JAMA Network Open

Original Investigation | Health Informatics

Large Language Model Influence on Diagnostic Reasoning
A Randomized Clinical Trial

Ethan Goh, MBBS, MS; Robert Gallo, MD; Jason Hom, MD; Eric Strong, MD; Yingjie Weng, MHS; Hannah Kerman, MD; Joséphine A. Cool, MD; Zahir Kanjee, MD, MPH; Andrew S. Parsons, MD, MPH; Neera Ahuja, MD; Eric Horvitz, MD, PhD; Daniel Yang, MD; Arnold Milstein, MD; Andrew P. J. Olson, MD; Adam Rodman, MD, MPH; Jonathan H. Chen, MD, PhD

Abstract

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

(continued)

Key Points

Question Does the use of a large language model (LLM) improve diagnostic reasoning performance among physicians in family medicine, internal medicine, or emergency medicine compared with conventional resources?

Findings In a randomized clinical trial including 50 physicians, the use of an LLM did not significantly enhance diagnostic reasoning performance compared with the availability of only conventional resources.

Meaning In this study, the use of an LLM did not necessarily enhance diagnostic reasoning of physicians beyond conventional resources; further development is needed to effectively integrate LLMs into clinical practice.

+ Visual Abstract

+ Invited Commentary

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

technology and workforce development to realize the potential of physician-artificial intelligence collaboration in clinical practice.

TRIAL REGISTRATION ClinicalTrials.gov Identifier: NCT06157944

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

Introduction

Diagnostic errors are common, contribute to substantial patient harm, and result from a combination of cognitive and systems factors.[1-5] Effective interventions to improve diagnostic performance and reduce diagnostic errors will need to focus on both systems factors and cognitive factors, often referred to as clinical reasoning. Strategies that have been advanced to improve clinical reasoning include a variety of educational, reflective, and team-based practices, as well as clinical decision support tools.[6] The impact of these interventions has been limited, and even the most useful methods, such as reflective practice, are difficult to integrate clinically at scale.[7,8] Artificial intelligence (AI) technologies have long been pursued as promising tools for assisting physicians with diagnostic reasoning.

Large language models (LLMs)—machine learning systems that produce humanlike responses from written language—have shown the ability to solve complex cases, exhibit humanlike clinical reasoning, take patient histories, and display empathetic communication.[9-14] Due to their generalizable nature, LLMs are actively being integrated into multiple health care settings.[15-20] Despite the impressive performance of these emerging technologies in benchmarking tasks, current integrations of LLMs require human participation, with the LLM augmenting, rather than replacing, human expertise and oversight.[21] Understanding the implications of deploying these systems in patient care with limited workforce training and integration requires human-computer user studies with richer measures of diagnostic reasoning.

We performed a randomized clinical trial to compare the diagnostic reasoning performance of physicians using a commercial LLM AI chatbot (ChatGPT Plus [GPT-4]; OpenAI) compared with conventional diagnostic resources (eg, UpToDate, Google). Many studies of diagnostic performance only assess shallow measures of accuracy without attention to the quality of the diagnostic process used to arrive at that diagnosis. To develop a deeper assessment of how new tools affect physician reasoning, we further adapted structured reflection—a measure of factors contributing to a diagnostic decision—as a novel assessment tool of the diagnostic process.[22]

Methods

This study was reviewed and determined to be exempt from approval by institutional review boards at Stanford University, Beth Israel Deaconess Medical Center, and the University of Virginia. Informed consent was obtained prior to enrollment and randomization. Resident participants were offered $100 and attending participants were offered up to $200 for completing the study. This study follows the Consolidated Standards of Reporting Trials (CONSORT) reporting guideline for randomized clinical trials. The study protocol is available in Supplement 1.

We recruited attending and resident physicians with training in a general medical specialty (internal medicine, family medicine, or emergency medicine) through email lists 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 hour. The participant flow is depicted in the **Figure**. A visual iteration is presented in eFigure 2 in Supplement 2.

