

Performance of Claude 3.5 Sonnet Compared to Human Readers of *The New England Journal of Medicine* Clinical Case Challenges

Jack Litwin, BS | Daniel J. Parente, MD, PhD

PRiMER. 2026;10:40.

Published: 10/16/2026 | DOI:[10.22454/PRiMER.2025.0129](https://doi.org/10.22454/PRiMER.2025.0129)

Abstract

Introduction: Large language models (LLMs) can pass the US medical licensing exam and outperform residents on the family medicine in-training exam. The diagnostic performance of LLMs when applied to more complex medical cases, however, remains unknown.

Methods: We evaluated the diagnostic accuracy of Claude (Anthropic), an LLM, on 20 case challenges published in *The New England Journal of Medicine* (NEJM). For each case, we presented Claude 3.5 Sonnet with multimodal input, including the text and all corresponding images from the cases. Claude was instructed to give an answer among multiple choices. We compared Claude's diagnostic accuracy to human respondents on the NEJM website.

Results: Claude 3.5 Sonnet diagnosed cases correctly about half the time, and at a significantly higher rate than NEJM readers (beta-binomial odds ratio 2.79; 95% CI [profile-likelihood] 1.09–7.01; $P = 0.031$). Claude was highly self-consistent in repeated exposure to questions, with perfect accuracy (100%) in nine cases and no accuracy (0%) in 10 cases, with one case at 90% accuracy. Human accuracy ranged from 9%–63% across cases. Claude's performance was better on easier cases (ie, those with human accuracy >25%).

Conclusions: Claude 3.5 Sonnet demonstrated significantly higher diagnostic accuracy compared to medical journal readers, though with notable variability in performance. Further research comparing consistency and reliability of artificially intelligent diagnostic capabilities to that of licensed physicians is needed.

Introduction

Large language models (LLMs) are artificially intelligent (AI) systems that use artificial neural network machine learning on large sets of data to statistically describe language. Applications of AI to family medicine and medical education are being widely discussed.¹⁻⁴¹ LLMs are largely black boxes: Inspection of the inner workings has only limited ability to explain observed outputs. Thus, LLMs must be evaluated based on the quality of their outputs alone. Fortunately, medical educators have long experience in evaluating black-box neural networks: using exams to evaluate human brains.

Indeed, LLMs have done well on medical exams. GPT-4 (OpenAI) passed the US medical licensing exam (USMLE)⁴² and scored higher on the family medicine in-training exam (ITE) than third-year residents.²⁰ While this is exciting, the USMLE and ITE are artificial scenarios that are, by design, contrived to eliminate the uncertainty of real-world practice. Less research exists on how well these models can perform in clinical settings with higher complexity.

The public conversation around LLMs was spurred by the release of ChatGPT (OpenAI) in 2022.^{43,44} However, many other models exist. A recent comparison between Google Bard, Perplexity AI, ChatGPT, and Claude (Anthropic) found that Claude scored higher in relevance and completeness when analyzing medical cases.⁴⁵ In light of these new studies, we therefore focused on the performance of Claude 3.5 Sonnet.

Here, we evaluated the diagnostic capabilities of Claude on complex cases drawn from the published literature,⁴⁶⁻⁶⁵ using an approach previously applied to ChatGPT.⁶⁶ We hypothesized that Claude would significantly outperform human readers in answering multiple-choice questions based on published, real-world, complex cases.

Methods

We retrieved all 20 freely available (at time of curation) *The New England Journal of Medicine* (NEJM) case challenges released between 2020 and 2024.⁴⁶⁻⁶⁵ Each case, based on real patient encounters, included a history and physical with pertinent imaging and/or laboratory results. Readers selected the most likely diagnosis from a list of six answer choices.

We presented Claude 3.5 Sonnet with a standardized prompt (Appendix), followed by the text of the case and any related images in JPEG format, excluding figure/table captions but including table footnotes. The prompt instructed the AI model on how to approach the task, emphasizing careful reasoning and selection of a single most likely diagnosis from the options provided.

