 95% CI, −2.0 to 39.3%; $P = 0.076$), and physicians with conventional resources (20.2% higher, 95% CI, −0.4 to 40.9%; $P = 0.055$).

Diagnostic probabilistic reasoning cases

We used five cases on primary care topics given to a nationally representative sample of 553 medical practitioners (290 resident physicians, 202 attending physicians, and 61 nurse practitioners or physician assistants) tasked with estimating pretest and posttest probabilities compared with scientific reference probabilities, which were derived from an expert panel who used literature-based estimates to define the ground truth, as detailed in a prior study (*21*). As shown in fig. S2 and table S6, o1-preview performed similarly to GPT-4 in estimating pretest and posttest probabilities, with o1-preview modestly outperforming GPT-4 overall. In general, clinicians display substantially wider variability in estimates than both GPT-4 and o1-preview (fig. S2). Notably, o1-preview substantially outperformed both GPT-4 and human clinicians estimating posttest probabilities for the cardiac ischemia case.

Emergency room cases

We compared the ability of o1, 4o, and two attending physicians to provide differential diagnoses across 76 cases from the Beth Israel Deaconess Medical Center, divided into three clinically meaningful diagnostic touchpoints [initial emergency room (ER) triage, ER physician, and admission to the medical floor or intensive care unit (ICU)]. Overall, o1 outperformed both 4o and two expert attending physicians, as assessed by two other attending physicians who both were blinded to the source of the differential diagnosis (human or AI model) (Fig. 5 and fig. S3); however, both models displayed considerable uncertainty (fig. S4). The physician raters exhibited moderate agreement in quality scores [agreement on 496/911 (54%), $\kappa = 0.51$]. Blinding was successful: Physician accuracy in guessing AI or human was 15.2% for one physician (83.6% "Can't tell") and 3.1% for the other (94.4% "Can't tell"), as displayed in table S7. At each diagnostic touchpoint, o1 either performed nominally better than or on par with the two attending physicians and 4o, with significant differences found at the first two touchpoints (Fig. 5). Performance differences were especially pronounced at the first diagnostic touchpoint (initial ER triage), where there is the least information available about the patient and the most urgency to make the correct decision.

The o1 model identified the exact or very close diagnosis (Bond scores of 4 to 5) in 67.1% of cases during the initial ER triage, 72.4% during the ER physician encounter, and 81.6% at admission to the medical floor or ICU—surpassing the two physicians (55.3, 61.8, and 78.9% for Physician 1; 50.0, 52.6, and 69.7% for Physician 2) at each stage.

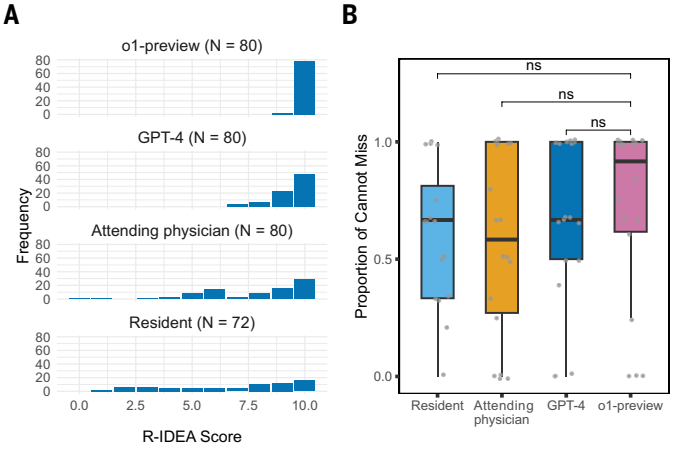

Fig. 3. Comparison of o1-preview, GPT-4, and physicians for clinical diagnostic reasoning. (**A**) Distribution of 312 R-IDEA scores stratified by respondents on 20 *NEJM* Healer cases. (**B**) Box plot of the proportion of cannot-miss diagnoses included in differential diagnosis for the initial triage presentation. The total sample size in this figure is 70, with 18 responses from attending physicians, GPT-4, and o1-preview, and 16 responses from residents. Two cases were excluded because the cannot-miss diagnoses could not be identified. ns, not significant.