Clinical Vignettes

Clinical vignettes were adapted from a landmark study that set the standard for the evaluation of computer-based diagnostic systems.[23] All cases in this study were based on actual patients and included information available on initial diagnostic evaluation, such as history, physical examination, and laboratory test results. The cases have never been publicly released to protect the validity of the test materials for future use, and therefore are excluded from training data of the LLM. A representative example is included in eTable 1 in Supplement 2. We used the nominal group technique to select a cross-section of cases; 4 physician authors (E.G., J.A.C., A.P.J.O., and J.H.C.) met to agree on case selection guidelines including preference for a broad range of pathologic settings, avoiding simplistic cases with limited plausible diagnoses, and excluding exceedingly rare cases.[24] Each member independently reviewed at least 50 of the 105 available cases to identify a minimum of 10 cases that satisfied selection guidelines. After individual ratings, the group convened again to come to a consensus on a prioritized list of cases to consider. In pilot tests, participants completed a maximum of 6 cases in 1 hour, leading us to select 6 final cases for this study. Cases were edited to modernize laboratory data reporting conventions and to replace pathognomic phrases (eg, livedo reticularis) with general descriptions (eg, purple, red, lacy rash).

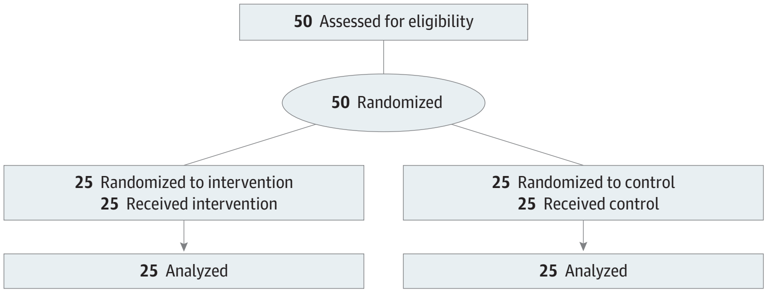
A common, but limited, evaluation benchmark in clinical decision support diagnostic studies is accuracy of differential diagnosis. While we assessed overall differential diagnosis accuracy as a secondary outcome similar to prior studies, the complex phenomena of human-computer interactions warrant richer evaluations of diagnostic reasoning skills. We therefore chose to develop an assessment from the clinical reasoning literature: structured reflection.[25]

Structured reflection aims to improve the process by which physicians consider reasonable diagnoses and clinical features that support or oppose their diagnoses, similar to how physicians may explain their reasoning in the assessment and plan component of clinical notes.[25,26] We adapted a structured reflection grid (eTable 1 in Supplement 2) with participants providing free-text responses on their top 3 differential diagnoses, the factors in the case that favor or oppose each of their 3 diagnoses, their final most likely diagnosis, and up to 3 next steps (eg, diagnostic tests) they would use to further evaluate the patient.

Assessing Performance

We built on previous studies of structured reflection by scoring the grid itself, not just final diagnosis accuracy. For each case, we assigned up to 1 point for each plausible diagnosis. Findings supporting each diagnosis and findings opposing the diagnosis were also graded based on correctness, with 0 points for incorrect or absent answers, 1 point for partially correct, and 2 points for completely correct responses. The final diagnosis was graded as 2 points for the most correct diagnosis and 1 point for a plausible diagnosis or a correct diagnosis that was not specific enough compared with the most correct final diagnosis. The participants then were instructed to describe up to 3 next steps to further evaluate the patient, with 0 points awarded for an incorrect response, 1 point awarded for a partially correct response, and 2 points awarded for a completely correct response (eTable 2 in

Figure. Participant Flowchart



Supplement 2). While incorrect differential diagnoses were not awarded points, appropriate reasoning based on those diagnoses were not penalized. Raters were blinded to participant group assignments.

Study Design

We used a randomized single-blind study design with stratified randomization. Participants were randomized to use the LLM interface (intervention group) or conventional resources (control group). They were given access to study accounts for the LLM, and transcripts from their use were saved. Both groups were instructed to access any conventional resources they normally use for clinical care, but the control group was explicitly instructed not to use LLMs. Participants had 1 hour to complete as many of the 6 diagnostic cases as they could, with instructions to prioritize quality of their responses over completing all cases.

The study was conducted using a survey tool (Qualtrics), with cases presented in random order for each participant. In a secondary analysis, we included a comparison arm using the LLM alone to answer the cases. Using established principles of prompt design, we iteratively developed a 0-shot prompt; the same language was used along with the clinical vignette questions for each case.[27] The researcher physician inputting prompts to the model did not alter model responses. eTable 4 in Supplement 2 gives an example prompt. These prompts were run 3 times in separate sessions, and the results from each run were included for blinded grading alongside the human outputs before unblinding or data analysis.

