nature medicine

Article

https://doi.org/10.1038/s41591-024-03456-y

GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial

Received: 5 August 2024

Accepted: 10 December 2024

Published online: 5 February 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

While large language models (LLMs) have shown promise in diagnostic reasoning, their impact on management reasoning, which involves balancing treatment decisions and testing strategies while managing risk, is unknown. This prospective, randomized, controlled trial assessed whether LLM assistance improves physician performance on open-ended management reasoning tasks compared to conventional resources. From November 2023 to April 2024, 92 practicing physicians were randomized to use either GPT-4 plus conventional resources or conventional resources alone to answer five expert-developed clinical vignettes in a simulated setting. All cases were based on real, de-identified patient encounters, with information revealed sequentially to mirror the nature of clinical environments. The primary outcome was the difference in total score between groups on expert-developed scoring rubrics. Secondary outcomes included domain-specific scores and time spent per case. Physicians using the LLM scored significantly higher compared to those using conventional resources (mean difference = 6.5%, 95% confidence interval (CI) = 2.7 to 10.2, $P < 0.001$). LLM users spent more time per case (mean difference = 119.3 s, 95% CI = 17.4 to 221.2, $P = 0.02$). There was no significant difference between LLM-augmented physicians and LLM alone (−0.9%, 95% CI = −9.0 to 7.2, $P = 0.8$). LLM assistance can improve physician management reasoning in complex clinical vignettes compared to conventional resources and should be validated in real clinical practice. ClinicalTrials.gov registration: NCT06208423.

Large language models (LLMs) show considerable abilities in diagnostic reasoning, outperforming previous artificial intelligence (AI) models and human physicians in their ability to construct helpful differential diagnoses, explain reasoning and collect historical information from standardized patients[1–5]. LLMs have not yet been shown to perform similarly in management reasoning, which encompasses decision-making around treatment, testing, patient preferences, social determinants of health and cost-conscious care, all while managing risk[6–8].

While there is overlap, clinical reasoning is often considered to include both diagnostic and management reasoning. The study of diagnostic reasoning has a century-long history with many metacognitive frameworks and assessment methods, while management reasoning processes are a comparatively recent area of study[9–11]. Current frameworks in management reasoning include context-dependent concepts, such as shared decision-making, dynamic relationships and competing priorities between medical systems and individuals, the physician–patient relationship and time constraints inherent in modern clinical encounters[8,12–14]. Unlike diagnostic reasoning, which can be thought of as a classification task with often a single right answer, management reasoning may have no right answers and involves weighing trade-offs between inherently risky courses of action; even inaction through 'watchful waiting' is a deliberate choice with potential risks

e-mail: jonc101@stanford.edu

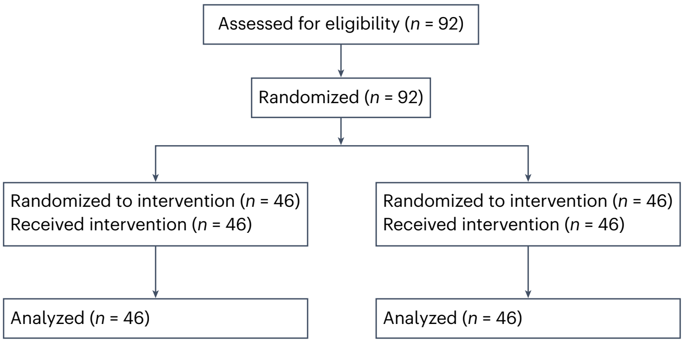

Assessed for eligibility (n = 92)
Randomized (n = 92)
Randomized to intervention (n = 46)
Received intervention (n = 46)
Randomized to intervention (n = 46)
Received intervention (n = 46)
Analyzed (n = 46)
Analyzed (n = 46)

Fig. 1 | Study flow diagram. The study included 92 practicing attending physicians and residents with training in internal medicine, family medicine or emergency medicine. Five expert-developed cases were presented, with scoring rubrics created using a Delphi process. Physicians were randomized to use either GPT-4 via ChatGPT plus in addition to conventional resources (for example, UpToDate, Google), or conventional resources alone. The primary outcome was the difference in total score between groups on expert-developed scoring rubrics. Secondary outcomes included domain-specific scores and time spent per case.

and benefits. From a cognitive psychology perspective, management reasoning often uses heuristics called management scripts, which allow clinicians to quickly make decisions[15]. However, these scripts are susceptible to the same fallibilities that affect other domains of human reasoning. With few exceptions, these scripts must be adapted to specific situations to balance all factors that influence management reasoning, as well as continually updated with new and emerging information.

Previous generations of non-LLM AI systems can improve human management decisions in some situations, especially when a human user treats an AI suggestion as a second opinion[16]. One of the theoretical strengths of LLMs is their ability to serve as a cooperation partner, augmenting human cognition[17]. LLMs may offer differing points of view that would assist in aligning patients' and clinicians' values and goals into a cohesive plan. We designed a prospective, randomized, controlled trial to assess whether physicians using an LLM performed better than physicians using standard resources on a series of complex clinical management questions. We then compared physician answers to the output of the LLM (without a human) alone. All cases were derived from real, de-identified patient encounters. Rather than presenting complete case information upfront, physicians received information sequentially to mirror the complex nature of clinical progression. This design choice enabled physicians to iteratively formulate and adjust their management plan as new data emerged.

Results

We enrolled 92 physicians to participate in the study, which was conducted from 30 November 2023 to 21 April 2024. Participants were randomized evenly between the LLM and conventional resources groups (Fig. 1); 73% (67 of 92) were attending physicians while 27% (25 of 92) were residents (Table 1). Seventy-four percent (68 of 92) specialized in internal medicine, 20% (18 of 92) emergency medicine and 6.5% (6 of 92) family medicine. The mean time in practice of all physicians was 7.6 years while the median was 5.8 years (interquartile range (IQR) = 3.0 to 9.0 years). Only 24% (22 of 92) self-described as frequent users of LLMs; 20.8% (19 of 92) had either used it only once or never used it.

From these 92 physicians, 400 cases were scored in total, 176 from the group of physicians using the LLM, 199 from physicians using only conventional resources and 25 from the LLM alone. Three graders agreed on the scoring of 328 of 400 cases (82%), with a pooled kappa statistic (κ) of 0.80, reflecting substantial agreement between graders (case 1 $\kappa = 0.58$, case 2 $\kappa = 0.83$, case 3 $\kappa = 0.82$, case 4 $\kappa = 0.90$, case 5 $\kappa = 0.89$) (Supplementary Item 1 for an example of case, rubric and a high scoring and low scoring participant case response).

