npj Digital Medicine

https://doi.org/10.1038/s41746-026-02545-1

Article in Press

From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis

Received: 6 July 2025

Accepted: 4 March 2026

Cite this article as: Everett, S.S., Bunning, B.J., Jain, P. *et al.* From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02545-1

Selin S. Everett, Bryan J. Bunning, Priyank Jain, Ivan Lopez, Anup Agarwal, Manisha Desai, Robert Gallo, Ethan Goh, Vinay B. Kadiyala, Zahir Kanjee, Jacob M. Koshy, Andrew Olson, Adam Rodman, Kevin Schulman, Eric Strong, Jonathan H. Chen & Eric Horvitz

From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis

Selin S. Everett[1,2], Bryan J. Bunning[3,4,5], Priyank Jain [6,7], Ivan Lopez[1,2,3], Anup Agarwal[6,7], Manisha Desai[5,7], Robert Gallo[9], Ethan Goh[2], Vinay B. Kadiyala[6,10], Zahir Kanjee [6,10], Jacob M. Koshy[6,10], Andrew Olson[8], Adam Rodman[6,10], Kevin Schulman[11], Eric Strong[11], Jonathan H. Chen[2,3,4,11†], Eric Horvitz[12,13†*]

[1]Stanford University School of Medicine, Stanford, CA, USA
[2]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
[3]Stanford Department of Biomedical Data Science, Stanford, CA, USA
[4]Stanford Clinical Excellence Research Center, Stanford University, Stanford, CA, USA
[5]Stanford Quantitative Sciences Unit, Stanford, CA, USA
[6]Harvard Medical School, Boston, MA, USA
[7]Department of Medicine, Cambridge Health Alliance, Cambridge, MA, USA
[8]University of Minnesota Medical School, Minneapolis, MN, USA
[9]Center for Innovation to Implementation, VA Palo Alto Health Care System, Palo Alto, CA, USA
[10]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
[11]Division of Hospital Medicine, Stanford University, Stanford, CA, USA
[12]Office of the Chief Scientific Officer, Microsoft, Redmond, WA, USA
[13]Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, Stanford, CA, USA
*Corresponding author: horvitz@microsoft.com
†Co-senior authors

Abstract

Early studies of large language models (LLMs) in clinical settings have largely treated artificial intelligence (AI) as a tool rather than an active collaborator. As LLMs demonstrate expert-level diagnostic performance, the focus shifts from whether AI can offer valuable suggestions to how it integrates into physicians' diagnostic workflows. We conducted a randomized controlled trial (n=70 clinicians) to assess a custom system designed for collaborative diagnostic reasoning. The design involved independent diagnostic assessments by the clinician and AI, followed by an AI-generated synthesis integrating both perspectives, highlighting agreements, disagreements, and offering commentary. We evaluated two collaborative workflows: AI as first opinion (preceding clinician) and AI as second opinion (following clinician). Both improved clinician diagnostic accuracy over conventional resources, (85% and 82% vs. 75%). Performance was comparable across workflows and not statistically different from AI-alone accuracy (90%), highlighting the potential of collaborative AI to complement clinician expertise. Qualitative analyses illustrate how workflow design shapes human-AI interaction. ClinicalTrials.gov: NCT06911645.

Introduction

The use of large language models (LLMs) to support physicians' diagnostic reasoning remains in its infancy. Evaluations of GPT-4 on medical challenge problems have demonstrated that LLMs can reach near-expert performance on standardized competency exams such as the United States Medical Licensing Examination (USMLE), fueling growing interest and debate about their potential readiness for real-world clinical use.[1,2]

This study was sparked by unexpected findings in Goh and Gallo et al. that showed physicians using an LLM perform significantly worse than the LLM alone.[3,4] Insufficient skill and experience with prompting LLMs was proposed as an explanation for the poor AI-assisted human performance. The authors highlighted that, “development in human-computer interactions is needed to realize the potential of AI in clinical decision support.”[3] We took that charge as the starting point for our investigation, focusing on how system prompting and workflow design could instantiate structured workflows and interactive protocols that reshape a general-purpose LLM into a collaborative decision-support system.

As a central design consideration, we examined how changing whether the clinician or the AI model makes the initial assessment in generating diagnoses and management steps might influence the quality of diagnostic reasoning. We considered results from cognitive psychology studies that document biases in judgment and decision making, particularly *anchoring* effects when subjects are primed with information prior to their own assessments.[5] Such findings highlight the potential influence of clinical workflow sequencing in AI-assisted diagnosis. Recent studies in medical informatics have identified such anchoring effects, including in radiology workflows in which AI inferences were presented before radiologists’ assessments.[6-8]

Human factors research on human’s tendency to place inappropriate trust in automation[9] also informed our approach. Designs of AI for high-stakes domains such as medicine must account for risks of *automation bias* and of clinicians becoming overreliant on AI, which can lead to acceptance of erroneous AI inferences, erosion of independent reasoning, and clinician deskilling.[10-17] Concerns about overreliance motivate mitigation strategies, such as techniques that provoke users’ critical thinking.[18]

Additional insights about effective human-AI interaction were drawn from studies of collaboration in cognitive psychology, particularly from research on how collaborators establish mutual understanding or “grounding,” [19] and from related work that has directly examined human-AI grounding.[20-23]. Related work has also examined the mental models people hold about how AI systems operate, with the goal of enhancing human-AI performance.[6,24]

We also drew on methods and findings from the emerging field of research focused on developing mechanisms to explicitly support human-AI collaboration. [6,25-29] This work includes design guidelines for human-AI collaboration, providing a broad set of design goals around transparency and control;[15,26] principles of mixed-initiative interaction[25,30] regarding when and how to interleave AI into human problem solving; research on AI intelligibility and the construction and evaluation of explanations of AI inferences;[13,31-33] and methods for identifying and leveraging the complementary capabilities of humans and AI systems, focusing on the relative strengths and weaknesses of people and AI for the tasks at hand. [27,34-37]

These strands of prior work motivated our exploration of clinician-AI collaboration designs that jointly consider the clinicians' and the AI model's reasoning, provide syntheses of the two perspectives, and support reflection and discussion on the way to convergence. Such a strategy is in contrast to the use of AI models to solely generate recommendations, the most prevalent approach to deploying AI for medical decision support. In this spirit, we examined work on AI in medical decision support in which AI is harnessed to critique clinicians' reasoning and care plans rather than to provide its independent reasoning about the case at hand. [38,39]

Building on these ideas, we designed a custom GPT-4 system for collaborative diagnostic reasoning and evaluated alternative clinician–AI workflows in a randomized controlled trial.

In this study, the custom GPT-4 system was guided by a system-level prompt to support clinician-AI collaboration on diagnostic reasoning. We did not perform fine-tuning of the model; all customization was done through development of a custom, multi-part prompt. The system prompt is provided in **Supplementary Figure 1**.

The custom GPT is instructed to provide structured clinician-AI workflows, in which the clinician and AI each independently analyze a case, followed by integration and critique of their respective perspectives. We explored two workflow variants: an *AI-as-first-opinion* workflow, where the clinician sees the AI's output on differential diagnosis and next diagnostic steps for a given case before offering their own analysis, and an *AI-as-second-opinion* workflow, where the clinician first provides their answers before reviewing the AI output.

Regardless of whether the clinician or the AI analyzes the case first, the system subsequently generates and makes available an integrated *synthesis view* that highlights areas of agreement and disagreement and provides a critique of each of the inferred diagnoses (see **Figure 1**). Following the display of the integrated view, clinicians are invited to engage in open dialogue with the system on any aspect of the independent analyses or the synthesis.

The custom GPT provides a set of capabilities not available in the standard ChatGPT interface. The system-level prompt shapes the GPT into a collaborative actor with a workflow aimed at promoting human-AI grounding and provoking critical thinking via explicit comparison and consideration of the clinician's and the AI model's inferences. In addition to explaining its own reasoning, the AI model critiques the clinician's analysis and invites open-ended dialogue to support reflection on differing perspectives.

Results

71 clinicians participated in the study. One participant was removed from analysis because of exposure to the same clinical vignettes via participating in a prior study. 70 U.S.-licensed physicians were included in the analysis, 39 residents and 31 attendings. Our trial population consists almost entirely (97%) of internal medicine specialists. Groups were balanced with

respect to specialty and level of training. Regarding generative AI experience, in the AI-as-second-opinion arm, 42% used generative AI frequently, 33% occasionally, 16% rarely, and 9.3% had never used it; in the AI-as-first-opinion arm, 33% used generative AI frequently, 26% occasionally, 33% rarely, and 7.4% had never used it (**Table 1**).

Full transcripts were collected from each participant. 6 of 70 participants (8.6%), all from the AI-as-second-opinion arm, did not adhere strictly to the procedural instructions for every case attempted (e.g., copying and pasting only the vignette rather than their responses into the AI, or failing to initiate a new transcript for each case). Their data were retained in the analysis in accordance with the intention-to-treat principle.