Assessment Tool Validation

To establish validity, we collected 2 sets of pilot data with 13 participants not included in the final study. This included a total of 65 cases completed, based on a sampling of multiple case vignettes, including the 6 used in the final study. The 3 primary scorers (J.H, A.P.J.O., and A.R.), all board-certified physicians with experience in the evaluation of clinical reasoning at the postgraduate medical level, graded these independently to assess consistency. Based on iterative feedback from both graders and pilot participants, as well as grader concordance, the study case vignettes were selected and rubrics were further refined before data were collected for the final study. After data collection, each case was graded independently by 2 scorers who were blinded to the assigned treatment group. Disagreement between scorers was predefined as a difference of more than 10% of the final score. eTable 5 in Supplement 2 gives variance by subcomponents. When scorers disagreed, they met to discuss differences in their assessments and seek consensus. We designed scoring to intentionally acknowledge ambiguity in diagnostic processes, allowing for multiple variations of correct answers determined by scorer consensus. Final diagnosis scoring was adjudicated by 2 scorers to obtain agreement for the secondary outcome of diagnostic accuracy. We calculated a weighted Cohen κ value to show concordance in grading and Cronbach α value to determine the internal reliability of this measure.[28,29]

Study Outcomes

Our primary outcome was the final score as a percentage across all components of the structured reflection tool. Secondary outcomes were time spent per case (in seconds) and final diagnosis accuracy. Final diagnosis was treated as an ordinal outcome with 3 groups (incorrect, partially correct, and most correct). Since the difference between the most correct response and partially correct responses may not be clinically meaningful, we additionally analyzed the outcome as binary (incorrect compared with at least partially correct).

Statistical Analysis

The target sample size of 50 participants was prespecified based on a power analysis using 2 validation sets of data, scored before study enrollment. Using PASS 2023, version 23.0.2 software (NCSS LLC), our power analysis showed more than 80% power to detect an 8% score difference with

200 to 250 cases completed (4 to 5 cases per participant) with a 2-sided α value of .05. We used a mixed-effects model suitable for cluster-randomized designs, with an intraclass correlation coefficient ranging from 0.05 to 0.15 and an SD of 16.2%.

All analyses followed the intention-to-treat principle and were conducted at the case level, clustered by the participant. Linear mixed-effects models were applied to assess the difference in the primary outcome of diagnostic performance and the secondary outcome of time spent per completed case, with normality assumptions verified. Ordinal and logistic mixed-effects models were used for comparisons of other secondary outcomes including ordinal and binary final diagnosis accuracy. A random effect for the participant was included in the models 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. Family-wise type I error (α) was controlled at .05 for the primary outcome of diagnostic performance considered as a continuous variable. Analysis of the secondary outcomes was exploratory without adjusting for multiple comparisons. A preplanned sensitivity analysis evaluated the effect of including incomplete cases on the primary outcome. Subgroup analyses were conducted based on training status and experience with the LLM product used. In a secondary analysis, cases completed by the LLM alone were treated as a third group, with cases clustered in a nested structure of 3 attempts under a single participant. These were compared with cases from real participants with each case considered as a single attempt under a single participant using a similar nested structure.

All statistical analysis was performed using R, version 4.3.2 (R Foundation for Statistical Computing). Further details regarding the trial protocol and statistical analysis plan are provided in Supplement 1.

Results

Fifty US-licensed physicians were recruited and participated (26 attendings, 24 residents) from November 29 to December 29, 2023; of these, 39 (78%) participated in virtual encounters and 11 (22%) were in-person. Median years in practice was 3 (IQR, 2-8). Further information on participants is included in **Table 1**.

Table 1. Baseline Participant Characteristics

	Participants, No. (%)		
Participant characteristic	Overall (N = 50)	Physicians plus LLM (n = 25)	Physicians plus conventional resources (n = 25)
Career stage			
Attending	26 (52)	13 (52)	13 (52)
Resident	24 (48)	12 (48)	12 (48)
Specialty			
Internal medicine	44 (88)	22 (88)	22 (88)
Family medicine	1 (2)	1 (4)	0
Emergency medicine	5 (10)	2 (8)	3 (12)
Years in practice, median (IQR)	3 (2-8)	3 (2-7)	3 (2-9)
LLM experience			
I've never used it before	8 (16)	5 (20)	3 (12)
I've used it once ever	6 (12)	4 (16)	2 (8)
I use it rarely (less than once per month)	15 (30)	7 (28)	8 (32)
I use it occasionally (more than once per month but less than weekly)	13 (26)	6 (24)	7 (28)
I use it frequently (weekly or more)	8 (16)	3 (12)	5 (20)

Abbreviation: LLM, large language model.