Management performance

Physicians randomized to use the LLM performed better than the control group (43.0% compared to 35.7%, difference = 6.5%, 95% confidence interval (CI) = 2.7% to 10.2%, $P < 0.001$) (Table 2 and Fig. 2). The LLM alone scored comparably to humans using the LLM (43.7% versus 43.0%, difference = 0.9%, 95% CI = −7.2% to 9.0%, $P = 0.80$), while trending toward scoring higher than humans using conventional resources (43.7% versus 35.7%, difference = 7.3%, 95% CI = −0.7% to 15.4%, $P = 0.074$) (Fig. 3). We ran additional post hoc sensitivity analyses using repeated measures analysis of variance (Supplementary Item 5), which showed similar results to our primary analysis using mixed-effects models.

Question domain subgroups

The physicians using the LLM scored better than those using conventional resources alone in questions explicitly testing management decisions (40.5% versus 33.4%, difference = 6.1%, 95% CI = 2.5% to 9.7%, $P = 0.001$), questions testing diagnostic decisions (56.8% versus 45.8%, difference = 12.1%, 95% CI = 3.1% to 21%, $P = 0.009$) and context-specific questions (42.4% versus 34.9%, difference = 6.2%, 95% CI = 2.4% to 9.9%, $P = 0.002$). While we did not detect a difference in factual recall between the two groups (62.9% versus 53.8%, difference = 9.6%, 95% CI = −3.1% to 22.3%, $P = 0.14$) and general management knowledge (29.4% versus 26.5%, difference = 3.3, 95% CI = −1.3% to 7.9%, $P = 0.2$), they were directionally similar to the other subdomains.

Time

Physicians randomized to use the LLM spent 111.3 s more on each case (801.5 s versus 690.2 s, difference = 119.3 s, 95% CI = 17.4 to 221.2, $P = 0.022$) (Fig. 4). We performed an additional post hoc sensitivity analysis adjusting for time spent on each case (Extended Data Table 1), which showed a 5.4 percentage point (95% CI = 1.7 to 9.0, $P = 0.004$) increase in score per case even after adjustment for time spent on the case. Results were similar for subdomains. We further examined the unadjusted correlation between time spent and total scores with a positive association between time spent and total scores for both groups (Extended Data Table 2). Overall, we observed that for each additional minute spent on a case, there was a small but statistically significant increase of 0.6 points in the score per case (95% CI = 0.4 to 0.8, $P < 0.001$) using a mixed-effects model (Extended Data Fig. 1).

Response length

To address the potential influence of response length on scores, we conducted an additional post hoc sensitivity analysis adjusting our primary analysis for the character count of responses (Supplementary Item 3). This analysis revealed an attenuated but still positive effect, with the LLM group scoring 3.7 percentage points higher (95% CI = 0.7 to 6.7, $P = 0.02$). Notably, while longer responses tended to score higher (approximately 0.3 points per 100 characters), the LLM intervention arm outperformed the conventional resources arm even after this adjustment.

Likelihood and extent of harm

Analysis of potential harm revealed similar patterns between groups (Supplementary Item 6). In the LLM-assisted group, 8.5% and 4.2% of physician responses carried medium and high likelihood of harm, respectively, compared to 11.4% and 2.9% in the conventional resources group. Regarding harm severity, mild-to-moderate harm was observed in 4.0% of LLM-assisted responses compared to 5.3% in the conventional resources group. Severe harm ratings were nearly identical between groups (LLM = 7.7%; conventional = 7.5%).

Table 1 | Participant characteristics according to randomized group

Variable	Overall, *n*=92	Physicians+LLM, *n*=46	Physicians+conventional resources only, *n*=46	SMD
Career stage				0.05
Attending	67 (73%)	34 (74%)	33 (72%)	
Resident	25 (27%)	12 (26%)	13 (28%)	
Specialty				0.22
Internal medicine	68 (74%)	36 (78%)	32 (70%)	
Emergency medicine	18 (20%)	8 (17%)	10 (22%)	
Family medicine	6 (6.5%)	2 (4.3%)	4 (8.7%)	
Years in medical training				−0.02
Mean (s.d.)	7.6 (7.1)	7.6 (7.9)	7.7 (6.3)	
Median (IQR)	5.8 (3.0 to 9.0)	5.0 (3.1 to 8.8)	6.0 (3.0 to 9.8)	
Past GPT experience				0.11
I use it frequently (weekly or more)	22 (24%)	11 (24%)	11 (24%)	
I use it occasionally (more than once per month but less than weekly)	28 (30%)	15 (33%)	13 (28%)	
I use it rarely (less than once per month)	23 (25%)	11 (24%)	12 (26%)	
I've used it once ever	9 (9.8%)	4 (8.7%)	5 (11%)	
I've never used it before	10 (11%)	5 (11%)	5 (11%)	

SMD, standardized mean difference.

Table 2 | Comparisons of the primary and secondary outcomes for physicians with LLM and with conventional resources only (scores standardized to 0–100)

Outcomes	Physicians+LLM, *n*=178	Physicians+conventional resources only, *n*=197	Difference between physicians+GPT-4 and physicians+conventional resources only
Primary outcome			
Total score (*n*)	178	197	6.5 (2.7 to 10.2), *P*<0.001
Mean (s.d.)	43.0 (17.3)	35.7 (15.5)	
Median (IQR)	41.3 (30.6 to 54.1)	34.4 (22.5 to 47.8)	
Secondary outcomes			
Management (*n*)	178	197	6.1 (2.5 to 9.7), *P*=0.001
Mean (s.d.)	40.5 (19.1)	33.4 (17.3)	
Median (IQR)	37.5 (26.8 to 52.4)	30.0 (19.3 to 45.5)	
Factual (*n*)	69	78	9.6 (−3.1 to 22.3), *P*=0.14
Mean (s.d.)	62.9 (37.6)	53.8 (39.6)	
Median (IQR)	75.0 (37.5 to 100.0)	56.2 (15.6 to 100.0)	
Diagnostic (*n*)	72	77	12.1 (3.1 to 21.0), *P*=0.009
Mean (s.d.)	56.8 (37.6)	45.8 (26.7)	
Median (IQR)	66.7 (29.2 to 83.3)	50.0 (33.3 to 66.7)	
Specific (*n*)	178	197	6.2 (2.4 to 9.9), *P*=0.002
Mean (s.d.)	42.4 (20.2)	34.9 (17.9)	
Median (IQR)	42.6 (28.1 to 57.4)	35.2 (20.8 to 48.5)	
General (*n*)	70	80	3.3 (−1.3 to 7.9), *P*=0.2
Mean (s.d.)	29.4 (15.0)	26.5 (13.0)	
Median (IQR)	27.3 (18.2 to 39.8)	24.6 (17.5 to 33.3)	
Time spent per case (s)	178	197	119.3 (17.4 to 221.2), *P*=0.022
Mean (s.d.)	801.5 (417.2)	690.2 (372.4)	
Median (IQR)	719.8 (514.6 to 1,010.2)	570.9 (452.9 to 814.9)	

Estimated differences were derived from a generalized mixed-effects model with random effects for participant and case. Reported *P* values are two-sided.