298 cases were captured from 71 clinicians' interactions with the AI system. We excluded six cases completed by the participant who had previous exposure to the clinical vignettes. 38 unfinished cases were also excluded (**Figure 3**). The final primary analysis set included 254 cases from 70 participants (**Table 1**): 108 cases for AI-as-first-opinion (27 participants, 4 average cases completed per person) and 146 cases for AI-as-second-opinion (43 participants, 3.4 cases completed per person). The intraclass correlation coefficient (ICC) computed between graders' scores was 0.91 indicating very high agreement.

Use of AI significantly boosts clinician performance

Clinicians using conventional resources only had a significantly lower overall score (75%) than either AI arm, with the clinicians' using AI as a first opinion (85%, p=0.00039, mean difference=9.9%, 95% CI= 4.7%-15%), or AI as a second opinion (82%, p=3.3e-6, mean difference=6.8%, 95% CI=4.0%-9.6%). AI alone was not significantly different (87%, p=0.20), but its average score was numerically highest among all groups **(Figure 4, Figure 5)**.

No difference in overall performance for the two AI workflows

After controlling for case and clinician variability, we found that scores for AI as a first opinion were not significantly different from AI as a second opinion (p=0.22, mean difference =3.0% favoring 1st opinion, 95% CI -1.72%-7.63%) **(Figure 4, Figure 5).** Sensitivity analysis showed that clinicians' number of years of experience was not significant and did not change the model.

AI as first opinion is superior to AI as second opinion for clinically actionable decisions

There was a significant difference in the scores for the clinically actionable decisions (final diagnosis and next steps sections of the quiz shown in **Supplementary Figure 2, Parts 2 and 3**). Clinicians in the AI-as-first-opinion arm scored 8.9% better than AI-as-second-opinion arm (p=0.026, 95% CI 1.1%-16%). There was no significant difference between arms in the scores derived from the subsection seeking evidence for and against each assessed diagnosis (P=0.74, Mean score difference= 0.37%, 95% CI=-2.0%-2.6%) (**Supplementary Figure 2, Part 1)**.

AI as second opinion boosts clinically actionable decisions over conventional resources alone

We found that there was a significant 14.9% (p=0.00092, 95% CI 6.0%-23%) increase in the score related to the clinically actionable decisions for the AI-as-second-opinion arm versus clinicians using only conventional resources (**Figure 6**)**.** In the AI-as-second-opinion arm, the actionable score increased in 52 cases. Of these 52, 8 cases increased by 4 points, 5 cases increased by 3 points, 16 cases increased by 2 points and 23 cases increased by 1 point. Comparatively, the actionable score decreased in 12 cases. In 11 cases it decreased by 1 point and in 1 case decreased by 2 points. A total of 83 cases had no change in score.

Faster case times observed when AI goes first

The time spent on each case between the two clinician-AI arms was not significantly different (p=0.11, mean difference=57 seconds), with the mean time of the AI-as-first-opinion arm, 631 seconds, being slightly faster than the AI-as-second-opinion arm, 688 seconds. In a post hoc per-protocol analysis, after removal of all cases from 6 participants who did not adhere to the study instructions (all in AI-as-second-opinion arm), this difference became statistically significant with mean time in the AI-as-first-opinion arm remaining 631 seconds, but the AI-as-second-opinion arm increasing to 723 seconds (p=0.016, mean difference=92 seconds).

Workflow influences physician-AI dialogue

After submitting a case and reviewing the AI analysis, participants could interact in an unstructured way with the AI system to ask questions about AI output or to engage in open explorations. Clinicians did not interact with the AI system beyond the required input of the vignettes and assessments for 3 of 27 (11%) in the AI-as-first-opinion arm and 14 of 43 (33%) in the AI-as-second-opinion arm. Despite refraining from interaction with AI beyond the required input, all 14 of the AI-as-second-opinion participants still made a change to their answers after review of the AI inferences.

Qualitative coding of participant transcripts (**Supplementary Figure 4A**) revealed that workflow sequencing shaped both the frequency and style of AI engagement. Participants in the AI-as-second-opinion more frequently used anthropomorphizing language, including prompts like, *“Yes, that is a great thought”*, indicating a more humanized, conversational stance toward the system. They also expressed gratitude and effective responses including, *“Thanks for your help!”*, *“Looks great, thanks!”* more often than AI-as-first-opinion participants (**Supplementary Figure 4B**).

Clinicians are more open to using AI in clinical reasoning after hands-on experience

Our before-and-after survey asked participants about their beliefs and attitudes toward AI. After using the AI tool, participants were significantly more likely to agree with the statement, “I am open to using AI to help with complex clinical reasoning,” compared to before using the tool (99% vs. 91%, p = 0.011). For both arms (AI-as-first-opinion, AI-as-second-opinion), the vast majority of participants enjoyed working with the tool (96%, 98%), agreed that the tool provided a valuable collaborative experience (100%, 95%), would use the tool in their daily job (96%, 95%), and agreed that seeing the AI tool's recommendations increased their confidence in their differential (96%, 97%) (**Supplementary Table 1**).

AI anchoring on clinician Input

In our post hoc analysis, in a random sample of 58 cases (29 AI-as-first-opinion and 29 AI-as-second-opinion, matched by vignette), we noted complete overlap in 3% of the AI-as-first-opinion cases, where all three of the clinician's initial diagnoses appeared in the AI's independent output. In contrast, we saw such complete overlap in 48% of AI-as-second-opinion cases. Similarly, we noted complete overlap in recommendations on next steps in 24% of AI-as-first-opinion cases and in 52% of AI-as-second-opinion cases (**Supplementary Figure 5, Supplementary Table 2**).

Discussion

Because this work represents an exploratory, early-stage evaluation conducted using structured vignettes rather than real clinical encounters, the observed effects should be interpreted as hypothesis-generating rather than confirmatory. Future studies in clinical environments will be required to assess whether similar dynamics arise in practice.

The use of AI was associated with performance gains over the use of conventional resources. Clinicians using only conventional resources achieved a mean score of 75%. In contrast, clinicians employing the collaborative workflow achieved a mean score of 85% (9.9% increase) for AI-as-first-opinion and 82% (6.8% increase) for AI-as-second opinion. We found no significant difference in the overall performance of participants between the AI-as-first-opinion and AI-as-second-opinion arms, which may reflect limited precision to detect subtle between-arm differences between the two AI workflows. The observed difference was ~3% and we were powered to detect a 10% difference. Taken together, these findings suggest that, across workflows, the collaborative system can reliably improve diagnostic performance relative to conventional tools.

In this vignette-based, controlled setting, AI support raised the floor of diagnostic performance by reducing particularly low-scoring cases. Whether used as a first opinion or as a second opinion, AI compressed the lower tail of the case score distribution and shifted it upward (**Figure 5**). This suggests that AI can help clinicians avoid their lowest-performing diagnoses, even if the average score remains similar. The benefit of collaborative interaction with the LLM was most pronounced in the assessments of final diagnosis and next steps (**Figure 6**), the portions of the case that represent clinically actionable decisions (**Supplementary Figure 2, Parts 2 and 3**). We found that performance significantly improved for actionable decisions in both workflows, with scores notably higher in the AI-as-first-opinion arm. In the AI as second-opinion arm, 36% (n=52) of cases demonstrated improvements in scores on the clinically actionable decisions, following engagement with the AI (**Figure 6**).

We note that AI did not uniformly enhance clinician's performance. Scores on clinically actionable decisions *decreased* after engagement with the AI in 8% of cases, highlighting the possibility that AI recommendations can interfere with performance. While early assessment of AI performance is focused on efficacy of the technology, our understanding of the safety of these tools is nascent. A safety assessment will require estimating the frequency and severity of adverse events (such as decreases in clinical performance) and considering their likelihood of causing patient harm. Decisions about deploying and using AI tools, such as the custom GPT we have studied, would be better informed by such risk–benefit assessments.

Our prior study, Goh and Gallo et al. (2024), showed no significant improvement of physician performance with the use of the unmodified LLM.[3] In contrast, use of the custom collaborative GPT significantly increased the performance of clinicians' diagnostic capabilities (9.9% in AI-as-first-opinion arm and 6.8% in AI-as-second-opinion arm). The two studies used identical vignettes, case structure, and scoring scheme. Comparing the studies contextualizes our controls and improvements, focused on identifying the value of a collaborative workflow. In both studies, clinicians using only conventional resources had nearly identical baseline scores, a median of 74% in the previous study versus 75% in this study. Additionally, our AI-alone performance (median score of 89.5%) was similar to the previous study's AI-alone benchmark (median score of 92%), despite differences in expert graders, the specific prompt used, stochastic variability in LLM outputs, and the potential impact of intervening model updates. Together, these results underscore that gains in clinician performance in the current trial are attributable not to improvements in base model capability but to tailoring the LLM for clinical collaboration.