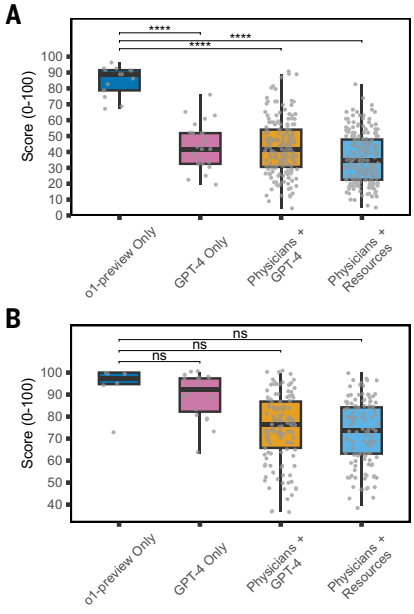

Fig. 4. Comparison of o1-preview, GPT-4, and physicians for management and diagnostic reasoning. (**A**) Box plot of normalized management reasoning points by LLMs and physicians on *Grey Matters* Management Cases. Five cases were included. We generated three o1-preview responses for each case. The prior study collected five GPT-4 responses to each case, 178 completed cases from 46 physicians with access to GPT-4, and 197 completed cases from 46 physicians with access to conventional resources. *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$. (**B**) Box plot of normalized diagnostic reasoning points by LLMs and physicians. Six diagnostic challenges were included. We generated one o1-preview response for each case. The prior study collected three GPT-4 responses to all cases, 125 cases completed by 25 physicians with access to GPT-4, and 119 cases completed by 25 physicians with access to conventional resources.

Discussion

We systematically evaluated the medical reasoning abilities of an LLM across six diverse experiments, comparing the model with hundreds of expert physicians. Overall, the model outperformed physicians across experiments, including in cases utilizing real and unstructured clinical data taken directly from the health record in an emergency department. These diagnostic touchpoints mirror the high-stakes decisions taken in emergency medicine departments, where nurses and clinicians make time-sensitive choices with limited information. Our results showed that humans, GPT-4o, and o1 all improved their diagnostic abilities as more information was available; o1 outperformed humans at multiple touchpoints, with the widest gap at initial ER triage, where there is the least information available.

The rapid pace of improvement in LLMs has substantial implications for the science and practice of clinical medicine. Although applying AI to assist with clinical decision support is sometimes viewed as a high-risk endeavor (*22*, *23*), greater use of these tools might serve to mitigate the human and financial costs of diagnostic error, delay, and lack of access (*24*, *25*). Our findings suggest the urgent need for prospective trials to evaluate these technologies in real-world patient care settings and for health care systems to prepare for investments for computing infrastructure and design for clinician-AI interaction that can facilitate the safe integration of AI tools into patient-care workflows. This includes the development of robust monitoring frameworks to oversee the broader implementation of AI clinical decision support systems (*22*), monitoring not just final diagnostic accuracy but other metrics crucial for successful deployment, including safety, efficiency, and cost.

We emphasize that our study addresses only text-based performance for both humans and machines; clinical medicine is multifaceted and awash with nontext inputs, including auditory (such as the patient's level of distress) and visual information (for example, interpretation of medical imaging studies) that clinicians routinely use. Existing studies suggest that current foundation models are more limited in reasoning over nontext inputs (*26*, *27*); future work is needed to assess how humans and machines may effectively collaborate (*28*) in use of nontext signals. This requires new benchmarks, trials, and technological solutions to more faithfully measure clinical encounters. Existing investment in increasingly pervasive ambient AI scribes and other passive monitoring technologies holds promise to serve as the basis for such investigations. Such studies may help in further investigating an important limitation of existing benchmarks of medical AI, including several studied here, namely their reliance on the careful work of clinicians to curate and "clean up" cases and on questions originally developed for educational purposes, which may therefore overstate the performance of AI models when using "messy" data available in more realistic clinical workflows (*13*).