Primary Outcome: Diagnostic Performance

A total of 244 cases were completed by all participants (125 cases in LLM group, 119 cases in control group). The median number of completed cases per participant was 5 (IQR, 4-6). Analysis of the transcripts showed that 100% (22 of 22) of physicians randomized to use the LLM did so; 3 transcripts were lost due to technical issues and not included. The median score per case was 76% (IQR, 66%-87%) for the LLM group and 74% (IQR, 63%-84%) for the control group. The mixed-effects model showed a difference of 2 percentage points (95% CI, −4 to 8 percentage points; *P* = .60) between the LLM and control groups, as presented in **Table 2**. A sensitivity analysis including all cases, complete and incomplete, showed a similar result with a difference of 2 percentage points (95% CI, −4 to 8 percentage points; *P* = .50) between the LLM and control group. The distribution of diagnostic performance scores by group is given in eFigure 1 in Supplement 2.

Secondary Outcomes

Median time spent per case was 519 (IQR, 371-668) seconds for the LLM group and 565 (IQR, 456-788) seconds for the control group (**Table 3**). The linear mixed-effects model resulted in an adjusted difference of −82 seconds (95% CI, −195 to 31 seconds; *P* = .20).

Accuracy of the final diagnosis (eTable 3 in Supplement 2) using the ordinal scale showed the LLM intervention group had 1.4 times higher odds (95% CI, 0.7-2.8; *P* = .39) of a correct diagnosis than the control group. In assessing the accuracy of final diagnoses, treating them as binary (correct vs incorrect) variables did not qualitatively change the results (odds ratio, 1.9; 95% CI, 0.9-4.0; *P* = .10).

Subgroup Analyses

Table 2 and Table 3 include the analyses by subgroups, including level of training and level of experience with the LLM. Subgroup analyses were qualitatively similar to the analyses for the whole cohort.

Table 2. Diagnostic Performance Outcomes

	Median (IQR), %			
Group	Physicians plus LLM	Physicians plus conventional resources	Difference (95% CI), percentage points[a]	*P* value
All participants	76 (66 to 87)	74 (63 to 84)	2 (−4 to 8)	.60
Level of training				
Attending	79 (63 to 87)	75 (61 to 87)	0.5 (−9 to 1)	.92
Resident	76 (68 to 84)	74 (63 to 84)	3 (−6 to 11)	.50
LLM experience				
Less than monthly	76 (63 to 84)	76 (63 to 87)	−0.5 (−8 to 7)	.90
More than monthly	79 (68 to 90)	74 (63 to 84)	5 (−7 to 16)	.40

Abbreviation: LLM, large language model.

[a] Differences between groups are reported from the multilevel analysis accounting for clustering of cases by participant.

Table 3. Time Spent per Case

	Median (IQR) time, s			
Group	Physicians plus LLM	Physicians plus conventional resources	Difference (95% CI)[a]	*P* value
All participants	519 (371 to 668)	565 (456 to 788)	−82 (−195 to 31)	.15
Level of training				
Attending	533 (389 to 672)	563 (435 to 778)	−73 (−204 to 58)	.26
Resident	478 (356 to 654)	565 (458 to 800)	−76 (−284 to 131)	.45
LLM experience				
Less than monthly	556 (415 to 742)	572 (474 to 778)	−46 (−219 to 127)	.59
More than monthly	462 (305 to 627)	556 (427 to 810)	−140 (−294 to 13)	.07

Abbreviation: LLM, large language model.

[a] Differences between groups are reported from the multilevel analysis accounting for clustering of cases by participant.

LLM Alone

In the 3 runs of the LLM alone, the median score per case was 92% (IQR, 82%-97%). Comparing LLM alone with the control group found an absolute score difference of 16 percentage points (95% CI, 2-30 percentage points; *P* = .03) favoring the LLM alone.