Our results suggest different workflow challenges may result based on two forms of anchoring bias: (1) clinicians anchoring on AI responses in the AI-as-first-opinion arm and (2) AI anchoring on the clinician's input in the AI-as-second-opinion. On the former, we hypothesize that AI inferences are more persuasive when they are presented before the clinician completes their assessments, similar to the anchoring effect identified in a prior study in radiology.[6] On the latter, we noted that the clinician's initial input can interfere with the independent reasoning of the LLM in the AI-as-second-opinion workflow (**Supplementary Figure 5**). Each of the identified anchoring biases, one by clinicians and one by AI, will require different strategies to address in the context of specific workflow designs.

Our findings, resonating with prior work on the sycophancy of LLMs[40], highlight a consideration for interaction design with many consumer LLMs: providing clinician input before soliciting the AI's assessment may compromise its intended independent analysis. We found that, even when explicitly instructed to disregard the clinician's earlier input, the model was at times influenced by human input. Our findings on anchoring of AI inferences when exposed initially to human judgments, analogous to the phenomena of human anchoring, may reduce the AI's ability to provide complementary inferences.

Studies have demonstrated that GPT-4 can seek to be likable,[41] sometimes at the expense of truthfulness or epistemic rigor. This sycophantic behavior has been linked to reinforcement learning with human feedback (RLHF),[40] a training procedure that rewards outputs preferred by human evaluators. Because agreement and flattery may tend to be rated more favorably, models can prioritize agreeing with users rather than challenging them when they are incorrect.[42]

Our findings reinforce the perspective that developing AI systems for clinical practice will require efforts to optimize not only the accuracy of the system, but innovations with human-computer interaction design. Given the strong diagnostic performance of today's LLMs alone, we see opportunities for further innovation in the realm of clinician-AI interaction. Deficits in human-AI collaboration for medical applications may be a key limiting factor in mainstream adoption of AI for diagnostic decision support, even when model performance is high.

We also see great opportunities ahead for leveraging advances in core capabilities of AI methods and findings in ongoing work in human-AI collaboration. For example, future effective designs for clinician-AI collaboration could benefit from advances that enable AI systems to infer and share well-calibrated confidences.[1,42] This would support formal and qualitative syntheses of AI and human expertise by explicitly considering the diagnostic confidence of both clinicians and the AI.

We found that clinicians assisted by AI did not perform better than AI alone. However, we see promise in enabling "better-together-than-alone" performance via collaborative interfaces that leverage the complementary skills of the AI system and clinicians. Bejnordi et al. (2017) found that an AI system performed well on large-scale image analysis in radiology, while humans remained better at edge cases involving ambiguity, context, or prior experience.[37] Wilder et al. (2020) employed a machine learning procedure to characterize the capabilities and performance gaps of human pathologists in detecting metastatic breast cancer in lymph node tissue.[27] They employed a methodology to boost the discriminatory power of the predictive model so it would be most effective where humans are weakest, and to learn when to inquire about or defer to human expertise in generating diagnostic recommendation. These examples illustrate how explicitly modeling complementarity can enable human-AI teams to outperform either alone.

The degree and nature of complementarity can depend on the particular expertise of the clinicians deploying the system.[40] We note that considerations of complementarity will be sensitive to the march of progress in AI capabilities. Regardless of AI capabilities today and in the future, the handling of complementarity and the sequencing of AI and clinician initiative and attention will also be influenced by policies and goals regarding the primacy of human responsibility and agency in medical decisions.

Our qualitative analyses showed that clinicians across both workflow arms reported exceptionally high satisfaction, perceived collaborative value, and confidence gains from using the tool, indicating strong enthusiasm for integrating LLM-based decision support into clinical reasoning. These findings suggest that clinicians are motivated to consult and respond to the AI's input regardless of whether the AI presents its opinion first or second.

These findings of our exploratory study suggest that workflow order may influence how clinicians relate to AI. Our findings suggest that workflow order influences how clinicians relate to AI. When the AI was positioned as a second opinion, participants were more likely to adopt a conversational, even humanized, stance toward the system. They frequently used anthropomorphizing language (e.g., “Yes, that is a great thought”) and expressed gratitude (“Thanks for your help!”, “Looks great, thanks!”), suggesting an effective engagement that extended beyond transactional tool use (**Supplementary Figure 4**). These nuanced behavioral and linguistic differences with different workflows may have downstream consequences for the calibration of trust, diagnostic reasoning dynamics, and other detail of clinician-AI teaming[3,6]. Future design choices should consider how workflow design and sequencing shape overall human-AI performance, as well as clinicians’ autonomy, appropriate trust, and engagement in collaboration with an AI system.

When comparing participants’ prior use of LLMs, 32% of participants in the prior study (2024) were first-time users of generative AI compared to 8% of our participants in this study (2025). This change over one year demonstrates that AI use has become more widespread, though it is not yet universal–even among those open to participating in AI studies. This highlights a key difference between the two study populations: participants in the 2024 study were less experienced using LLMs than those in this 2025 study. Baseline familiarity with generative AI could affect external validity of our results.

We identified key limitations pertaining to the AI system itself that interfered with the intended design for collaborative interaction. The custom prompted GPT was designed to produce structured joint reasoning, including summaries of agreement and disagreement and critiques of each diagnostic hypothesis, while also enabling free-form discussion following an initial phase of assessments by AI and clinician. Our transcript review revealed that in 10% of AI-as-second-opinion cases, AI omitted displaying the joint analysis (see example of the synthesis view, shown in **Figure 2**). These system failures diminished the intended collaborative functionality and point to the need for system-level mechanisms to ensure prompt fidelity or detect and address deviations. LLMs producing an incomplete or incorrect response remains a limitation of the current technology’s grounding ability. For example, one participant in the AI-as-first-opinion group noticed that the AI system claimed the patient in Case 1 (Supplementary Figure 3) had a fever when they had a temperature of 99.6F (a fever is defined by the CDC as ≥100.4 °F).

> **User said:**
> *“The vignette does not say the patient had a fever”*
>
> **ChatGPT said:**
>
> *“You're absolutely correct—thank you for pointing that out! The vignette describes the patient as feeling **febrile and tired**, but the measured temperature is **99.6°F**, which is technically **not a fever**. This is an important distinction because it refines our interpretation of the systemic symptoms.”*

Future work should investigate physicians’ reactions to these errors and their implications for physician-AI interactions and trust.

We also observed non-determinism, whereby the AI system could generate different recommended diagnostic assessments across participants for the same vignette input, raising concerns about reliability and replicability in clinical decision support. Recent evaluations of LLMs consider such stochasticity as a measured metric used to evaluate the reliability of different models.[44] Such variability and unpredictability are problematic in clinical settings, where consistency and replicability are essential. We hypothesize that improvements and methods development into the prompting, guardrails, or response template for the model synthesis may improve upon this technical limitation and is a rich avenue for future work.

Another limitation of the study stems from the different numbers of answers generated by the LLM versus those requested as answers from the clinicians on the case. We had initially designed the system to provide seven next steps, versus the three next steps required by the case instructions. While this design decision was implemented to broaden clinical thinking, it created ambiguity when scoring the AI-alone condition because the system prompt did not instruct the AI to rank its suggested next steps, as it did for the differential. For scoring purposes, we assumed the first three steps listed by the LLM were its top recommendations.

A further limitation within our study stems from the six clinical vignettes designed to simulate diagnostic reasoning tasks within a one-hour timeframe, consistent with standard exam formats. Although this format is convenient and controlled, vignette-based cases are not representative of real-world clinical practice, where clinicians must actively gather information through history-taking, physical exams, and test selection—core components of diagnostic reasoning that were assumed, not assessed, in this study. Furthermore, because LLMs are often trained on text corpora that include structured vignettes, this format may confer an unintended advantage to the AI compared to real-world encounters. Finally, our restricted case selection may limit generalizability, as model performance could differ on a broader set of typical presentations or on atypical, diagnostically challenging cases.

Sampling bias, particularly in interpreting clinicians' attitudes toward the AI tool, could have played a significant role in our qualitative analysis of participant attitudes toward AI. Additionally, familiarity with AI and level of AI literacy within our population may threaten generalizability, as it may be higher than in the general population of clinicians. The observed high levels of acceptance may be partially attributable to self-selection effects: clinicians who voluntarily participate in a diagnostic reasoning study involving AI are likely to hold pre-existing favorable views toward such technologies.

Overall, this trial should be understood as an exploratory study, intended to characterize workflow-dependent dynamics and the influence of a design for clinician-AI collaboration rather than as producing definitive evidence for clinical benefit.

Given the early and rapidly evolving understandings of the value of generalist LLMs in medical decision support, our exploratory study aimed to provide initial evidence on how workflow design may shape diagnostic reasoning. We found that moving from a consumer LLM to a GPT customized to foster collaboration can meaningfully improve clinician diagnostic accuracy in a setting of clinical vignettes. The study provides evidence of unrealized opportunities at the intersection of design, engineering, and medicine for enhancing clinician-AI collaboration. Additional cross-disciplinary research and methods development are needed to realize the possibilities. As AI continues to evolve, a key question is not whether AI will replace clinicians, but how clinicians and AI can best work together to boost human learning, deliberation, efficiency, and decision-making prowess, and ultimately and most importantly, to enhance patient outcomes. Our findings contribute to the broader conversation on the evolving role of AI in medicine.