Our study has several limitations. First, whereas some of the experiments were originally performed with human-computer interaction, our current study reflects only model performance and primarily focuses on the preview version of the o1 model, which has since been

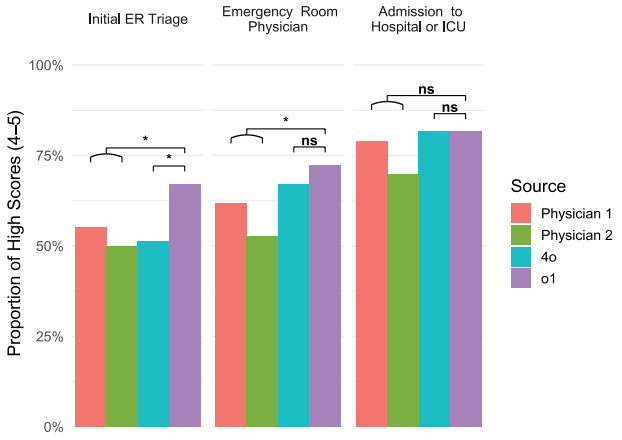

Fig. 5. Blinded assessment of AI and human expert second opinions on real ER cases. Bar plot comparing two internal medicine attending physicians, o1, and GPT-4o diagnostic performance on 76 clinical cases at three diagnostic touchpoints (triage in the ER, initial evaluation by a physician, and admission to the hospital or ICU). Differential diagnoses were capped at five diagnoses for all participants. The source of the differential diagnosis was blinded and scored by two separate attending internal medicine physicians using the Bond scale. The proportion of responses scored 4 or 5 are shown, indicating a response that contains something exact or very close to the true diagnosis. *P* values were computed with a mixed-effects logistic regression model (materials and methods). *$P \leq 0.05$.

supplanted by newer models (for example, OpenAI's o3 model). Although we expect performance to be sustained or improved with newer models (*27*, *29*), further studies should be done to elucidate how performance varies across models and to study how humans and LLMs may collaborate. Second, our study examined only six aspects of clinical reasoning; researchers have identified dozens of other tasks that could be studied that may have even more impact on actual clinical care (*30*). Third, despite large numbers and varieties of cases included in our study, which were focused on internal medicine and emergency medicine, it is not representative of broader medical practice, which includes multiple specialties that require varying skill sets, such as decisions related to surgery. Performance may vary according to diagnoses, patient characteristics, or practice locations that were not interrogated in this study. Fourth, although the results of our emergency department experiment have face validity, the task we studied, namely providing a second opinion at predefined touchpoints, is best thought of as a proof of concept. Decisions in the emergency department are often centered around triage, disposition, and immediate management and not diagnostic accuracy. Lastly, we did not always find robust improvements in o1 performance compared with previous models, for example in the crucial cannot-miss diagnoses in the *NEJM* Healer cases and in the landmark diagnostic cases.

Overall, our findings show that LLMs now demonstrate substantial performance in differential diagnosis, diagnostic clinical reasoning, and management reasoning, and exceed both prior model generations and even human clinicians across multiple domains. These same performance gains are seen in providing second opinions in real, unstructured medical cases in the emergency department, where clinicians must act quickly with limited and often missing information.

More than 65 years ago, Ledley and Lusted described the standard for evaluating the diagnostic abilities of AI (*1*). The broad challenge they laid out of reasoning over complex clinical case vignettes has guided the development and evaluation of computational systems for much of the past century. Our findings suggest that LLMs have now eclipsed most benchmarks of clinical reasoning, motivating the urgent need for human-computer interaction studies and prospective clinical trials to rigorously assess the potential of AI systems to improve clinical practice and patient outcomes.