Assessment Tool Validation

The weighted Cohen κ value between all 3 graders was 0.66, indicating substantial agreement within the expected range for diagnostic performance studies.[30] The overall Cronbach α value was 0.64. The variances of individual sections of the structured reflection rubric are presented in eTable 5 in Supplement 2. After removing the final diagnosis, which had the highest variance, the Cronbach α value was 0.67.

Discussion

This randomized clinical trial found that physician use of a commercially available LLM chatbot did not improve diagnostic reasoning on challenging clinical cases, despite the LLM alone significantly outperforming physician participants. The results were similar across subgroups of different training levels and experience with the chatbot. These results suggest that access alone to LLMs will not improve overall physician diagnostic reasoning in practice. These findings are particularly relevant now that many health systems offer Health Insurance Portability and Accountability Act–compliant chatbots that physicians can use in clinical settings, often with no to minimal training on how to use these tools.[15,17-19]

Our data did not confirm any differences in time spent solving cases. With wide variability observed in time to complete cases, future studies with substantially larger sample sizes would be necessary to evaluate whether physicians with experience using LLMs spend less time on diagnostic reasoning.

An unexpected secondary result was that the LLM alone performed significantly better than both groups of humans, similar to a recent study with different LLM technology.[31] This may be explained by the sensitivity of LLM output to prompt formulation.[32] There are numerous frameworks for prompting LLMs and an emerging consensus on prompting strategies, many of which focus on providing details on the task, context, and instructions; our prompt was iteratively developed using these frameworks. Training clinicians in best prompting practices may improve physician performance with LLMs. Alternatively, organizations could invest in predefined prompting for diagnostic decision support integrated into clinical workflows and documentation, enabling synergy between the tools and clinicians. Prior studies on AI systems show disparate effects depending on the component of the diagnostic process they are used in.[33,34] Given the conversational nature of chatbots, changes in how the LLM interacts with humans, for example by specifically pointing out features that do not fit the differential diagnosis, might improve diagnostic and reflective performance.[35,36] More generally, we see opportunity with deliberate consideration and redesign of medical education and practice frameworks that adapt to disruptive emerging technologies and enable the best use of computer and human resources to deliver optimal medical care.

Results of this study should not be interpreted to indicate that LLMs should be used for diagnosis autonomously without physician oversight. The clinical case vignettes were curated and summarized by human clinicians, a pragmatic and common approach to isolate the diagnostic reasoning process, but this does not capture competence in many other areas important to clinical reasoning, including patient interviewing and data collection.[37] Furthermore, this study was acontextual, and clinicians' understanding of the clinical environment is fundamental for high-quality decision-making. While early studies show that LLMs might effectively collect and summarize patient information, these capabilities need to be studied more thoroughly.[12,16] Additionally, improvement in rubric scoring here represents an important signal of clinical reasoning, but broader clinical trials are necessary to assess for meaningful differences in downstream clinical impact.

This study developed a measure based on structured reflection, inspired by research on physician cognition.[38] Scoring the adapted structured reflection tool as a primary outcome represents a novel contribution of this study to offer a richer evaluation framework of diagnostic reasoning skills. This assessment tool demonstrated substantial agreement between graders and internal reliability similar or superior to other measures used in the assessment of reasoning.[39-42] This advances the field beyond early LLM research, which has focused on benchmarks with limited clinical utility, such as multiple-choice question banks used for medical licensing or curated case vignettes of diseases rarely seen in clinical practice, such as clinicopathologic case conferences.[11,43] While having obvious advantages in ease of measurement, these tasks are not consistent with clinical reasoning in practice. As AI research progresses and nears clinical integration, it will become even more important to reliably measure diagnostic performance using the most realistic and clinically relevant evaluation methods and metrics.

Limitations

This trial has limitations. We focused our investigation around a single LLM, given its commercial availability and integration into clinical practice.[15,17-19] Multiple alternative LLM systems are rapidly emerging, although the one studied currently remains among the most performant tools for the applications studied.[44,45] Participants were given access to the chatbot without explicit training in prompt engineering techniques that could have improved the quality of their interactions with the system; however, this is consistent with current integrations and thus requires this representative evaluation.[15,17-19] Furthermore, even though all of the physicians in the LLM arm at least tried to use the system based on chat logs, they were not forced to use the system in any consistent way. This was a purposeful design to better reflect an effectiveness evaluation in the clinical practice setting.