Methods

We conducted a randomized controlled trial to evaluate clinicians' diagnostic performance when using the custom GPT system for diagnostic decision-making. Our primary aim was to assess clinicians' diagnostic performance on a set of clinical vignettes under two workflow conditions: AI-as-first-opinion and AI-as-second-opinion. In both conditions, clinicians subsequently had access to a synthesis view that integrated the clinician's and AI's inferences. In the AI-as-second-opinion workflow, the synthesis view was displayed immediately following the AI's recommendations, whereas in the AI-as-first-opinion arm it was made optionally available after clinicians had reviewed the AI output. A secondary aim was to examine how the interactive design and workflows influenced clinicians' engagement with the AI and their attitudes toward its role in diagnostic decision-making.

We enrolled participants from December 16, 2024 to January 24, 2025. The inclusion criterion was being a U.S.-licensed internal medicine or family medicine physician. The exclusion criterion was participation in any previous study using the clinical vignettes used in our study. We recruited attending and resident physicians via our networks and email lists at Stanford University, Beth Israel Deaconess Medical Center, Vanderbilt University Medical Center, New York-Presbyterian Hospital, and Cambridge Health Alliance. Resident participants were offered $100, and attending participants were offered $199 for completing a one-hour session. Participants joined remotely in small groups (≤ 5). The same study team member (S.E.) facilitated each session, and randomization occurred at the session level. The randomization sequence was generated by B.B. Participant flow is illustrated in **Figure 2.**

Participants were encouraged to critically evaluate, challenge, or refine the AI model's diagnostic conclusions as needed and were reminded that the system outputs could contain errors or gaps. Participants were instructed to start a new AI dialogue for each diagnostic reasoning case to clear the context under consideration by the LLM. Participants completed as many cases as possible in the hour, with instructions to prioritize quality over quantity.

The study was conducted via a secure survey platform (Qualtrics). At the start of each session, the facilitator provided a brief, live demonstration of the case workflow to clearly show that the vignette was to be entered into the AI system via a copy-paste operation. For the facilitator's full script used in each session, please refer to the standard operating procedures included in **Supplementary Figure 6**. Support from the facilitator was available throughout the session. All participants evaluated a set of up to six clinical vignettes, with their order randomized. The evaluation format, response structure, and case vignettes were identical to the approach presented previously.[3] Unfinished cases in the study period were excluded. A case was considered unfinished if the following were both true: (1) it was the final vignette in a session, and (2) the participant spent less than 30 seconds reviewing information.

The vignettes are based on real patients. The content was deidentified and included information about history, physical examination, and laboratory test results (**Supplementary Figure 3).** The cases have not been publicly released and therefore are excluded from the training data of the LLM. A brief description of each case, outlining the age, sex, and organ systems involved, is included in **Supplementary Table 1**.

Two workflows

The custom GPT provided two distinct workflows, depending on the arm a clinician was randomized to. Of note, the system was intentionally designed to broaden clinician reasoning, offering five differential diagnoses rather than the three requested in the instructions, and to suggest seven management steps instead of three. The model was instructed to rank the differential diagnoses by likelihood. Case response structure is shown in **Supplementary Figure 2**.

For the AI-as-first-opinion arm, physicians began by inputting the full case and had the opportunity to use the AI system from the outset to generate ideas, interpret information, and construct their responses. In the AI-as-second-opinion, physicians initially worked independently, forming their own assessments with access to conventional (non-AI/LLM) resources such as UpToDate, PubMed, or Google. Participants were instructed to paste the vignette along with their initial answers into the system to initiate a collaborative interaction. After reviewing the output of AI as a second opinion, clinicians were required to either revise or retain their original responses. After completing the initial stage of the structured workflow and viewing the integrated summary, participants were invited by the system to engage in free-form dialogue before submitting their final answers.

Baselines for clinicians and AI. To establish a baseline for clinician performance, we scored the initial responses provided by clinicians in the AI-as-second-opinion arm. For a baseline assessment of AI-alone performance, we analyzed the AI's initial responses in the AI-as-first-opinion arm. For each of the six vignettes, five separate responses were randomly selected to yield 30 cases for analysis. For scoring, the AI's top three diagnoses, supporting and opposing evidence, and first three proposed next steps were taken as its answers to the case.

Diagnostic accuracy scoring and rubric

For each case, two internal medicine board-certified physician scorers (J.K., V.K., A.A.) graded the responses using a previously established 19-point scoring schema.[3] The rubric evaluates clinical reasoning based on the plausibility of differential diagnoses, the appropriateness of supporting and opposing findings, accuracy of the final diagnosis, and the relevance of proposed next steps, with assessments made by our physician scorers using a standardized rubric.

Scorers assigned up to one point for each plausible differential diagnosis. Assessments of findings that support and oppose the diagnosis were graded based on being clinically reasonable, with zero points for incorrect or absent answers, one point for partially correct, and two points for completely correct responses. The final diagnosis was graded as two points for selecting the most correct diagnosis, or one point for a plausible diagnosis or a diagnosis that was not incorrect, but not specific in matching the most correct final diagnosis. The participants were instructed to describe up to three next steps to further evaluate the case. Zero points were awarded for an incorrect response, one point was awarded for a partially correct response, and two points were awarded for the correct response. The physician graders were given the rubric with sample answers and were asked to use their expert judgment on plausibility and correctness of other answers.

Scorers were blinded to participant group assignments as well as the 30 AI-alone cases. For scoring the responses to the diagnostic reasoning cases, the identity of the source of responses (clinician versus AI) was concealed during the blinded scoring process, ensuring consistency and minimizing bias prior to unblinding or analysis.[3] With the exception of the final diagnosis, there were no predefined correct answers, and correctness was left to the graders' expert opinion. To ensure consistency relative to inter-rater reliability, graders reconciled their scores if there was a difference greater than two points between their assessments.

Dialogue analysis

We characterized clinician engagement with the AI system across different workflows by systematically applying qualitative codes to participants' free-form contributions to the conversation. All clinician-entered text and AI responses were captured from the transcripts of the dialogue following each study session. We developed a post-hoc coding framework based on the language and structure observed in the clinician-AI conversations **(Supplementary Figure 4A)**. All example prompts in **Supplemental Figure 4A** are excerpts from our participants' interactions with the AI system.

Clinician perception of AI

After providing assessments for the diagnostic cases, clinicians were asked to rate their experiences across multiple dimensions, including enjoyment, perceived collaboration, confidence, and willingness to use such a tool in clinical practice **(Supplementary Table 2)**. Additionally, participants completed pre- and post-study surveys measuring their openness to using AI for complex clinical reasoning.

Exploration of AI anchoring on clinician input

Motivated by the facilitator's observation of the transcripts, we conducted an exploratory post hoc analysis to assess the potential influence of clinicians' initial responses on AI's "independent" analysis of the vignette in the AI-as-second-opinion arm. In particular, we saw evidence in the transcripts suggesting that the AI system disobeyed instructions to generate output independent of the clinician's input, as specified in the system prompt **(Supplementary Figure 1):**

> "*Start by reviewing the full patient case and conducting your independent analysis… BEFORE making any consideration of the physician's input information that came via the input of their assessments…"*.

During the customization process, we explicitly included this instruction with the goal of ensuring the AI system maintained independence before merging the clinician's and AI assessments into a singular view. If there was true independence between the initial clinician's responses and the AI analysis that followed right after in AI-as-second-opinion workflows, we would expect similar responses from the AI model in both arms. We compared the overlap of AI responses in the two arms with the clinicians' initial responses obtained from the AI-as-second-opinion arm **(Supplementary Figure 5)**.

To evaluate whether clinicians' submitted assessments influenced the AI's so-called "independent" analyses, we retrieved paired clinician reasoning and AI outputs from cases in both workflows. We randomly selected a subsampling of 29 clinician–case encounters from each study arm (14 from one case, 15 from another), yielding 58 encounters in total: 29 AI-as-first-opinion and 29 AI-as-second-opinion. For each encounter we retrieved the AI's independent analysis, as well as the clinicians' own diagnostic reasoning in the AI-as-second-opinion arm.

We manually retrieved differential diagnoses and next-step recommendations from the clinician and AI independent analyses for our evaluation. For each clinician-generated item, we used GPT-4o to determine whether an identical or semantically equivalent phrase appeared in the corresponding AI output (prompts are shown in **Supplementary Table 2**). The number of overlaps between the human and AI's differential and next steps is shown in **Supplementary Figure 4.**

Statistical analyses

The primary outcome, the total graded vignette score between the two AI-enabled workflows, was determined with a linear mixed-effects model with a random intercept to account for variation between participants, and fixed effects to account for variation between cases.

$$\text{Total Score} = \text{Arm} + \text{Case} + (1|\text{participant})$$

To evaluate power, a target sample size of 225 cases was prespecified based on simulation using data and variances seen in the Goh and Gallo et al. (2024) study. A total of 225 cases would provide >80% power to show a 10% difference between treatment arms in the overall score of the vignette. Power was calculated at the case level, not the participant level. As a sensitivity analysis, we included the level of training of the participants in the model.