Claude's output is partially nondeterministic: The same input might result in different output. For any given question, Claude might always select a correct answer, might always select an incorrect answer, or might answer correctly only some of the time. We therefore presented each case to Claude 10 times and calculated the accuracy rate across these trials (20 cases × 10 repetitions). We performed these analyses between September 24, 2024, and October 20, 2024. We calculated Claude's per-case and overall accuracy.

We compared Claude's accuracy with the accuracy of NEJM readers, as reported on the NEJM website, using beta-binomial regression with the logit of human accuracy included as a per-case offset.

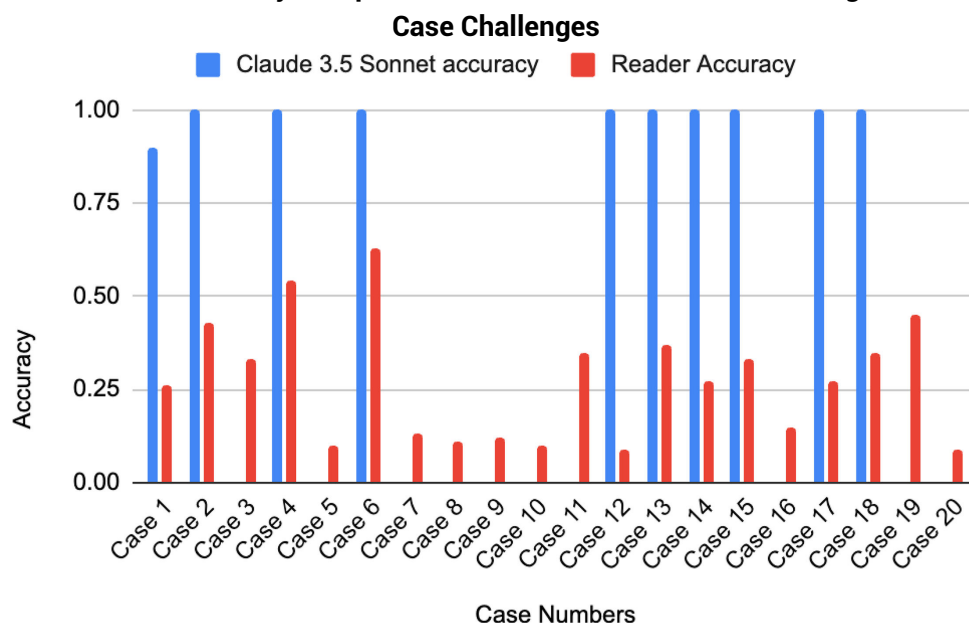
The University of Kansas Medical Center Institutional Review Board determined this study was not human subjects research.

Results

Sample output is shown in the Appendix. Claude 3.5 Sonnet made the correct diagnosis in 99 of 200 (49.5%) trials. Within cases, Claude was almost entirely self-consistent: 10 cases having 0% accuracy, one case having 90% accuracy, and nine cases having 100% accuracy (Figure 1).

A median of 7,576 (IQR: 5,751–9,193) human readers responded to each of these cases on the NEJM website. Human respondents had an overall accuracy of 27.4%. Accuracy on individual cases was—as with Claude—highly variable, ranging from 9%–63% across cases. Claude's accuracy was significantly higher than human readers (beta-binomial odds ratio 2.79; 95% profile-likelihood CI 1.09–7.01; $P = 0.031$).

Figure 1. Claude 3.5 Sonnet Accuracy Compared to Human Readers of the New England Journal of Medicine



Among cases humans answered with greater than 25% accuracy (“easier”), Claude achieved 74.1% (89/120) accuracy, as compared to only 12.5% (10/80) accuracy on “harder” cases.