REFERENCES AND NOTES

1. R. S. Ledley, L. B. Lusted, *Science* **130**, 9–21 (1959).
2. K. Brodman, A. J. Erdmann Jr., I. Lorge, C. P. Gershenson, H. G. Wolff, *J. Clin. Psychol.* **8**, 119–124 (1952).
3. F. T. de Dombal, D. J. Leaper, J. R. Staniland, A. P. McCann, J. C. Horrocks, *BMJ* **2**, 9–13 (1972).
4. E. Shortliffe, in *Proceedings: Symposium on Computer Applications in Medical Care*, Washington, DC, 3 to 5 October 1977 (Hanley & Belfus, 1977), pp. 66–69.
5. E. B. Ing, M. Balas, G. Nassrallah, D. DeAngelis, N. Nijhawan, *Ophthalmic Plast. Reconstr. Surg.* **39**, 461–464 (2023).
6. R. A. Miller, H. E. Pople Jr., J. D. Myers, *N. Engl. J. Med.* **307**, 468–476 (1982).
7. Z. Kanjee, B. Crowe, A. Rodman, *JAMA* **330**, 78–80 (2023).
8. D. McDuff *et al.*, *Nature* **642**, 451–457 (2025).
9. H. Nori *et al.*, arXiv:2411.03590 [cs.CL] (2024).
10. OpenAI, arXiv:2412.16720 [cs.AI] (2024).
11. P. Lee, S. Bubeck, J. Petro, *N. Engl. J. Med.* **388**, 1233–1239 (2023).
12. S. Bubeck *et al.*, arXiv:2303.12712 [cs.CL] (2023).
13. S. Johri *et al.*, *Nat. Med.* **31**, 77–86 (2025).
14. A. Rodman, L. Zwaan, A. Olson, A. K. Manrai, *NEJM AI* **2**, AIe2500143 (2025).
15. R.-E. E. Abdulnour *et al.*, *N. Engl. J. Med.* **386**, 1946–1947 (2022).
16. S. Cabral *et al.*, *JAMA Intern. Med.* **184**, 581–583 (2024).
17. V. Schaye *et al.*, *J. Gen. Intern. Med.* **37**, 507–512 (2022).
18. E. Goh *et al.*, *Nat. Med.* **31**, 1233–1238 (2025).
19. E. Goh *et al.*, *JAMA Netw. Open* **7**, e2440969 (2024).
20. E. S. Berner *et al.*, *N. Engl. J. Med.* **330**, 1792–1796 (1994).
21. D. J. Morgan *et al.*, *JAMA Intern. Med.* **181**, 747–755 (2021).
22. R. M. Ratwani, D. W. Bates, D. C. Classen, *JAMA Health Forum* **5**, e235514 (2024).
23. Q. Jin *et al.*, *NPJ Digit. Med.* **7**, 190 (2024).
24. D. E. Newman-Toker *et al.*, Agency for Healthcare Research and Quality, US Department of Health and Human Services, "Diagnostic errors in the emergency department: A systematic review" (AHRQ, 2022); https://effectivehealthcare.ahrq.gov/products/diagnostic-errors-emergency-updated/research.
25. A. D. Auerbach *et al.*, *JAMA Intern. Med.* **184**, 164–173 (2024).
26. T. Buckley, J. A. Diao, P. Rajpurkar, A. Rodman, A. K. Manrai, arXiv:2311.05591 [cs.CV] (2024).
27. T. A. Buckley *et al.*, arXiv:2509.12194 [cs.AI] (2025).
28. F. Yu *et al.*, *Nat. Med.* **30**, 837–849 (2024).
29. G. Dhaliwal *et al.*, *N. Engl. J. Med.* **393**, 1421–1434 (2025).
30. M. Goldszmidt, J. P. Minda, G. Bordage, *Acad. Med.* **88**, 390–397 (2013).
31. T. Buckley, 2v/llm-physician-tasks: v1.0.0, Zenodo (2026); https://doi.org/10.5281/zenodo.18292046.