Secondary outcomes included (1) difference in total score compared to clinicians using conventional resources, (2) time spent on each case, (3) change in score of the evidence section of the vignette (supporting and opposing evidence) (**Supplementary Figure 2, Part 1)**, and (4) change in score of the *clinically actionable* decisions, which we define as the portion of the quiz on the final diagnosis and next steps (**Supplementary Figure 2, Parts 2 and 3**). Differences in scores compared to using conventional resources were assessed using the linear mixed effects model described above.

Descriptive statistics were used to summarize the diagnostic accuracy scores and the transcript analyses results. All analyses followed the intention-to-treat principle and were conducted at the case level unless otherwise noted.

Transcript analysis included the frequency with which users engaged in additional interactions across arms including the number of prompts issued per case and, for the participants in the AI-as-second-opinion arm, which parts of the case they altered after review of AI inferences (the diagnoses with supporting and opposing evidence for each, final diagnosis, and next steps). This enabled us to identify patterns in how clinicians formulate queries and respond to AI inferences.

Sections of the case and time spent outcomes were assessed using a t-test with a significance threshold of 0.05. For the anchoring analysis, the proportion of cases showing full overlap, as a percent, is reported. For the perception analysis, Likert-scale responses, grouped by arm, were analyzed using a Wilcoxon rank sum test. All statistical analysis was performed using R, version 4.4.2 (R Foundation for Statistical Computing).

Human subjects research

The study was submitted and approved by Stanford University's institutional review board (IRB# 71319). The randomized trial was registered at clinicaltrials.gov (ID: NCT06911645) beginning on December 16, 2024. Informed consent from participating physicians was obtained prior to enrollment and randomization.

Data Availability

The diagnostic challenge problems and datasets generated and analyzed during the study are not publicly available as their disclosure would risk their inclusion in training datasets of future models. The data can be made available on reasonable request to the corresponding author.

Code Availability

The system prompt for the custom GPT is available in the supplemental information. Additional information can be made available to qualified researchers on reasonable request to the corresponding author.

Author Contributions

SE: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Project administration, Writing – original draft, Writing – review & editing

BB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing

PJ: Conceptualization, Data curation, Investigation, Methodology, Project administration, Validation, Writing – review & editing

IL: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – review & editing

AA: Data curation, Writing – review & editing

MD: Methodology, Formal analysis, Writing – review & editing

RG: Writing – review & editing

EG: Methodology, Writing – review & editing

VK: Data curation, Writing – review & editing

ZK: Writing – review & editing

JK: Data curation, Writing – review & editing

AO: Writing – review & editing

AR: Writing – review & editing

KS: Writing – review & editing

ES: Writing – review & editing

JC: Supervision, Methodology, Funding acquisition, Writing – review & editing

EH: Conceptualization, Formal analysis, Investigation, Software, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing

Competing Interests:

The authors declare no competing interests.

Acknowledgements

We are grateful to Jason Hom, MD, Curtis Langlotz, MD PhD, Natalie Pageler, MD, Mihaela Vorvoreanu, PhD, Jenna Wiens, PhD, and Daniel Yang, MD for their insightful feedback. We thank Isabel Weng, MHS for guidance on the statistical analyses. This work was supported by the Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford Medical Scholars Research Program, Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program, the Gordon and Betty Moore Foundation [Grant #12409], and the National Library of Medicine [2T15LM007033].

References

1. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. 2023. doi:10.48550/arXiv.2303.13375
2. Cabral S, Restrepo D, Kanjee Z, et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. JAMA Intern Med. 84(5):581-583. 2024. doi:10.1001/jamainternmed.2024.0295
3. Goh E, Gallo R, Hom J, et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA Netw Open. 7(10):e2440969-e2440969. 2024. doi:10.1001/jamanetworkopen.2024.40969
4. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature. 1-7. 2025. doi:10.1038/s41586-025-08869-4
5. A, Kahneman D. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. Science. 185(4157):1124-1131. 1974. doi:10.1126/science.185.4157.11
6. Fogliato R, Chappidi S, Lungren M, et al. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1362–1374. 2022. doi:10.1145/3531146.3533193
7. Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., & Gogate, V. (2021). Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. 26th International Conference on Intelligent User Interfaces, 340–350. https://doi.org/10.1145/3397481.3450639
8. Yin J, Ngiam KY, Tan SSL, Teo HH. Designing AI-Based Work Processes: How the Timing of AI Advice Affects Diagnostic Decision Making. Manage Sci. 2022;–:–. doi:10.1287/mnsc.2022.01454
9. Sellen A, Horvitz E. The rise of the AI co-pilot: Lessons for design from aviation and beyond. Communications of the ACM. 67(7):18-23. 2024. doi:10.1145/3637865
10. Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-computer Interaction, 5(CSCW1), 1-21.
11. Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., & Satzger, G. (2023, March). Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. In Proceedings of the 28th International Conference on Intelligent User Interfaces (pp. 453-463).
12. Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Information Systems Research, 33(2), 678-696.
13. Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. Proceedings of the 2015 International Conference on Healthcare Informatics, 160–169.

14. Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: Susceptibility in deployment of clinical decision-aids. Npj Digital Medicine, 4(1), 1–8. https://doi.org/10.1038/s41746-021-00385-9
15. Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual Differences in the Calibration of Trust in Automation. Human Factors, 57(4), 545–556. https://doi.org/10.1177/0018720814564422
16. Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305. https://doi.org/10.1145/3351095.3372852
17. Passi, S., Dhanorkar, S., Vorvoreanu, M. (2025). Addressing Overreliance on AI. In: Xu, W. (eds) Handbook of Human-Centered Artificial Intelligence. Springer, Singapore. https://doi.org/10.1007/978-981-97-8440-0_98-1
18. Drosos I, Sarkar A, Toronto N. “ It makes you think”: Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work. ArXiv Prepr. Published online 2025. doi:10.48550/arXiv.2501.17247
19. Herbert H Clark. 1996. Using language. Cambridge University Press.
20. Shaikh O, Mozannar H, Bansal G, Fourney A, Horvitz E. Navigating Rifts in Human-LLM Grounding: Study and Benchmark. ACL 2025: Proceedings of 63rd Annual meeting of the Association for Computational Linguistics. 2025. doi:10.48550/arXiv.2503.13975
21. Susan E Brennan. 2014. The grounding problem in conversations with and through computers. In Social and cognitive approaches to interpersonal communication, pp. 201–225. Psychology Press
22. Bohus, Dan, and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. 2010.
23. Traum, D. R. (1994). A Computational Theory of Grounding in Natural Language Conversation. PhD thesis, Department of Computer Science, University of Rochester. Also available as TR 545, Department of Computer Science, University of Rochester.
24. Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, Horvitz E. Beyond accuracy: The role of mental models in human-AI team performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7(1), 2-11. 2019. doi:10.1609/hcomp.v7i1.5285
25. Horvitz E. Principles of mixed-initiative user interfaces. Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '99). Association for Computing Machinery, New York, NY, USA, 159–166. 1999. doi:10.1145/302979.303030
26. Amershi S, Weld D, Vorvoreanu M, et al. Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery, New York, NY, USA, Paper 3, 1–13. 2019. doi:10.1145/3290605.3300233

27. Wilder B, Horvitz E, Kamar E. Learning to complement humans. IJCAI'20: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. 212:1526 - 1533. 2020. doi:10.24963/ijcai.2020/212
28. G Bansal, B Nushi, E Kamar, E Horvitz, DS Weld. Is the most accurate AI the best teammate? Optimizing AI for teamwork AAAI 35 (13), 11405-11414
29. Calisto FM, Abrantes JM, Santiago C, Nunes NJ, Nascimento JC. Personalized explanations for clinician-AI interaction in breast imaging diagnosis by adapting communication to expertise levels. Int J Hum-Comput Stud. 2025;197:103444. doi:10.1016/j.ijhcs.2025.103444
30. Mozannar H, Satyanarayan A, Sontag D. Teaching Humans When to Defer to a Classifier via Exemplars. Proceedings of the AAAI Conference on Artificial Intelligence. 36, 5323-5331. 2022. doi:10.1609/aaai.v36i5.20469.
31. DS Weld, G Bansal. The challenge of crafting intelligible intelligence Communications of the ACM 62 (6), 70-79
32. Bansal, G., Tongshuang, W. U., Zhou, J., Raymond, F. O. K., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2020). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Article 81, 1–16. https://doi.org/10.1145/3411764.3445717.
33. E. Horvitz, D. Heckerman, B. Nathwani, L.M. Fagan, The use of a heuristic problem-solving hierarchy to facilitate the explanation of hypothesis-directed reasoning, October 1986, In: Proceedings of Medinfo, Washington, DC, North Holland: New York, pp. 27-31.
34. E. Horvitz and T. Paek. Complementary Computing: Policies for Transferring Callers from Dialog Systems to Human Receptionists. User Modeling and User Adapted Interaction 17 (2007).
35. E. Kamar, S. Hacker, E. Horvitz. Combining Human and Machine Intelligence in Large-scale Crowdsourcing, AAMAS 2012, Valencia, Spain, June 2012.
36. Mozannar H, Bansal G, Fourney A, Horvitz E. When to show a suggestion? Integrating human feedback in AI-assisted programming. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, No. 9, pp. 10137-10144. 2024. doi:10.1609/aaai.v38i9.28878
37. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 318(22):2199-2210. 2017. doi:10.1001/jama.2017.14585
38. Langlotz CP, Shortliffe EH. Adapting a consultation system to critique user plans. Int J Man-Mach Stud. 19(5):479-496. 1983. doi:10.1016/S0020-7373(83)80067-4
39. Miller PL. ATTENDING: Critiquing a physician’s management plan. IEEE Trans Pattern Anal Mach Intell. (5):449-461. 1983. doi:10.1109/TPAMI.1983.4767424
40. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 35:27730-27744. 2022.
41. Salecha A, Ireland ME, Subrahmanya S, Sedoc J, Ungar LH, Eichstaedt JC. Large language models display human-like social desirability biases in Big Five personality surveys. PNAS Nexus. 3(12):pgae533. 2024. doi:10.1093/pnasnexus/pgae533.