No sample of clinical vignettes can comprehensively cover the variety of cases in the field of medicine. Our study included 6 cases that could feasibly be completed within a single study session while remaining comparable to standard practices in national licensing and objective structured clinical examinations to use a small, but broad sample of clinical cases.[6,46-49] This is not meant to comprehensively assess a participant's knowledge, but rather to evaluate their general clinical reasoning across a set of cases. To maximize a range of coverage, we deliberately selected cases to capture a broad and relevant cross-section of disciplines and a range of clinical problems.

Conclusions

The availability of an LLM as a diagnostic aid did not improve physician performance compared with conventional resources in a diagnostic reasoning randomized clinical trial. The LLM alone outperformed physicians even when the LLM was available to them, indicating that further development in human-computer interactions is needed to realize the potential of AI in clinical decision support systems.

ARTICLE INFORMATION

Accepted for Publication: August 2, 2024.

Published: October 28, 2024. doi:10.1001/jamanetworkopen.2024.40969

Corresponding Author: Ethan Goh, MD, MS, Stanford Clinical Excellence Research Center, Stanford Center for Biomedical Informatics Research, Stanford University School of Medicine 453 Quarry Rd Palo Alto, CA 94304 (ethangoh@stanford.edu).

Author Affiliations: Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California (Goh, Chen); Stanford Clinical Excellence Research Center, Stanford University, Stanford, California (Goh, Milstein, Chen); Center for Innovation to Implementation, VA Palo Alto Health Care System, Palo Alto, California (Gallo);

Department of Hospital Medicine, Stanford University School of Medicine, Stanford, California (Hom, Strong, Ahuja); Quantitative Sciences Unit, Stanford University School of Medicine, Stanford, California (Weng); Department of Hospital Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts (Kerman, Cool, Kanjee, Rodman); Department of Hospital Medicine, Harvard Medical School, Boston, Massachusetts (Kerman, Cool, Kanjee, Rodman); Department of Hospital Medicine, School of Medicine, University of Virginia, Charlottesville (Parsons); Microsoft Corp, Redmond, Washington (Horvitz); Stanford Institute for Human-Centered Artificial Intelligence, Stanford, California (Horvitz); Department of Hospital Medicine, Kaiser Permanente, Oakland, California (Yang); Department of Hospital Medicine, University of Minnesota Medical School, Minneapolis (Olson); Division of Hospital Medicine, Stanford University, Stanford, California (Chen).

Author Contributions: Drs Goh and Gallo had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Goh and Gallo contributed equally to this work. Drs Olson, Rodman, and Chen contributed equally to this work.

Concept and design: All authors.

Acquisition, analysis, or interpretation of data: Goh, Gallo, Hom, Weng, Kerman, Cool, Kanjee, Parsons, Ahuja, Horvitz, Rodman, Chen.

Drafting of the manuscript: Goh, Gallo, Weng, Cool, Horvitz, Olson, Rodman, Chen.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Goh, Gallo, Weng.

Obtained funding: Goh, Hom, Ahuja, Horvitz, Yang, Milstein, Rodman, Chen.

Administrative, technical, or material support: Goh, Hom, Parsons, Ahuja, Horvitz, Olson, Rodman, Chen.

Supervision: Goh, Horvitz, Olson, Rodman, Chen.

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 Nation 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.

Funding/Support: This study was funded by the Gordon and Betty Moore Foundation.

Role of the Funder/Sponsor: The funding organization had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 3.

REFERENCES

1. Shojania KG, Burton EC, McDonald KM, Goldman L. Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *JAMA*. 2003;289(21):2849-2856. doi:10.1001/jama.289.21.2849

2. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. *JAMA Intern Med*. 2013;173(6):418-425. doi:10.1001/jamainternmed.2013.2777

3. Auerbach AD, Lee TM, Hubbard CC, et al; UPSIDE Research Group. Diagnostic errors in hospitalized adults who died or were transferred to intensive care. *JAMA Intern Med*. 2024;184(2):164-173. doi:10.1001/jamainternmed.2023.7347

4. Balogh EP, Miller BT, Ball JR, eds; *Improving Diagnosis in Health Care*. National Academies Press; December 29, 2015. doi:10.17226/21794

5. Newman-Toker DE, Peterson SM, Badihian S, et al. Diagnostic errors in the emergency department: a systematic review. Agency for Healthcare Research and Quality. December 2022 report No.:22(23)-EHC043. Accessed September 23, 2024. https://www.ncbi.nlm.nih.gov/books/NBK588118/pdf/Bookshelf_NBK588118.pdf