ACKNOWLEDGMENTS

We thank the NEJM Group for permission to use the CPCs and Healer cases. **Funding:** We gratefully acknowledge support from NIH/NIEHS award R01ES032470 (A.K.M.), the Harvard Medical School Dean's Innovation Award for Artificial Intelligence (A. K. M.), Macy Foundation awards B25-15 and P25-04 (A.R. and J.C.), Moore Foundation award 12409 (A.R., J.C., and Z.K.), NIH/NIAID 1R01AI17812101 (J.C.), NIH-NCATS UM1TR004921 (J.C.), the Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (J.C.), NIH U01 NS134358 (J.C.), and a Stanford RAISE Health Seed Grant 2024 (J.C.). **Author contributions:** P.G.B., T.A.B., A.K.M., and A.R. conceived the study and conducted the analyses; A.K.M. and A.R. obtained funding and supervised the work; and all authors contributed to interpretation and writing of the manuscript. **Competing interests:** A.R. is a Visiting Researcher at Google DeepMind. E.H. is employed by Microsoft. J.C. is cofounder of Reaction Explorer LLC, a paid medical expert witness from Elite Experts, and received one-time honoraria or travel expenses for invited presentations by Insitro, General Reinsurance Corporation, AASCIF, and other industry conferences, academic institutions, and health systems. Z.K. discloses royalties from Oakstone Publishing and Wolters Kluwer. A.O. discloses employment of his spouse by Exact Sciences. R.-E.A. is employed by the Massachusetts Medical Society and has consulted for Lumeris. The other authors declare that they have no competing interests. **Data, code, and materials availability:** Analysis code and rubrics are available at Zenodo, a public repository (*31*). For case data from *NEJM*, please contact NEJM Group (permissions@nejm.org). For access to the *Grey Matters* management cases and landmark diagnostic cases, please contact AR (arodman@bidmc.harvard.edu). The probabilistic reasoning cases are publicly available in the supplement of a prior publication (*21*). Internal patient data used in our study of ER cases cannot be made publicly available because of patient privacy and data use restrictions. No new materials were generated in this study.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adz4433
Materials and Methods; Supplementary Text; Figs. S1 to S4; Tables S1 to S7; References (*32*–*35*); MDAR Reproducibility Checklist

Submitted 2 June 2025; accepted 23 February 2026

10.1126/science.adz4433

Performance of a large language model on the reasoning tasks of a physician

Peter G. Brodeur, Thomas A. Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdulnour, Adrian D. Haimovich, Jason A. Freed, Andrew Olson, Daniel J. Morgan, Jason Hom, Robert Gallo, Liam G. McCoy, Haadi Mombini, Christopher Lucas, Misha Fotoohi, Matthew Gwiazdon, Daniele Restifo, Daniel Restrepo, Eric Horvitz, Jonathan Chen, Arjun K. Manrai, and Adam Rodman

Science **392** (6797), . DOI: 10.1126/science.adz4433

Editor's summary

Computational tools for medical decision support have been advancing over time, mainly by serving as resources for limited applications. Machine learning tools for autonomous interpretation of clinical cases have also been gradually improving over time. Brodeur *et al*. pitted a large language model, the OpenAI o1 series, directly against hundreds of physicians at different levels of training and experience on a variety of clinical cases ranging from published patient vignettes to evaluations of brand-new emergency room patients, as well as on clinical tasks including both diagnosis and planning of clinical management (see the Perspective by Hopkins and Cornelisse). Across a variety of scenarios and applications, the large language model outperformed both human physicians and older models, suggesting its potential utility for clinical care. —Yevgeniya Nusinovich