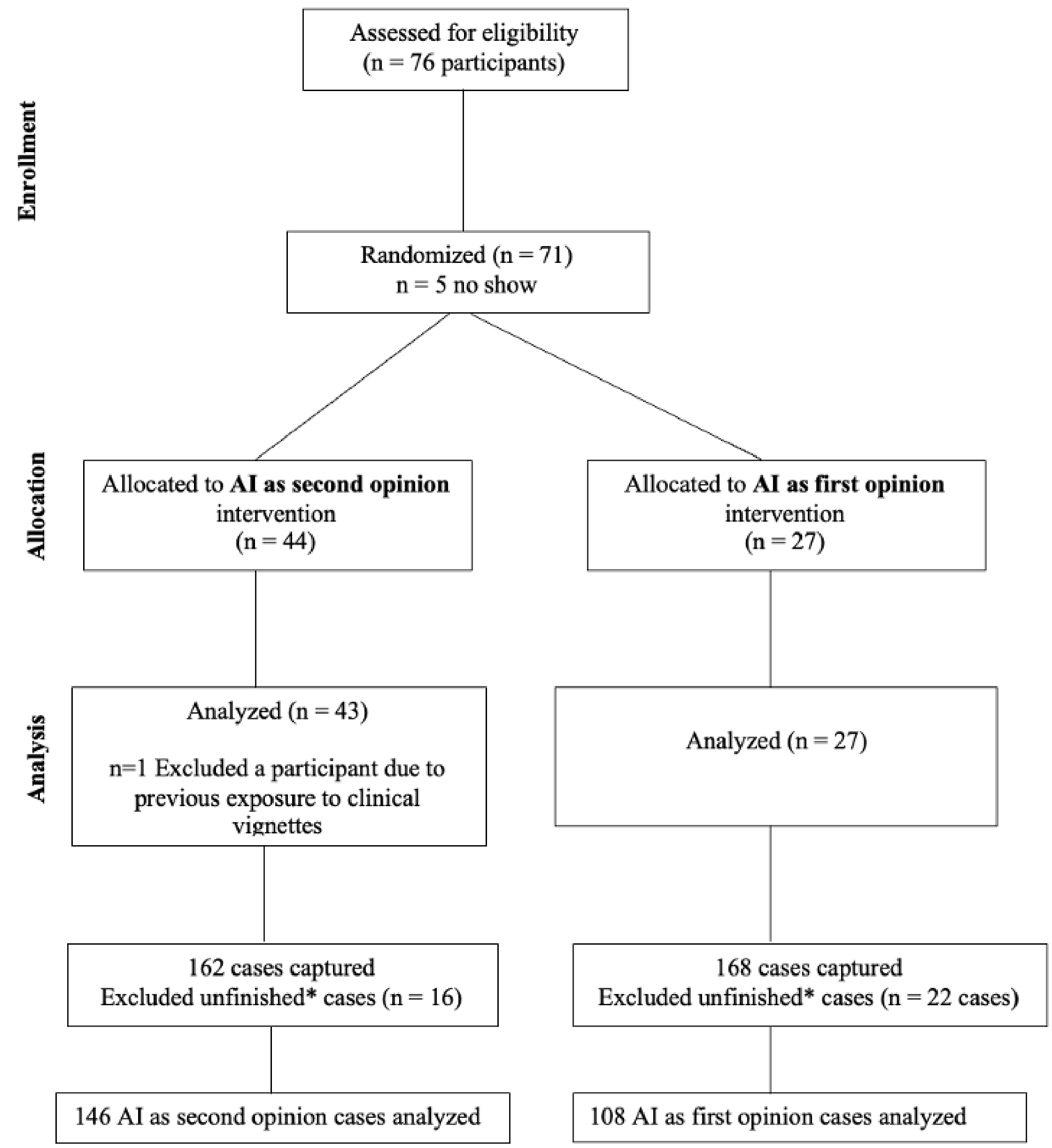
42. Sharma M, Tong M, Korbak T, et al. Towards understanding sycophancy in language models. ArXiv Prepr. Published online 2023. doi:10.48550/arXiv.2310.13548
43. Savage T, Wang J, Gallo R, et al. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. J Am Med Inform Assoc. 32(1):139-149. 2025. doi:10.1093/jamia/ocae254
44. Balachandran V, Chen J, Joshi N, et al. Eureka: Evaluating and understanding large foundation models. ArXiv Prepr. Published online 2024. doi:10.48550/arXiv.2409.10566

Synthesis of Reasoning (AI + Clinician)

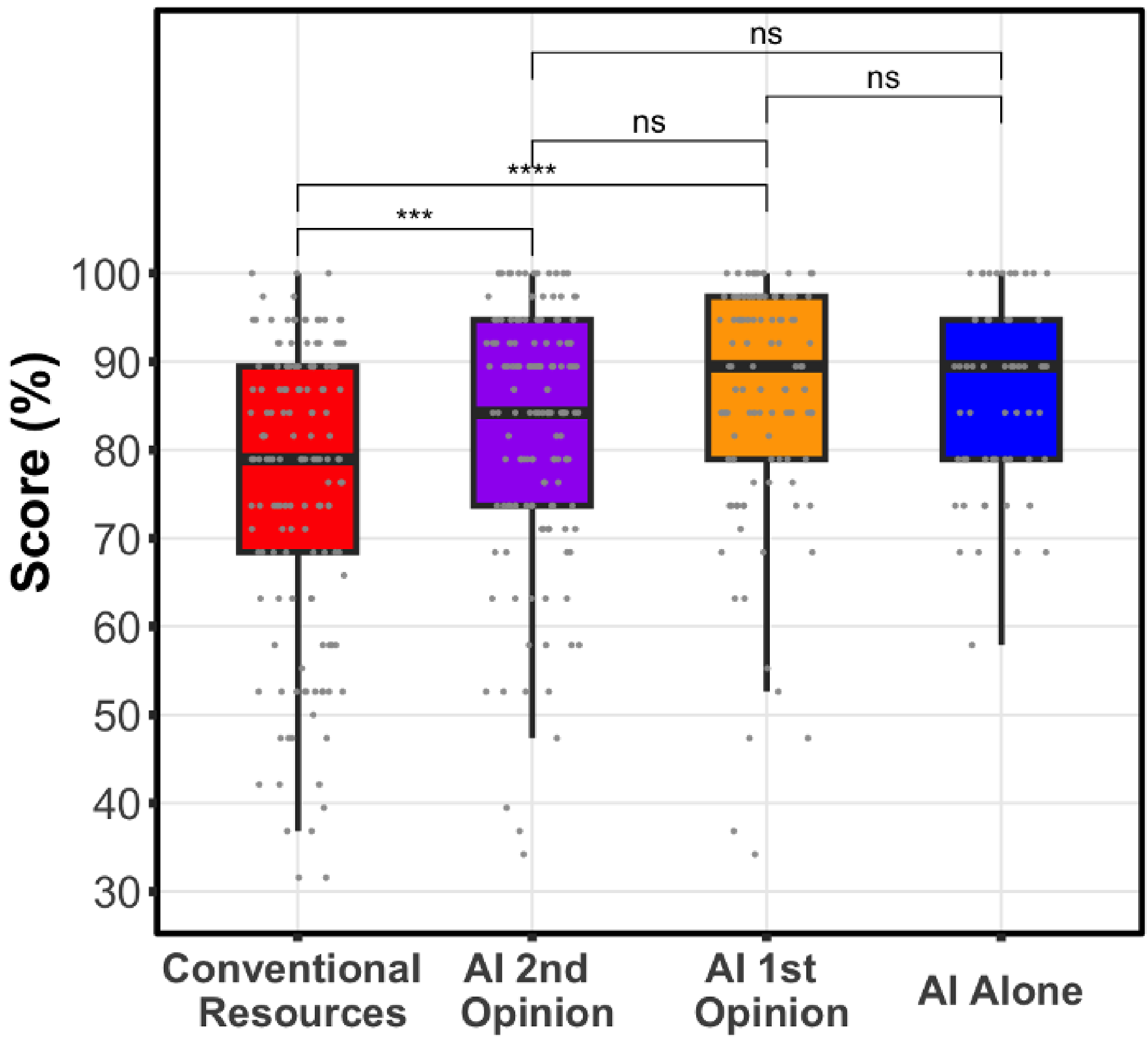
Diagnosis	Origin	Comments
Polycythemia Vera	AI & Clinician	Best explained by low EPO, pruritus, hyperuricemia, and panmyelosis
EPO-secreting tumor	Clinician	Low EPO makes this unlikely; however, rare analogs (e.g., EPO-like substances) are possible but far less likely
Lymphoma	AI & Clinician	Supported by pruritus and uric acid; not supported by rest of picture
Primary Biliary Cholangitis	Clinician	Itch without cholestasis is atypical; unlikely
Other myeloproliferative neoplasm	AI	Considered if PV ruled out