Discussion

We hypothesized that Claude 3.5 Sonnet would have superior diagnostic capabilities, as compared with human NEJM readers, when presented with complex clinical case challenges; this hypothesis was borne out. First, we concluded that this hypothesis suggests that LLMs may help physicians approach real-world situations that include diagnostic uncertainty. However, Claude’s performance was dependent upon case difficulty, struggling with questions humans likewise find difficult.

Second, AI is not a diagnostic panacea: Claude generated incorrect solutions in half of the cases. Our findings are concordant with similar studies evaluating ChatGPT-4.⁶⁶ AI systems should be viewed with open-minded skepticism unless and until their reasoning is reliably correct.

Third, despite Claude’s nondeterministic output, its answers were almost entirely consistent across trials (199/200 trials). Claude’s responses were, thus, repeatable, and asking Claude once for an opinion seems sufficient, rather than prompting multiple times.

These conclusions are nevertheless preliminary due to limitations inherent in this study. Most importantly, we cannot assess the medical background of the NEJM readers and thus cannot extrapolate the data in comparison with licensed physicians, who would be the gold-standard reference population. Second, this study evaluated diagnostic accuracy on multiple-choice questions; given that these cases presented six multiple-choice answers, the model and the readers had a 16.7% chance of guessing correctly. Real-world applications of LLMs would not have multiple choice.

While diagnostic accuracy is an important part of a physician’s job, physicians require other skills to practice medicine. Neither AI nor journal readers were asked to take a thorough history or perform a detailed physical examination. Furthermore, all relevant labs and tests had seemingly already been ordered, and explanations were provided for many of the radiological findings. Real medical practice requires deciding not only how to react to already-collected information but also deciding what information (if any) to collect next.

Future studies comparing Claude to licensed physicians on complex cases would be worthwhile. Likewise, to ensure the best models are used, future studies should assess not only a specific LLM's diagnostic capabilities in comparison to physicians but also in comparison to other LLMs. Furthermore, LLMs are only one component of artificial intelligence, and exploring how the integration of various tools can be used in clinical practice will be important.

Presentations

Litwin J. Claude 3.5 Sonnet Medical Case Review: Analyzing the Diagnostic Capabilities of Claude 3.5 Sonnet on Complex Clinical Cases. Oral Presentation at KUMC Student Research Forum, April 8–9, 2025, Kansas City, KS.

Corresponding Author

Daniel J. Parente

School of Medicine, University of Kansas Medical Center, Kansas City, KS

dparente@kumc.edu

Author Affiliations

Jack Litwin, BS - School of Medicine, University of Kansas Medical Center, Kansas City, KS

Daniel J. Parente, MD, PhD - School of Medicine, University of Kansas Medical Center, Kansas City, KS | Family Medicine and Community Health, University of Kansas Medical Center, Kansas City, KS