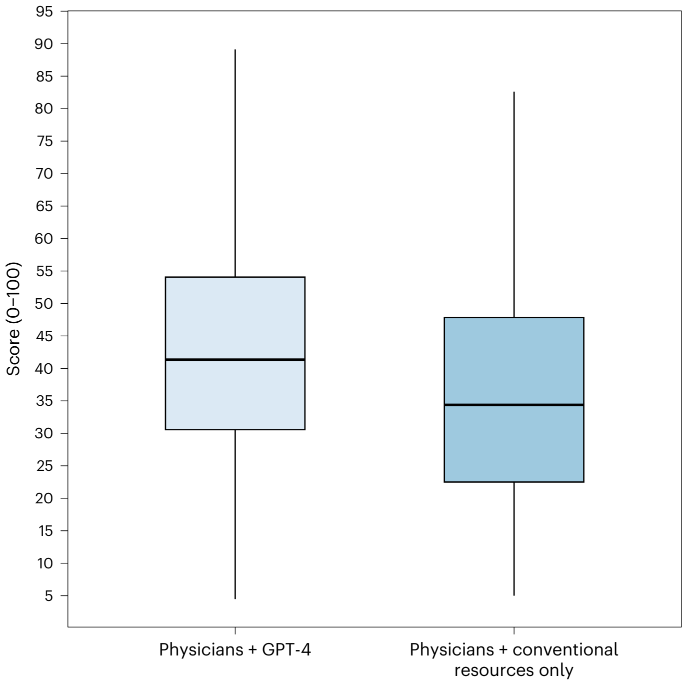


Fig. 2 | Comparison of the primary outcome for physicians with LLM and with conventional resources only (total score standardized to 0–100). Ninety-two physicians (46 randomized to the LLM group and 46 randomized to conventional resources) completed 375 cases (178 in the LLM group, 197 in the conventional resources group). The center of the box plot represents the median, with the boundaries representing the first and third quartiles. The whiskers represent the furthest data points from the center within 1.5 times the IQR.

Discussion

In this randomized, controlled trial, the availability of an LLM improved physician management reasoning compared to conventional resources only, with comparable scores between physicians randomized to use AI and AI alone. This suggests a future use for LLMs as a helpful adjunct to clinician judgment, while also highlighting the potential for standalone LLM applications in certain clinical scenarios. Delineating specific contexts where LLM assistance provides added value to physicians versus areas where AI might be useful independently is becoming increasingly important. From a cognitive psychology perspective, it is surprising that an LLM would enhance management reasoning. The abilities of LLMs to make diagnostic decisions probably result from their underlying token prediction architecture and its similarities to how physicians cluster and activate semantic illness scripts in making diagnoses[18]. Management scripts, on the other hand, are highly contextual and individualized, and include many factors outside the biomedical encounter. Thus, the best decision for a patient in a given situation may be different than another patient with the same condition in a different context. For example, the appropriate management of an incidentally discovered 2.0-cm upper lobe lung nodule in a hospitalized inpatient might be immediate biopsy in a patient unlikely to follow up; scheduled outpatient biopsy in a health system capable of organizing and ensuring continuity; outpatient positron emission tomography scan in a patient reticent to undergo an invasive procedure; or serial imaging in a patient with limited life expectancy. The knowledge that such a large nodule in the upper lobe has a high chance of representing malignancy is only the first step in formulating a follow-up plan—patient preferences, knowledge of the healthcare system and the patient's social situation are similarly important factors.

The group using the LLM spent more time solving cases, a finding that aligns with historical studies of diagnostic support systems[19,20], but contrasts with recent findings of LLM use in diagnostic reasoning[4,5]. While this increased time may be due to the combined effects of case problem-solving and LLM interaction, engaging with the LLM may have served as a beneficial 'time out' to better consider the patient context. For example, we observed that physicians using the LLM exhibited apparent empathy to other providers and patients in difficult situations more frequently. We suspect that some of these emergent abilities come from the fine-tuning process called reinforcement learning through human feedback, in which empathetic and patient-centered responses are rated as favorable by humans[21]. Similar to studies showing increased empathetic communication phrasing from LLMs to patient queries, this study indicates that LLMs may influence physicians to better consider human factors in their management reasoning[22–24]. Improved humanistic and patient-centered behaviors of clinicians when they collaborate with an LLM is an important and even reassuring finding, even at the expense of taking more time.

Of note, the persistent advantage of the LLM group after adjusting for time spent per case and response length (Supplementary Item 4) suggests that the improved performance cannot be attributed solely to these factors. Further exploration into whether the LLM is merely encouraging users to slow down and reflect more deeply, or whether it is actively augmenting the reasoning process, would be valuable. While our findings suggest a combination of both influences, future studies could control for this variable more explicitly by introducing a group prompted to pause and consider alternate factors without LLM support, as well as evaluating more systematically how users directly interact with LLMs.

This study has multiple limitations. First, the cases are clinical vignettes, based on, but not actual patient cases. While our scoring rubrics show substantial interrater reliability, validity evidence for

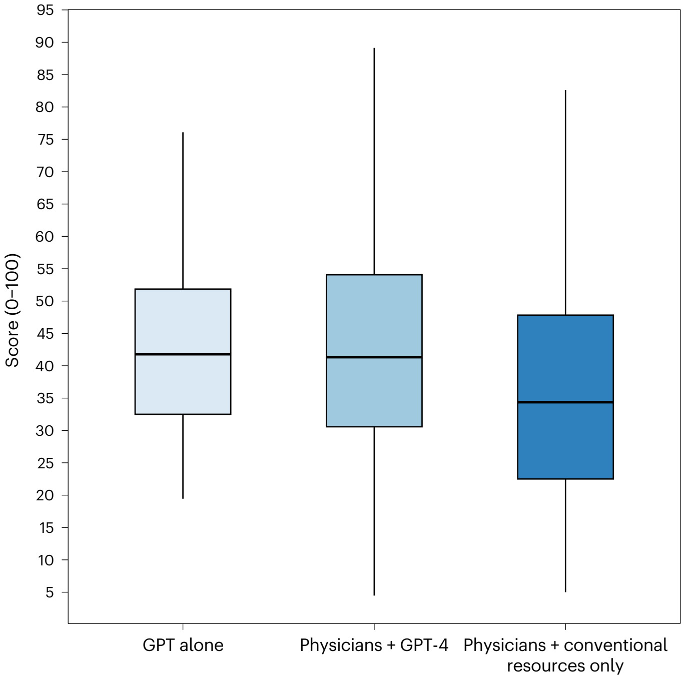


Fig. 3 | Comparison of the primary outcome according to GPT alone versus physician with GPT-4 and with conventional resources only (total score standardized to 0–100). The GPT-alone arm represents the model being prompted by the study team to complete the five cases, with the models prompted five times for each case for a total of 25 observations. The physicians with GPT-4 group included 46 participants that completed 178 cases, while the physician with conventional resources group included 46 participants that completed 197 cases. The center of the box plot represents the median, with the boundaries representing the first and third quartiles. The whiskers represent the furthest data points from the center within 1.5 times the IQR.