Demographics	**AI as Second Opinion (n=43)**	**AI as First Opinion (n=27)**
Specialty		
Internal Medicine	42 (97.7%)	26 (96.3%)
Family medicine	1 (2.3%)	1 (3.7%)
Position		
Resident	24 (55.8%)	15 (55.6%)
Attending	19 (44.2%)	12 (44.4%)
Site		

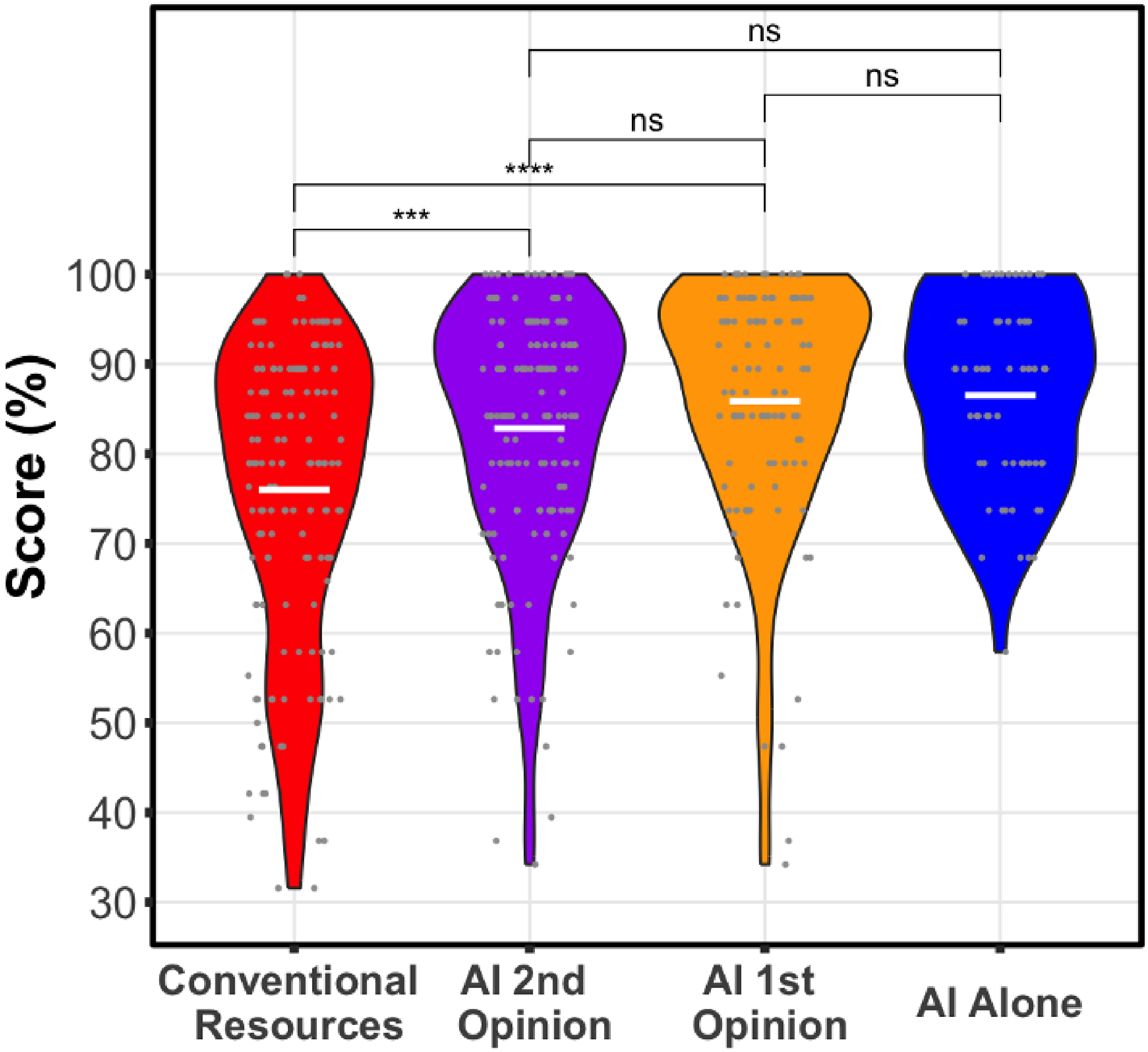
Beth Israel Deaconess Medical Center	15 (34.9%)	15 (55.6%)
Brigham and Women's Hospital	0 (0%)	1 (3.7%)
Cambridge Health Alliance	3 (7.0%)	2 (7.4%)
Columbia University Irving Medical Center	13 (30.2%)	4 (14.8%)
Stanford University Hospital	9 (20.9%)	2 (7.4%)
Vanderbilt University Medical Center	2 (4.7%)	3 (11.1%)
Prior Generative AI Experience		
Use frequently (weekly or more)	18 (41.9%)	9 (33.3%)
Use occasionally (more than once per month but less than weekly)	13 (30.2%)	7 (25.9%)
Use rarely (less than once per month)	7 (16.3%)	9 (33.3%)
Never used before	4 (9.3%)	2 (7.4%)



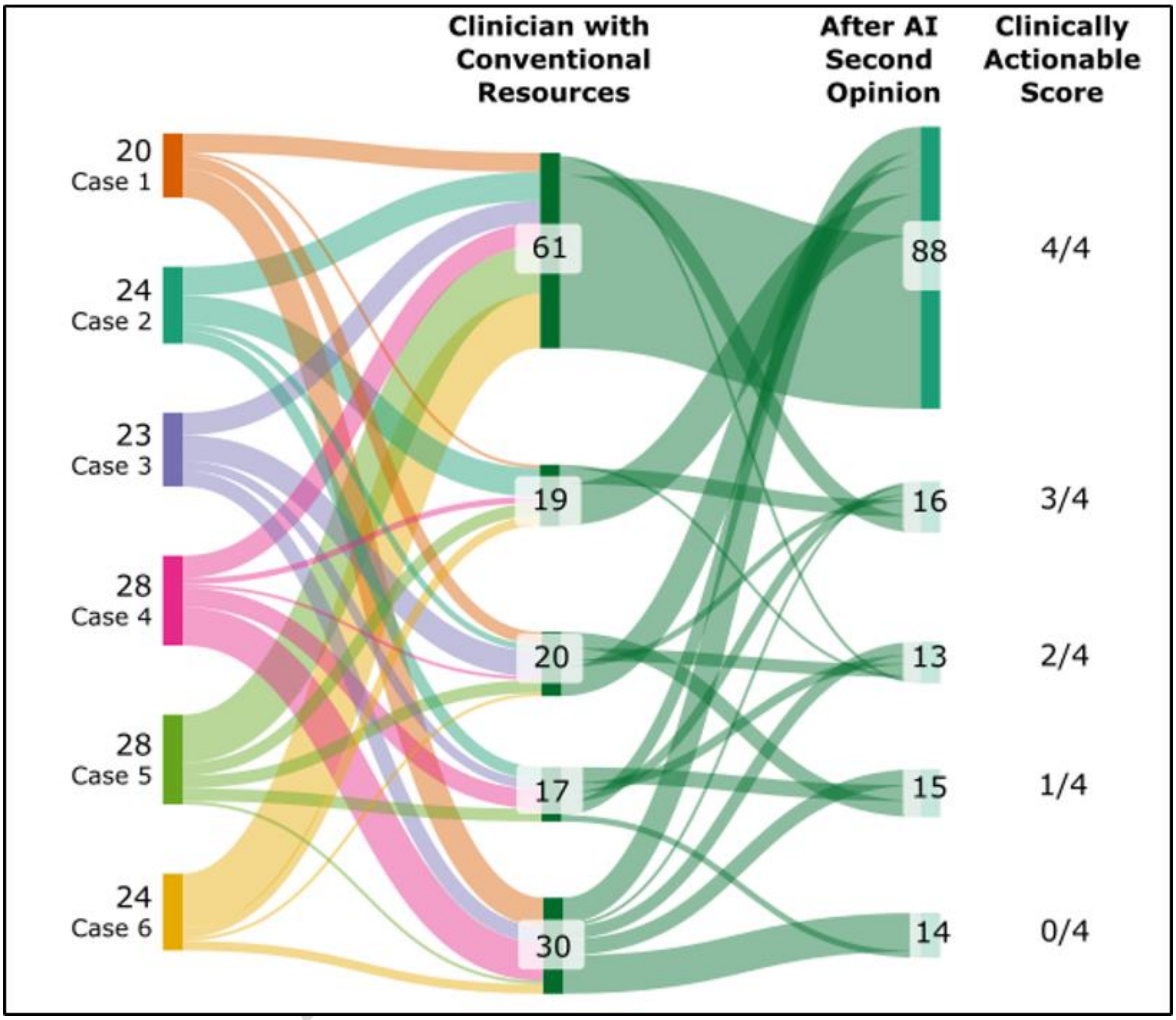
*unfinished was defined as (1): it was the final vignette in a session, and (2) the participant spent less than 30 seconds reviewing information.