References

1. Wang T, Price DW, Bazemore AW. Leveraging large language models to advance certification, physician learning, and diagnostic excellence. *J Am Board Fam Med*. 2025;38(3):599–602. doi:[10.3122/jabfm.2024.240385R1](https://doi.org/10.3122/jabfm.2024.240385R1)
2. Waldren S, Gerhart J, Liaw W, et al. A family medicine shared vision and road map for AI in primary care. *Ann Fam Med*. 2025;23(5):481–483. doi:[10.1370/afm.250510](https://doi.org/10.1370/afm.250510)
3. Tsai T, Lee JJ, Phillips R, Lin S. Data transformation to advance AI/ML research and implementation in primary care. *Ann Fam Med*. 2025;23(4):363–367. doi:[10.1370/afm.240459](https://doi.org/10.1370/afm.240459)
4. Schrager S, Seehusen DA, Sexton SM, et al. Use of AI in family medicine publications: a joint editorial from journal editors. *PRiMER*. 2025;9:3. doi:[10.22454/PRiMER.2025.889328](https://doi.org/10.22454/PRiMER.2025.889328)
5. Rodríguez JE, Lussier Y. The AI moonshot: what we need and what we do not. *Ann Fam Med*. 2025;23(1):7–8. doi:[10.1370/afm.240602](https://doi.org/10.1370/afm.240602)
6. Moschogianis S, Darley S, Coulson T, Peek N, Cheraghi-Sohi S, Brown BC. Seven opportunities for artificial intelligence in primary care electronic visits: qualitative study of staff and patient views. *Ann Fam Med*. 2025;23(3):214–222. doi:[10.1370/afm.240292](https://doi.org/10.1370/afm.240292)
7. Lin S. Establishing a national framework for family medicine AI centers of excellence. *Fam Med*. 2025;57(8):612–614. doi:[10.22454/FamMed.2025.142618](https://doi.org/10.22454/FamMed.2025.142618)
8. Daungsupawong H, Wiwanitkit V. Response to “Artificial intelligence-prompted explanations of common primary care diagnoses”. *PRiMER*. 2025;9:27. doi:[10.22454/PRiMER.2025.412803](https://doi.org/10.22454/PRiMER.2025.412803)
9. Clebak KT, Partin MT, Newman R, et al. Artificial intelligence (AI) adoption, policies, and goals in family medicine: a survey of department chairs. *J Am Board Fam Med*. 2025;38(4):740–744. doi:[10.3122/jabfm.2025.250003R1](https://doi.org/10.3122/jabfm.2025.250003R1)
10. Britz J, Bowman MA, Seehusen DA, Ledford CJW. The 4Cs of primary care, leveraging artificial intelligence, and improving clinical practice. *J Am Board Fam Med*. 2025;38(3):631–633. doi:[10.3122/jabfm.2025.250167R0](https://doi.org/10.3122/jabfm.2025.250167R0)