6. Daniel M, Rencic J, Durning SJ, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. *Acad Med*. 2019;94(6):902-912. doi:10.1097/ACM.0000000000002618

7. Ilgen JS, Bowen JL, McIntyre LA, et al. Comparing diagnostic performance and the utility of clinical vignette-based assessment under testing conditions designed to encourage either automatic or analytic thought. *Acad Med*. 2013;88(10):1545-1551. doi:10.1097/ACM.0b013e3182a31c1e

8. Mamede S, van Gog T, van den Berge K, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA*. 2010;304(11):1198-1203. doi:10.1001/jama.2010.1276

9. Goh E, Bunning B, Khoong E, et al. ChatGPT influence on medical decision-making, bias, and equity: a randomized study of clinicians evaluating clinical vignettes. *medRxiv*. Preprint posted online November 27, 2023. doi:10.1101/2023.11.24.23298844

10. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit Med*. 2024;7(1):20. doi:10.1038/s41746-024-01010-1

11. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*. 2023;330(1):78-80. doi:10.1001/jama.2023.8288

12. Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. *arXiv*. Preprint posted online January 11, 2024. doi:10.48550/arXiv.2401.05654

13. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

14. Strong E, DiGiammarino A, Weng Y, et al. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA Intern Med*. 2023;183(9):1028-1030. doi:10.1001/jamainternmed.2023.2909

15. Nigam Shah and partners roll out beta version of Stanford medicine SHC and SoM Secure GPT. Stanford Department of Biomedical Data Science. Published August 8, 2024. Accessed February 19, 2024. https://dbds.stanford.edu/2024/nigam-shaw-and-partners-roll-out-beta-version-of-stanford-medicine-shc-and-som-secure-gpt/

16. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal*. Published online February 21, 2024. doi:10.1056/CAT.23.0404

17. Brian. Washington University ChatGPT Beta is Now Available - Information Technology. Washington University in St. Louis. Published December 19, 2023. Accessed June 24, 2024. https://it.wustl.edu/2023/12/washington-university-chatgtp-beta-is-now-available/

18. AI Sandbox. Harvard University Information Technology. Accessed May 3, 2024. https://huit.harvard.edu/ai-sandbox

19. Generative AI at VUMC. Department of Biomedical Informatics. Vanderbilt University Medical Center. Accessed May 3, 2024. https://www.vumc.org/dbmi/GenerativeAI

20. Schwartz N. Google tests ChatGPT competitor at Mayo Clinic. Becker's Health IT. Published July 10, 2023. Accessed June 24, 2024. https://www.beckershospitalreview.com/innovation/google-tests-chatgpt-competitor-at-mayo-clinic.html

21. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls: a narrative review. *Ann Intern Med*. 2024;177(2):210-220. doi:10.7326/M23-2772

22. Mamede S, Schmidt HG. Deliberate reflection and clinical reasoning: founding ideas and empirical findings. *Med Educ*. 2023;57(1):76-85. doi:10.1111/medu.14863

23. Berner ES, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. *N Engl J Med*. 1994;330(25):1792-1796. doi:10.1056/NEJM199406233302506

24. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and nominal group in medical education research. *Med Teach*. 2017;39(1):14-19. doi:10.1080/0142159X.2017.1245856

25. Mamede S, van Gog T, Moura AS, et al. Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Med Educ*. 2012;46(5):464-472. doi:10.1111/j.1365-2923.2012.04217.x

26. Mamede S, Schmidt HG. Correlates of reflective practice in medicine. *Adv Health Sci Educ Theory Pract*. 2005;10(4):327-337. doi:10.1007/s10459-005-5066-2

27. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res*. 2023;25:e50638. doi:10.2196/50638

28. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull*. 1968;70(4):213-220. doi:10.1037/h0026256

29. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297-334. doi:10.1007/BF02310555

30. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-282. doi:10.11613/BM.2012.031

31. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *arXiv*. Preprint posted online November 30, 2023. doi:10.48550/arXiv.2312.00164

32. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv*. Preprint posted online November 28, 2023. doi:10.48550/arXiv.2311.16452