ns
ns
ns
****
***
100
90
80
70
60
50
40
30
Score (%)
Conventional Resources
AI 2nd Opinion
AI 1st Opinion
AI Alone



ns
ns
ns
****
***
100
90
80
70
60
50
40
30
Score (%)
Conventional Resources
AI 2nd Opinion
AI 1st Opinion
AI Alone


Clinician with Conventional Resources
After AI Second Opinion
Clinically Actionable Score
20
Case 1
24
Case 2
23
Case 3
28
Case 4
28
Case 5
24
Case 6
61
19
20
17
30
88
16
13
15
14
4/4
3/4
2/4
1/4
0/4

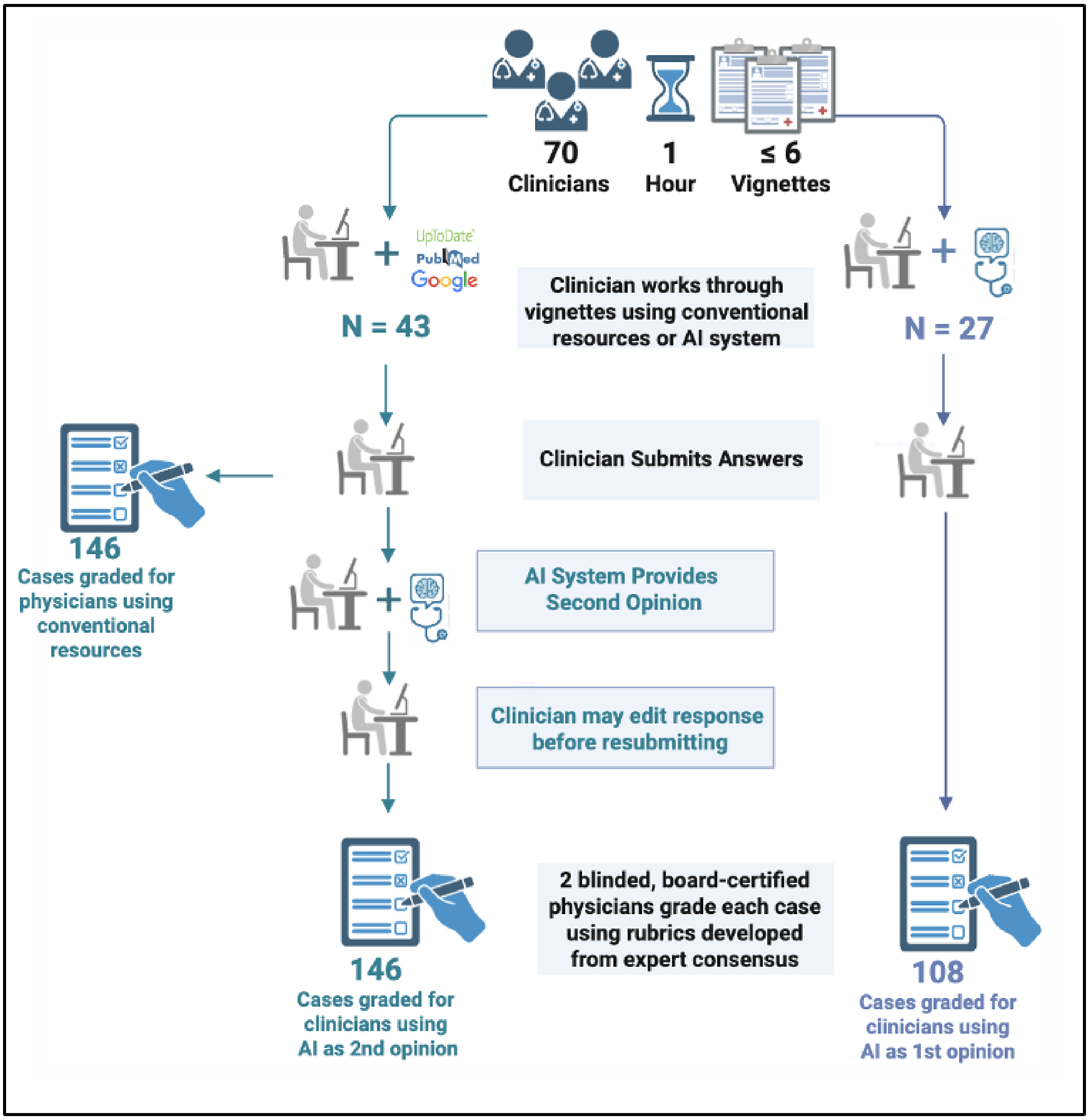


Table legends

Table 1. Demographics and prior generative AI experience of study participants.

Figure Legends

Figure 1. Sample display of synthesis of AI and clinician input with critiques.

The table illustrates an example of how clinician and AI contributions are integrated during case evaluation, showing candidate diagnoses, attribution of reasoning origin (clinician, AI, or both), and brief critiques summarizing supporting and conflicting evidence for each option.

Figure 2: CONSORT Flow Diagram of Study Enrollment, Allocation, and Analysis.

The diagram depicts clinician enrollment, randomization, allocation to AI as a first- or second-opinion workflow, and case-level inclusion in the final analysis, resulting in 146 AI second-opinion cases and 108 AI first-opinion cases after exclusion of unfinished cases and one ineligible participant.

Figure 3: Distribution of Diagnostic Performance Scores across Clinician-AI Workflows.

Box-and-jitter plots show overall diagnostic performance scores (%) for clinicians using conventional resources alone (red), AI as a second opinion (purple), AI as a first opinion (orange), and AI alone (blue). Points represent individual case scores. Boxes indicate the interquartile range (25th–75th percentiles), with the black horizontal line denoting the median and whiskers extending to the range of non-outlier values. Using a prespecified linear mixed-effects model controlling for both case- and participant-level variation, clinicians using conventional resources had significantly lower scores than those using AI as a first opinion ($p = 0.00039$; mean difference, 9.9%; 95% CI, 4.7%–15.0%) or as a second opinion ($p = 3.3 \times 10^{-6}$; mean difference, 6.8%; 95% CI, 4.0%–9.6%). No significant difference was observed between AI as a first versus second opinion ($p = 0.22$; mean difference, 3.0% favoring first opinion; 95% CI, −1.7% to 7.6%). Statistical annotations indicate pairwise comparisons (ns, not significant; *** and **** denote increasing levels of significance).

Figure 4: Distribution of Diagnostic Performance Scores Visualized as Violin Plots. .

Violin plots show the full distribution of diagnostic performance scores (%) for clinicians using conventional resources alone (red), AI as a second opinion (purple), AI as a first opinion (orange), and AI alone (blue). Individual dots represent case-level scores, and the width of each violin reflects score density. White horizontal lines indicate mean performance within each workflow. This visualization provides an alternate view of the results in Figure 4, demonstrating that AI-enabled workflows are associated with higher central tendency and a marked reduction in the lower-score tail compared with conventional resources alone, suggesting that AI support primarily mitigates poor-performing outcomes rather than uniformly increasing scores across all cases.

Figure 5: Changes in clinically actionable decision scores following AI second-opinion use.

Sankey diagram illustrating shifts in clinically actionable decision scores (0–4 points) across six clinical vignettes before and after AI was employed as a second opinion. Scores represent four expert-defined criteria encompassing the final diagnosis and next clinical steps. The left column shows score distributions when clinicians used only conventional resources, while the right column shows final scores after clinicians reviewed AI feedback and optionally revised their responses. Flow widths correspond to the number of cases transitioning between score levels. Overall, AI second-opinion use was associated with upward score shifts, including an increase in perfect scores (4/4) from 61 cases with conventional resources alone to 88 cases after AI assistance, indicating improved clinical actionability across multiple vignettes.

Figure 6: Study design and clinician–AI workflow.

Seventy clinicians completed up to six clinical vignettes within one hour, using either conventional resources alone, an AI system as a first opinion, or AI as a second opinion after an initial response. All submissions were independently graded by two blinded, board-certified physicians using expert-derived rubrics, yielding 146 conventionally assisted cases, 146 AI–second opinion cases, and 108 AI–first opinion cases.