11. Akyon SH, Arman IH, Onur G, Abecasis R, Unalan PC. Artificial intelligence readiness among young family doctors in Europe. *Ann Fam Med*. 2025;23(6):542–545. doi:10.1370/afm.250045
12. Young RA, Martin CM, Sturmberg JP, et al. What complexity science predicts about the potential of artificial intelligence/machine learning to improve primary care. *J Am Board Fam Med*. 2024;37(2):332–345. doi:10.3122/jabfm.2023.230219R1
13. Wang T, Mainous AG 3rd, Stelter K, O'Neill TR, Newton WP. Performance evaluation of the generative pre-trained transformer (GPT-4) on the family medicine in-training examination. *J Am Board Fam Med*. 2024;37(4):528–582. doi:10.3122/jabfm.2023.230433R1
14. Tulshian P. Artificial intelligence: reflections on its use to promote DEIA principles for residency recruitment. *Fam Med*. 2024;56(6):344–345. doi:10.22454/FamMed.2024.951121
15. Seehusen DA, Bowman MA, Britz J, Ledford CJW. Artificial intelligence and family medicine. *J Am Board Fam Med*. 2024;37(4):517–519. doi:10.3122/jabfm.2024.240233R0
16. Parente DJ. Generative artificial intelligence and large language models in primary care medical education. *Fam Med*. 2024;56(9):534–540. doi:10.22454/FamMed.2024.775525
17. Liaw W, Hischier B, James CA, Kakadiaris I, Kueper JK, Rahimzadeh V. Bridging the gap: transforming primary care through the artificial intelligence and machine learning for primary care (AIM-PC) curriculum. *Ann Fam Med*. 2024;22(6):570–571. doi:10.1370/afm.240537
18. Kleebayoon A, Wiwanitkit V. The hazards of using ChatGPT: additional comments. *PRiMER*. 2024;8:12. doi:10.22454/PRiMER.2024.203721
19. Kattih M, Bressler M, Smith LR, Schinelli A, Mhaskar R, Hanna K. Artificial intelligence-prompted explanations of common primary care diagnoses. *PRiMER*. 2024;8:51. doi:10.22454/PRiMER.2024.916089
20. Hanna RE, Smith LR, Mhaskar R, Hanna K. Performance of language models on the family medicine in-training exam. *Fam Med*. 2024;56(9):555–560. doi:10.22454/FamMed.2024.233738
21. Hanna K, Chartash D, Liaw W, et al. Family medicine must prepare for artificial intelligence. *J Am Board Fam Med*. 2024;37(4):520–524. doi:10.3122/jabfm.2023.230360R1
22. Hake J, Crowley M, Coy A, et al. Quality, accuracy, and bias in ChatGPT-based summarization of medical abstracts. *Ann Fam Med*. 2024;22(2):113–120. doi:10.1370/afm.3075
23. Gotler RS, Snyder B, Smith CK, et al. Medical students' views of the future in a rapidly changing world. *Fam Med*. 2024;56(9):541–547. doi:10.22454/FamMed.2024.918294
24. Gilfoyle M, Bosworth KT, Adesanya TMA, et al. Exploring artificial intelligence and the future of primary care. *Ann Fam Med*. 2024;22(2):174–175. doi:10.1370/afm.3112
25. Garies S, Liang S, Weyman K, Ramji N, Alhaj M, Pinto AD. Developing an AI tool to derive social determinants of health for primary care patients: qualitative findings from a codesign workshop. *Ann Fam Med*. 2024;22(4):317–324. doi:10.1370/afm.3117
26. Bowman MA, Seehusen DA, Britz J, Ledford CJW. Research to improve clinical care in family medicine: big data, telehealth, artificial intelligence, and more. *J Am Board Fam Med*. 2024;37(2):161–164. doi:10.3122/jabfm.2024.240050R0
27. Upshaw TL, Craig-Neil A, Macklin J, et al. Priorities for artificial intelligence applications in primary care: a Canadian deliberative dialogue with patients, providers, and health system leaders. *J Am Board Fam Med*. 2023;36(2):210–220. doi:10.3122/jabfm.2022.220171R1
28. Seehusen DA, Bowman MA, Ledford CJW. Research representing the changing landscape of family medicine. *J Am Board Fam Med*. 2023;36(2):207–209. doi:10.3122/jabfm.2023.230024R0
29. Rahimi S, Chu C, Grad R, et al. Explainable machine learning model to predict COVID-19 severity among older adults in the province of Quebec. *Ann Fam Med*. 2023;21(21 Suppl 1):3619. doi:10.1370/afm.21.s1.3619