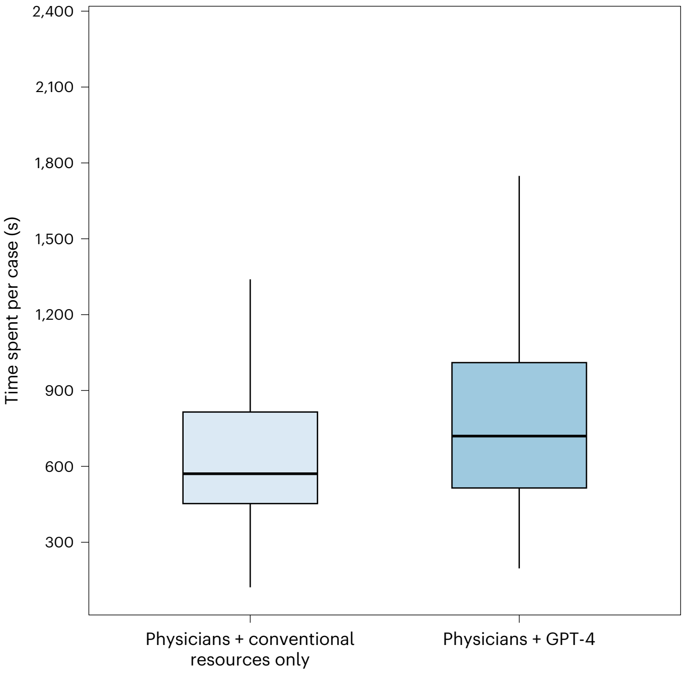


Fig. 4 | Comparison of the time spent per case by physicians using GPT-4 and physicians using conventional resources only. Ninety-two physicians (46 randomized to the LLM group and 46 randomized to the conventional resources) completed 375 cases (178 in the LLM group, 197 in the conventional resources group). The center of the boxplot represents the median, with the boundaries representing the first and third quartiles. The whiskers represent the furthest data points from the center within 1.5 times the IQR.

these rubrics has not been gathered outside this study. Only potentially correct answers were given credit, while wrong answers were not penalized. This approach, which is consistent with many standardized assessments of clinical reasoning, such as Step 2 Clinical Skills and the UK Practical Assessment of Clinical Examination Skills exam, was chosen to facilitate results interpretation and to focus on rewarding appropriate clinical decision-making rather than penalizing errors. While this method does not capture the potential harm of incorrect decisions, an exploratory secondary analysis suggests that there was little to no difference in either the likelihood of harm or the extent of harm when an LLM was available to physicians. The real-world implementation of LLMs in clinical settings necessitates careful consideration of how potential hallucinations and misinformation could impact patient care. Specifically, real-world LLM deployment may require physicians to serve as the sole reliable backstop for misinformation, which could affect both cognitive load and decision-making quality. Additionally, we acknowledge the inherent challenge of distinguishing between accuracy and thoroughness in responses. Our rubric design, informed by expert consensus through a modified Delphi process, attempted to balance these factors by setting maximum point thresholds for each question and rewarding the appropriateness rather than exhaustiveness of responses. Continued refinement of assessment tools may further enhance the ability to differentiate between these aspects of clinical reasoning.

With only five cases expected for participants to complete in a 1-h session, we intentionally selected content to represent a breadth of general medicine, in line with standardized evaluations such as objective structured clinical examinations. A wider variety of cases could show different outcomes. Finally, we provided only basic training on the use of LLMs to either group as well as technical support. While evidence suggests that prompting strategies can dramatically improve model performance on medical tasks, we intentionally chose to mimic current strategies around LLM deployment in healthcare settings, which have been provided with minimal formal training on prompting strategies[25,26].

This study found that the addition of LLM AI assistance improved physician management reasoning compared to conventional resources. Early implementation of LLMs into healthcare has largely been directed at clerical clinical workflows, including portal messaging and ambient listening. Our findings demonstrate that decision support—even in a task as complex as management reasoning—represents a promising application of LLMs that requires rigorous validation in real clinical settings to realize its potential for enhancing patient care.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03456-y.

References

1. Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* **330**, 78–80 (2023).
2. Cabral, S. et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Intern. Med.* **184**, 581–583 (2024).
3. Tu, T. et al. Towards conversational diagnostic AI. Preprint at https://arxiv.org/abs/2401.05654 (2024).
4. McDuff, D. et al. Towards accurate differential diagnosis with large language models. Preprint at https://arxiv.org/abs/2312.00164 (2023).
5. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* **7**, e2440969 (2024).
6. Zaboli, A., Brigo, F., Sibilio, S., Mian, M. & Turcato, G. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? *Am. J. Emerg. Med.* **79**, 44–47 (2024).
7. Truhn, D. et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci. Rep.* **13**, 20159 (2023).
8. Cook, D. A., Sherbino, J. & Durning, S. J. Management reasoning beyond the diagnosis. *JAMA* **319**, 2267–2268 (2018).
9. Ledley, R. S. & Lusted, L. B. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science* **130**, 9–21 (1959).
10. Bordage, G. Prototypes and semantic qualifiers: from past to present. *Med. Educ.* **41**, 1117–1121 (2007).
11. Bowen, J. L. Education educational strategies to promote clinical diagnostic reasoning. *N. Engl. J. Med.* **355**, 2217–2225 (2006).
12. Cook, D. A., Stephenson, C. R., Gruppen, L. D. & Durning, S. J. Management reasoning: empirical determination of key features and a conceptual model. *Acad. Med.* **98**, 80–87 (2023).
13. Mercuri, M. et al. When guidelines don't guide: the effect of patient context on management decisions based on clinical practice guidelines. *Acad. Med.* **90**, 191–196 (2015).
14. Schmidt, H. G., Norman, G. R., Mamede, S. & Magzoub, M. The influence of context on diagnostic reasoning: a narrative synthesis of experimental findings. *J. Eval. Clin. Pract.* **30**, 1091–1101 (2024).
15. Parsons, A. S., Wijesekera, T. P. & Rencic, J. J. The management script: a practical tool for teaching management reasoning. *Acad. Med.* **95**, 1179–1185 (2020).
16. Reverberi, C. et al. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.* **12**, 14952 (2022).