33. Kostopoulou O, Rosen A, Round T, Wright E, Douiri A, Delaney B. Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. *Br J Gen Pract*. 2015;65(630):e49-e54. doi:10.3399/bjgp15X683161

34. Sibbald M, Monteiro S, Sherbino J, LoGiudice A, Friedman C, Norman G. Should electronic differential diagnosis support be used early or late in the diagnostic process? A multicentre experimental study of Isabel. *BMJ Qual Saf*. 2022;31(6):426-433. doi:10.1136/bmjqs-2021-013493

35. Kostopoulou O, Rosen A, Round T, Wright E, Douiri A, Delaney B. Early diagnostic suggestions improve accuracy of GPs: a randomised controlled trial using computer-simulated patients. *Br J Gen Pract*. 2015;65(630):e49-e54. doi:10.3399/bjgp15X683161

36. Sibbald M, Monteiro S, Sherbino J, LoGiudice A, Friedman C, Norman G. Should electronic differential diagnosis support be used early or late in the diagnostic process? a multicentre experimental study of Isabel. *BMJ Qual Saf*. 2022;31(6):426-433. doi:10.1136/bmjqs-2021-013493

37. Olson A, Rencic J, Cosby K, et al. Competencies for improving diagnosis: an interprofessional framework for education and training in health care. *Diagnosis (Berl)*. 2019;6(4):335-341. doi:10.1515/dx-2018-0107

38. Mamede S, van Gog T, van den Berge K, et al. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA*. 2010;304(11):1198-1203. doi:10.1001/jama.2010.1276

39. Staal J, Hooftman J, Gunput STG, et al. Effect on diagnostic accuracy of cognitive reasoning tools for the workplace setting: systematic review and meta-analysis. *BMJ Qual Saf*. 2022;31(12):899-910. doi:10.1136/bmjqs-2022-014865

40. Schaye V, Miller L, Kudlowitz D, et al. Development of a clinical reasoning documentation assessment tool for resident and fellow admission notes: a shared mental model for feedback. *J Gen Intern Med*. 2022;37(3):507-512. doi:10.1007/s11606-021-06805-6

41. Omega A, Wijaya Ramlan AA, Soenarto RF, Heriwardito A, Sugiarto A. Assessing clinical reasoning in airway related cases among anesthesiology fellow residents using Script Concordance Test (SCT). *Med Educ Online*. 2022;27(1):2135421. doi:10.1080/10872981.2022.2135421

42. Groves M, Dick ML, McColl G, Bilszta J. Analysing clinical reasoning characteristics using a combined methods approach. *BMC Med Educ*. 2013;13(1):144. doi:10.1186/1472-6920-13-144

43. Nori H, King N, Mckinney SM, Carignan D, Horvitz E, Openai M. 2. Capabilities of GPT-4 on medical challenge problems. *arXiv*. Preprint posted online March 20, 2023. doi:10.48550/arXiv.2303.13375

44. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

45. Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv*. Preprint posted online November 28, 2023. doi:10.48550/arXiv.2311.16452

46. Harden RM. What is an OSCE? *Med Teach*. 1988;10(1):19-22. doi:10.3109/01421598809019321

47. Pell G, Fuller R, Homer M, Roberts T; International Association for Medical Education. How to measure the quality of the OSCE: a review of metrics—AMEE guide no. 49. *Med Teach*. 2010;32(10):802-811. doi:10.3109/0142159X.2010.507716

48. Khan KZ, 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*. 2013;35(9):e1437-e1446. doi:10.3109/0142159X.2013.818634

49. Chan SCC, Choa G, Kelly J, Maru D, Rashid MA. Implementation of virtual OSCE in health professions education: a systematic review. *Med Educ*. 2023;57(9):833-843. doi:10.1111/medu.15089

SUPPLEMENT 1.

Study Protocol

SUPPLEMENT 2.

eTable 1. Diagnostic Case #1 Vignette and Questions
eTable 2. Structured-Reflection Rubric for Diagnostic Case #1, With Example of high-Scoring vs Low-Scoring Example Responses
eFigure 1. Distribution of Diagnostic Performance Scores of Physician + GPT-4 vs Physician + Conventional Resources Only
eTable 3. GPT-4 Performance
eTable 4. GPT Prompt and Responses for Diagnostic Case 1
eTable 5. Assessment Tool Validation
eFigure 2. Visual Flow Diagram

SUPPLEMENT 3.

Data Sharing Statement