30. Nash DM, Thorpe C, Brown JB, et al. Perceptions of artificial intelligence use in primary care: a qualitative study with providers and staff of Ontario community health centres. *J Am Board Fam Med.* 2023;36(2):221–228. doi:10.3122/jabfm.2022.220177R2
31. Liaw W, Chavez S, Pham C, Tehami S, Govender R. The hazards of using ChatGPT: a call to action for medical education researchers. *PRiMER.* 2023;7:27. doi:10.22454/PRiMER.2023.295710
32. Horton JA, Ally I. Response to “Exploring the applications of ChatGPT in family medicine medical education.” *PRiMER.* 2023;7:28. doi:10.22454/PRiMER.203.940827
33. Hanna K. Exploring the applications of ChatGPT in family medicine education: five innovative ways for faculty integration. *PRiMER.* 2023;7:26. doi:10.22454/PRiMER.2023.985351
34. Ellertsson S, Hlynsson HD, Loftsson H, Sigurðsons EL. Triaging patients with artificial intelligence for respiratory symptoms in primary care to improve patient outcomes: a retrospective diagnostic accuracy study. *Ann Fam Med.* 2023;21(3):240–248. doi:10.1370/afm.2970
35. Terry A, Lizotte D, Brown J, et al. Is primary health care ready for artificial intelligence? stakeholder perspectives: worth the risk as long as you do it well. *Ann Fam Med.* 2022;20(Suppl 1):2905. doi:10.1370/afm.20.s1.2905
36. Lin S. A clinician’s guide to artificial intelligence (AI): why and how primary care should lead the health care AI revolution. *J Am Board Fam Med.* 2022;35(1):175–184. doi:10.3122/jabfm.2022.01.210226
37. Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I. Competencies for the use of artificial intelligence in primary care. *Ann Fam Med.* 2022;20(6):559–563. doi:10.1370/afm.2887
38. Lennon R, Calo W, Miller E, Zgierska A, Van Scoy L, Fraleigh R. Using artificial intelligence to support rapid, mixed-methods analysis: developing an automated qualitative assistant (AQUA). *Ann Fam Med.* 2022;20(Suppl 1):2893. doi:10.1370/afm.20.s1.2893
39. Kueper J, Lizotte D, Brown J, et al. Identifying priorities for artificial intelligence and primary care in Ontario: a multi-stakeholder engagement event. *Ann Fam Med.* 2022;20(Suppl 1):2943. doi:10.1370/afm.20.s1.2943
40. Hong G, Folcarelli A, Less J, Wang C, Erbas N, Lin S. Voice assistants and cancer screening: a comparison of Alexa, Siri, Google Assistant, and Cortana. *Ann Fam Med.* 2021;19(5):447–449. doi:10.1370/afm.2713
41. Holdsworth LM, Park C, Asch SM, Lin S. Technology-enabled and artificial intelligence support for pre-visit planning in ambulatory care: findings from an environmental scan. *Ann Fam Med.* 2021;19(5):419–426. doi:10.1370/afm.2716
42. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv.* April 2023. doi:10.48550/arXiv.2303.13375
43. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *arXiv.* Epub 10.48550/arXiv.2005.14165 November 25, 2023. doi:
44. Open AI. Introducing ChatGPT. November 30, 2022. Accessed November 27, 2023. <https://openai.com/blog/chatgpt>
45. Uppalapati VK, Nag DS. A comparative analysis of AI models in complex medical decision-making scenarios: evaluating ChatGPT, Claude AI, Bard, and Perplexity. *Cureus.* 2024;16(1):e52485. doi:10.7759/cureus.52485
46. Vinh JM, Ford JN, Jorge AM, Zonozi R. Case 7-2024: a 67-year-old woman with alternating sixth cranial nerve palsy. *N Engl J Med.* 2024;390(9):843–851. doi:10.1056/NEJMcp2312740
47. Hung AY, Gupta R, Silverman SB, Hilburn CF. Case 31-2023: a 79-year-old man with involuntary movements and unresponsiveness. *N Engl J Med.* 2023;389(15):1416–1423. doi:10.1056/NEJMcp2300973
48. Hochberg NS, Dinculescu VV, Nutman TB. Case 17-2023: a 58-year-old woman with fatigue, abdominal bloating, and eosinophilia. *N Engl J Med.* 2023;388(23):2180–2189. doi:10.1056/NEJMcp2201248