17. Kempt, H. & Nagel, S. K. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. *J. Med. Ethics* **48**, 222–229 (2022).
18. Restrepo, D., Rodman, A. & Abdulnour, R.-E. Conversations on reasoning: large language models in diagnosis. *J. Hosp. Med.* **19**, 731–735 (2024).
19. Friedman, C. P. et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. *JAMA* **282**, 1851–1856 (1999); erratum **285**, 2979 (2001).
20. Miller, R. A., Pople, H. E. Jr & Myers, J. D. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *N. Engl. J. Med.* **307**, 468–476 (1982).
21. Chen, Y. et al. SoulChat: improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. Preprint at https://arxiv.org/abs/2311.00273 (2023).
22. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
23. Tai-Seale, M. et al. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw. Open* **7**, e246565 (2024).
24. Chen, S. et al. The effect of using a large language model to respond to patient messages. *Lancet Digit. Health* **6**, e379–e381 (2024).
25. Pfeffer, M. A., Shah, N. H., Sharp, C. & Lindmark, C. Nigam Shah and partners roll out beta version of Stanford medicine SHC and SoM Secure GPT. *Stanford Medicine* https://dbds.stanford.edu/2024/nigam-shaw-and-partners-roll-out-beta-version-of-stanford-medicine-shc-and-som-secure-gpt/ (2024).
26. Nori, H. et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. Preprint at https://arxiv.org/abs/2311.16452 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ethan Goh[1,2,14], Robert J. Gallo[3,4,14], Eric Strong[4], Yingjie Weng[5], Hannah Kerman[6,7], Jason A. Freed[6,7], Joséphine A. Cool[6,7], Zahir Kanjee[6,7], Kathleen P. Lane[8], Andrew S. Parsons[9], Neera Ahuja[4], Eric Horvitz[10,11], Daniel Yang[12], Arnold Milstein[2], Andrew P. J. Olson[8], Jason Hom[4,15], Jonathan H. Chen[1,2,13,15] ✉ & Adam Rodman[6,7,15]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA. [2]Stanford Clinical Excellence Research Center, Stanford University, Stanford, CA, USA. [3]Center for Innovation to Implementation, VA Palo Alto Health Care System, Palo Alto, CA, USA. [4]Stanford University School of Medicine, Stanford, CA, USA. [5]Quantitative Sciences Unit, Stanford University School of Medicine, Stanford, CA, USA. [6]Beth Israel Deaconess Medical Center, Boston, MA, USA. [7]Harvard Medical School, Boston, MA, USA. [8]Division of Hospital Medicine, University of Minnesota Medical School, Minneapolis, MN, USA. [9]Division of Hospital Medicine, University of Virginia School of Medicine, Charlottesville, VA, USA. [10]Microsoft, Redmond, WA, USA. [11]Stanford Institute for Human-Centered Artificial Intelligence, Stanford, CA, USA. [12]Kaiser Permanente, Oakland, CA, USA. [13]Division of Hospital Medicine, Stanford University, Stanford, CA, USA. [14]These authors contributed equally: Ethan Goh, Robert J. Gallo. [15]These authors jointly supervised this work: Jason Hom, Jonathan H. Chen, Adam Rodman. ✉e-mail: jonc101@stanford.edu

Methods

Participants

We recruited practicing attending physicians and resident physicians with training in a general medical specialty (internal medicine, family medicine or emergency medicine) through email lists from Stanford University, Beth Israel Deaconess Medical Center and the University of Virginia. Written informed consent was obtained before enrollment and randomization. This study was reviewed and determined to be exempt from institutional review board oversight by institutional review boards at Stanford University, Beth Israel Deaconess Medical Center and the University of Virginia. Small groups of participants were proctored by study coordinators either remotely or at an in-person computer laboratory. Sessions lasted for 1 h. Resident physicians were offered US$100 and attending physicians were offered US$200 to complete the study.

Clinical case vignette construction

We constructed our cases from the series of 'Grey Matters' from the American College of Physicians podcast 'Core IM'[27]. As these cases were adapted specifically for this study, they were not available to either GPT-4 or the participants before our study. Each of these cases was constructed by a panel of subspecialty and generalist experts (including A.R., Z.K., E.S., J.H. and A.S.P.) to explore how physicians make decisions when there are no clear right answers. We intentionally chose a selection of cases that would explore the breadth of general medicine management decision-making. Through initial pilot studies (not included in the analysis), we determined that no participant finished more than five cases within 1 h, in line with standardized tests of physician reasoning, such as licensing exams and observed structured clinical examinations[28,29].

Development of scoring rubrics

The paramount challenge in the evaluation of management reasoning is the relatively wide variety of reasonable answers depending on contextual factors[30,31]. Unlike a confirmed pathological final diagnosis, there is often a range of acceptable answers for management reasoning. To capture this nuance of a variety of management perspectives, for each case, we convened an expert group of five individuals—a member of the study team, two generalists and two subspecialists in the field applicable to the case. Through an iterative modified Delphi process, we refined management rubrics to score each case[32]. These rubrics were designed to be as thorough as possible for the specific case, while also acknowledging that considerable variation of acceptable management was possible. Because of this, scores on the rubrics do not comport with standard cutoffs from educational interventions (for example, 40% neither reflects a 'passing' or 'failing' score, only a percentage of the total possible points in a comprehensive rubric; points were awarded for all answers determined reasonable by the panel and while a higher score reflects a more comprehensive answer, there is no clear cutoff for high-quality care). Each of these rubrics were tested in two pilot groups and further refined with user feedback. Because often there was no clear divide between the diagnostic and management domains of clinical reasoning, each question was independently labeled by two members of the study team (E.G. and H.K.) as reflecting case-specific reasoning or more generalized clinical reasoning that did not require case-specific information. Case questions were similarly categorized as representing a diagnostic decision (for example, a differential for an incidentally found lung nodule), a management decision (for example, the contextual factors that drive the next steps in the workup of a lung nodule) or knowledge recall (the risk factors that make a lung nodule more likely to be malignant). There was complete agreement on these labels.

Study design

We used a prospective, randomized, single-blind (to the rater) study design with participants randomized to either using GPT-4 via the ChatGPT plus (OpenAI) interface or the conventional resources group (Fig. 1). To mirror real-world implementation, participants received GPT-4 training comparable to current live deployments in clinical settings[25]. This included basic instruction on system access and use, and live technical support throughout the study from a proctor.

Both groups were instructed that they could use any point-of-care resources they normally use in clinical practice, such as UpToDate (Wolters Kluwer), Epocrates (Athenahealth) and other internet resources. The control group was instructed not to use any LLMs (for example, ChatGPT, Claude, Bard/Gemini). We instructed participants to finish as many of the five cases as they could in an hour, prioritizing the quality of responses over completing all cases. The study was conducted using a Qualtrics survey tool; participants received the cases in sections before moving on. Participants were not able to change their answers to prior prompts as new pieces of information were introduced.

Prompt design for the LLM-only arm

For the LLM-only arm, we used established principles of prompt design to iteratively develop a zero-shot prompt by copy and pasting the management cases along with questions (Supplementary Item 2)[33]. Each prompt was run five times and the results from the five runs were included for blinded grading alongside the human outputs before any unblinding or data analysis.

Rubric validation

Two preliminary sets of data from ten individuals were collected to validate the rubrics. The three graders (A.R., E.S. and K.P.L.) independently graded these two datasets. They then met in person and came to a consensus on grading these two validation datasets. After data collection was complete, each case was graded independently by two of three graders who were blinded to group assignment. When scorers disagreed (predefined as a difference greater than 10% of the final score), they met to discuss differences in their assessments and to seek consensus. We calculated a weighted Cohen's kappa to show concordance in grading, both for each individual case and for all cases pooled together.

Study outcomes

The primary study outcome was the mean score for each of the groups. Secondary outcomes included scores in predefined domains of the rubrics, including management, knowledge recall and diagnostic domains, case specificity or generality of decisions and time spent on cases.

As an exploratory evaluation, a single blinded reviewer (A.R.) rated all responses for potential harm using a similar methodology to a previously published assessment[34]. The likelihood of possible harm was rated as low, medium or high; the extent of possible harm was rated as none, mild/moderate or severe/death. These ratings were performed at the individual response level, rather than at the case level used for other analyses.

Statistical methods

The target minimum sample size of 84 participants was prespecified based on a power analysis using the preliminary data of 13 cases among three participants, scored before study enrollment, corresponding to an expected 252–336 cases completed (3–4 cases per participant). This minimum target sample size ensured sufficient power (>80%) for both the primary outcome and time spent on cases as the secondary outcome. All analyses were at the case level, clustered according to the participant. In the primary analysis, we only included cases with completed responses, that is, answered up to the final question. To account for the potential of clustering, generalized mixed-effects models were applied to assess the difference in primary and secondary outcomes of the LLM group compared to the conventional resources-only group. A random effect for the participant was included in the model to account for the potential correlation between cases for a participant.

Additionally, a random effect for cases was included to account for any potential variability in difficulty across cases. Cases completed by the LLM alone were treated as a third group, with cases clustered in a nested structure of 5 attempts under a single participant, since repeat prompting of LLMs can have significant dependency with repeats[35,36]. All statistical analyses were performed using R v.4.3.2 (R Foundation for Statistical Computing). Statistical significance was set as $P < 0.05$.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Example case vignettes, questions and grading are included in the manuscript. All the raw scores produced by study participants are available via Figshare at https://doi.org/10.6084/m9.figshare.27886788 (ref. 37). Source data are provided with this paper.

Code availability

No custom code or software development was required for the current research.

References

27. Core IM. *American College of Physicians* www.acponline.org/cme-moc/internal-medicine-cme/internal-medicine-podcasts/core-im (2024).
28. Pell, G., Fuller, R., Homer, M. & Roberts, T. How to measure the quality of the OSCE: a review of metrics—AMEE guide no. 49. *Med. Teach.* **32**, 802–811 (2010).
29. Khan, K. Z., Ramachandran, S., Gaunt, K. & Pushkar, P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. *Med. Teach.* **35**, e1437–e1446 (2013).
30. Cook, D. A., Durning, S. J., Stephenson, C. R., Gruppen, L. D. & Lineberry, M. Assessment of management reasoning: design considerations drawn from analysis of simulated outpatient encounters. *Med. Teach*. 1–15, https://doi.org/10.1080/0142159X.2024.2337251 (2024).
31. Singaraju, R. C., Durning, S. J., Battista, A. & Konopasky, A. Exploring procedure-based management reasoning: a case of tension pneumothorax. *Diagnosis* **9**, 437–445 (2022).
32. Jones, J. & Hunter, D. Consensus methods for medical and health services research. *BMJ* **311**, 376–380 (1995).
33. Meskó, B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J. Med. Internet Res.* **25**, e50638 (2023).
34. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
35. Gallo, R. J., Savage, T. & Chen, J. H. Affiliation bias in peer review of abstracts. *JAMA* **331**, 1234–1235 (2024).
36. Gallo, R. J. et al. Establishing best practices in large language model research: an application to repeat prompting. *J. Am. Med. Inform. Assoc.* **32**, 386–390 (2025).
37. Goh, E. et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Figshare* https://doi.org/10.6084/m9.figshare.27886788 (2025).

Acknowledgements

We thank M. Chua, M. Maddali, H. Magon and M. Schwede for providing feedback and participating in our pilot study.

Author contributions

E.G. and R.J.G. participated in study design, acquired and interpreted the data, prepared the manuscript and revised it critically. E.S. and H.K. participated in study design, and acquired and interpreted the data. Y.W., J.A.F., J.C., Z.K., K.P.L., A.S.P., D.Y. and A.P.J.O. participated in study design and interpreted the data. A.M. and N.A. acquired the funding and provided administrative support. E.H. participated in study design and provided critical revision of the manuscript. J.H. and J.H.C. participated in study design, analyzed and interpreted the data, carried out critical revision of the manuscript, supervised the study, acquired the funding and provided administrative support. A.R. participated in study design, analyzed and interpreted the data, carried out critical revision of the manuscript and supervised the study.