49. Granfone A, Applewhite BP, Palmer BF, Jobbagy S. Case 2-2023: a 76-year-old man with dizziness and altered mental status. *N Engl J Med.* 2023;388(3):264–272. doi:10.1056/NEJMcpc2201240
50. Chwalisz BK, Kelly HR, Flaherty AW, Jorge AM, Murali MR. Case 38-2023: a 68-year-old woman with abnormal movements and confusion. *N Engl J Med.* 2023;389(24):2277–2285. doi:10.1056/NEJMcpc2309349
51. Bolster MB, Cahalane AM, Grinspoon SK, Miloslavsky EM. Case 12-2023: a 44-year-old woman with muscle weakness and myalgia. *N Engl J Med.* 2023;388(16):1513–1520. doi:10.1056/NEJMcpc2211375
52. Zanger K, Nimkin K, Israel EJ, Shih AR. Case 3-2022: a 14-year-old boy with fever, joint pain, and abdominal cramping. *N Engl J Med.* 2022;386(4):375–383. doi:10.1056/NEJMcpc2115850
53. Ottolini MC, Shailam R, Sharma A, Rosales IA. Case 21-2022: a 17-year-old girl with fever and cough. *N Engl J Med.* 2022;387(2):168–176. doi:10.1056/NEJMcpc2115851
54. Leaf RK, Dhami RS, Chun NS, Ta R. Case 11-2022: an 80-year-old woman with pancytopenia. *N Engl J Med.* 2022;386(15):1453–1461. doi:10.1056/NEJMcpc2201232
55. Knipe RS, Alba GA, Harvey Barnes JM, Hariiri LP. Case 15-2022: a 57-year-old man with persistent cough and pulmonary opacities. *N Engl J Med.* 2022;386(20):1933–1944. doi:10.1056/NEJMcpc2201234
56. Forst DA, Restrepo JA, Gonzalez RG, Jones PS, Marshall MS. Case 7-2022: a 65-year-old woman with depression, recurrent falls, and inability to care for herself. *N Engl J Med.* 2022;386(10):977–986. doi:10.1056/NEJMcpc2115853
57. Cohen EM, Sobrin L, Figueiro Longo MG, Pier DB, Brown DR, Stagner AM. Case 33-2022: an 11-year-old girl with redness of the eyes. *N Engl J Med.* 2022;387(17):1598–1607. doi:10.1056/NEJMcpc2201235
58. Winston LG, Winkler ML, Kheterpal A, Villalba JA. Case 36-2021: a 22-year-old man with pain and erythema of the left hand. *N Engl J Med.* 2021;385(22):2078–2086. doi:10.1056/NEJMcpc2107357
59. Unizony SH, Kelly HR, O'Donnell EK, Sadigh S. Case 3-2021: a 48-year-old man with transient vision loss. *N Engl J Med.* 2021;384(4):363–372. doi:10.1056/NEJMcpc2002416
60. Moore AB, Wing JR, Goiffon RJ, Leaf RK, Tsao L, Misdraji J. Case 25-2021: a 48-year-old man with fatigue and leg swelling. *N Engl J Med.* 2021;385(8):745–754. doi:10.1056/NEJMcpc2100282
61. Bendapudi PK, Whalen MJ, Lahoud-Rahme M, Villalba JA. Case 7-2021: a 19-year-old man with shock, multiple organ failure, and rash. *N Engl J Med.* 2021;384(10):953–963. doi:10.1056/NEJMcpc2027093
62. Virkud YV, Fernandes ND, Lim R, Mitchell DM, Rothwell WT. Case 39-2020: a 29-month-old boy with seizure and hypocalcemia. *N Engl J Med.* 2020;383(25):2462–2470. doi:10.1056/NEJMcpc2027078
63. Mukerji SS, Ard KL, Schaefer PW, Branda JA. Case 32-2020: a 63-year-old man with confusion, fatigue, and garbled speech. *N Engl J Med.* 2020;383(16):1578–1586. doi:10.1056/NEJMcpc2004996
64. Gottlieb GS, Rosenberg JM, Gonzalez RG, Gandhi RT. Case 27-2020: a 53-year-old woman with headache and gait imbalance. *N Engl J Med.* 2020;383(9):859–866. doi:10.1056/NEJMcpc1913472
65. Armstrong KA, Cohen JV, Shepard JAO, Folch EE, Mansour MK, Stefely JA. Case 16-2020: a 47-year-old woman with recurrent melanoma and pulmonary nodules. *N Engl J Med.* 2020;382(21):2034–2043. doi:10.1056/NEJMcpc1916258
66. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complex clinical cases. *NEJM AI.* 2024;1(1). doi:10.1056/AIp2300031