Competing interests

E.G., J.H., E.S., J.C., Z.K., A.P.J.O., A.R. and J.H.C. disclose funding from the Gordon and Betty Moore Foundation (grant no. 12409). R.J.G. is supported by a VA Advanced Fellowship in Medical Informatics. Z.K. discloses royalties from Wolters Kluwer for books edited (unrelated to this study), former paid advisory membership for Wolters Kluwer on medical education products (unrelated to this study) and honoraria from Oakstone Publishing for CME delivered (unrelated to this study). A.S.P. discloses a paid advisory role for *New England Journal of Medicine* Group and National Board of Medical Examiners for medical education products (unrelated to this study). A.P.J.O. receives funding from 3M for research related to rural health workforce shortages. and consulting fees for work related to a clinical reasoning application from the *New England Journal of Medicine*. A.M. reports uncompensated and compensated relationships with care.coach, Emsana Health, Embold Health, ezPT, FN Advisors, Intermountain Healthcare, JRSL, The Leapfrog Group, the Peterson Center on Healthcare, Prealize Health and PBGH. J.H.C. reports cofounding Reaction Explorer, which develops and licenses organic chemistry education software, as well as paid consulting fees from Sutton Pierce, Younker Hyde Macfarlane and Sykes McAllister as a medical expert witness. He receives funding from the National Institutes of Health (NIH)/National Institute of Allergy and Infectious Diseases (1R01AI17812101), NIH/National Institute on Drug Abuse Clinical Trials Network (UG1DA015815—CTN-0136), Stanford Artificial Intelligence in Medicine and Imaging—Human-Centered Artificial Intelligence Partnership Grant, the NIH-NCATS-Clinical & Translational Science Award (UM1TR004921), Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) [R12], NIH/Center for Undiagnosed Diseases at Stanford (U01 NS134358) and the American Heart Association—Strategically Focused Research Network—Diversity in Clinical Trials. J.H. discloses a paid advisory role for Cognita Imaging. The other authors declare no competing interests. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03456-y.

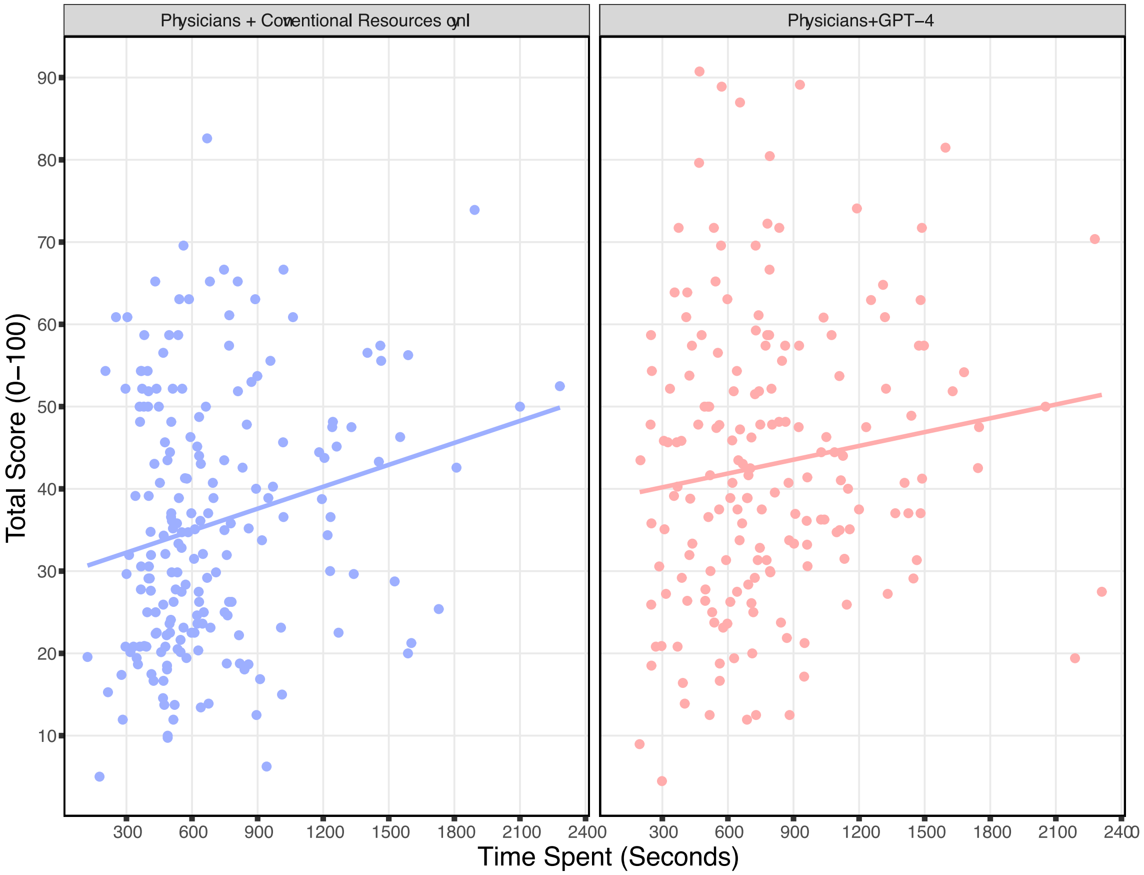
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03456-y.

Correspondence and requests for materials should be addressed to Jonathan H. Chen.

Peer review information *Nature Medicine* thanks Eric Oermann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

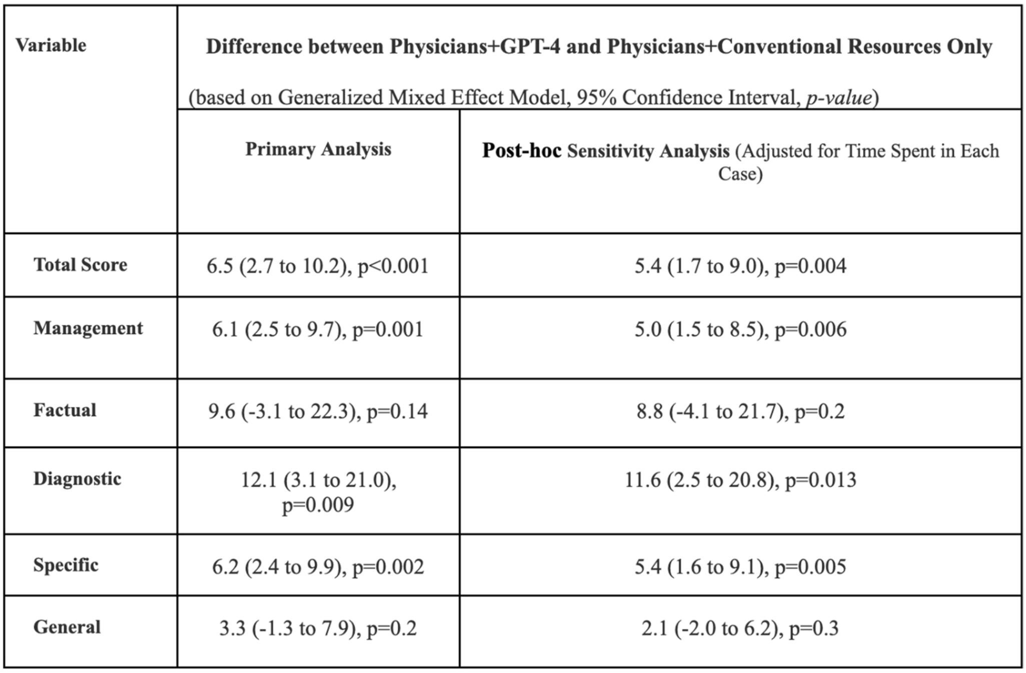
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41591-024-03456-y



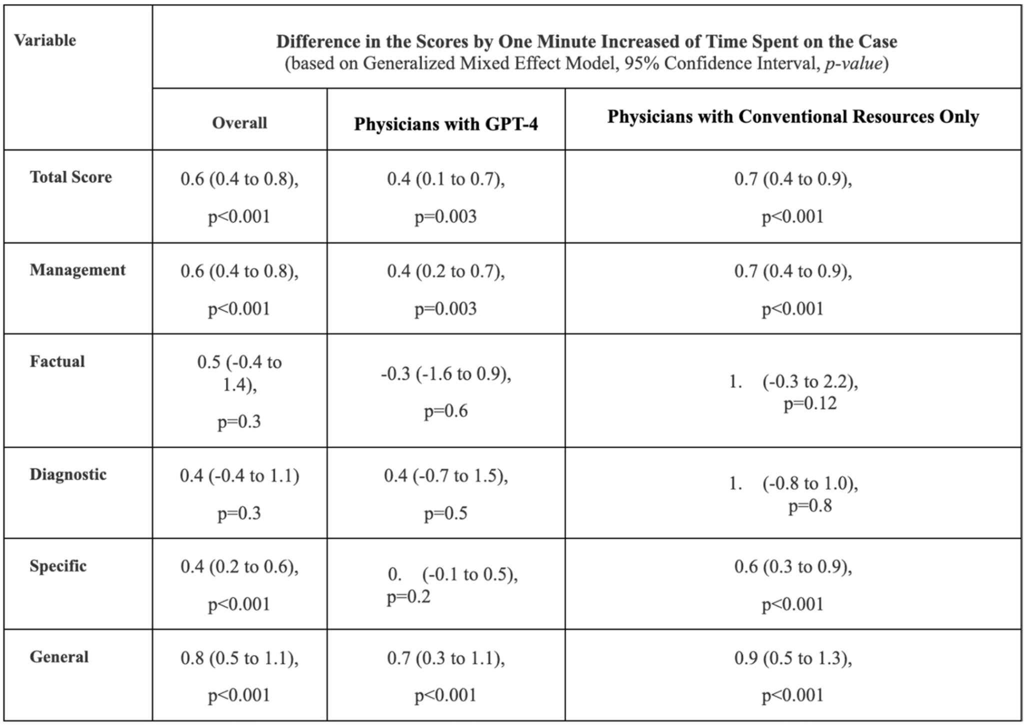
Extended Data Fig. 1 | Correlation between Time Spent in Seconds and Total Score. This figure demonstrates a sample medical management case with multi-part assessment questions, scoring rubric and example responses. The case presents a 72-year-old post-cholecystectomy patient with new-onset atrial fibrillation. The rubric (23 points total) evaluates clinical decision-making across key areas: initial workup, anticoagulation decisions, and outpatient monitoring strategy. Sample high-scoring (21/23) and low-scoring (8/23) responses illustrate varying depths of clinical reasoning and management decisions.

Extended Data Table 1 | Post-hoc Analysis Adjusted for Time Spent in Each Case



Variable	**Difference between Physicians+GPT-4 and Physicians+Conventional Resources Only** (based on Generalized Mixed Effect Model, 95% Confidence Interval, *p-value*)	
	Primary Analysis	**Post-hoc Sensitivity Analysis** (Adjusted for Time Spent in Each Case)
Total Score	6.5 (2.7 to 10.2), p<0.001	5.4 (1.7 to 9.0), p=0.004
Management	6.1 (2.5 to 9.7), p=0.001	5.0 (1.5 to 8.5), p=0.006
Factual	9.6 (-3.1 to 22.3), p=0.14	8.8 (-4.1 to 21.7), p=0.2
Diagnostic	12.1 (3.1 to 21.0), p=0.009	11.6 (2.5 to 20.8), p=0.013
Specific	6.2 (2.4 to 9.9), p=0.002	5.4 (1.6 to 9.1), p=0.005
General	3.3 (-1.3 to 7.9), p=0.2	2.1 (-2.0 to 6.2), p=0.3

Extended Data Table 2 | Post-hoc Analysis for the Associations between the Primary and Secondary Outcomes Overall



Variable	Difference in the Scores by One Minute Increased of Time Spent on the Case (based on Generalized Mixed Effect Model, 95% Confidence Interval, *p-value*)		
	Overall	**Physicians with GPT-4**	**Physicians with Conventional Resources Only**
Total Score	0.6 (0.4 to 0.8), p<0.001	0.4 (0.1 to 0.7), p=0.003	0.7 (0.4 to 0.9), p<0.001
Management	0.6 (0.4 to 0.8), p<0.001	0.4 (0.2 to 0.7), p=0.003	0.7 (0.4 to 0.9), p<0.001
Factual	0.5 (-0.4 to 1.4), p=0.3	-0.3 (-1.6 to 0.9), p=0.6	1. (-0.3 to 2.2), p=0.12
Diagnostic	0.4 (-0.4 to 1.1) p=0.3	0.4 (-0.7 to 1.5), p=0.5	1. (-0.8 to 1.0), p=0.8
Specific	0.4 (0.2 to 0.6), p<0.001	0. (-0.1 to 0.5), p=0.2	0.6 (0.3 to 0.9), p<0.001
General	0.8 (0.5 to 1.1), p<0.001	0.7 (0.3 to 1.1), p<0.001	0.9 (0.5 to 1.3), p<0.001

nature portfolio

Corresponding author(s): Jonathan Chen, Ethan Goh

Last updated by author(s): Nov 18, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

□ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: Data collected using Qualtrics survey tool. ChatGPT Plus used by participants as described in manuscript

Data analysis: Data analysis performed using R statistical software.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw score table available upon reasonable request, as this was the language included in participant informed consent forms.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	This information was not collected, as it was not considered relevant to the study question
Population characteristics	This information was not collected, as it was not considered relevant to the study question
Recruitment	67 (73%) attendings; 68 (74%) internal medicine; median 5.8 years in practice.
Ethics oversight	Participants recruited through professional network email lists.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Target minimum sample size of 84 participants was pre-specified based on a power analysis using preliminary data scored prior to study enrollment
Data exclusions	No data were excluded.
Replication	Replication not performed given human subjects research.
Randomization	Participants randomly assigned.
Blinding	Graders were blinded to study arm.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	ClinicalTrials.gov registration: NCT06208423
Study protocol	Reported in supplementary materials.

Data collection

From 11/2023-4/2024, participants proctored in small groups either remotely or in-person.

Outcomes

Primary outcome predefined as score on structured reflection rubric. Secondary outcome included time spent on cases.