**Diagnostic Performance of Generative AI and Physicians: A Systematic Review and Meta-Analysis.**

Hirotaka Takita, MD, PhD, Shannon L Walston, MS, Kenichi Saito, MD, Yasushi Tsujimoto, MD, MPH, Yukio Miki, MD, PhD, Daiju Ueda, MD, PhD

**Table of Contents:**

**Section S1: Supplementary Figures**

 Appendix Figure S1: Forest plot and funnel plot for overall accuracy of generative AI models

 Appendix Figure S2: Forest plots for model accuracy

 Appendix Figure S3: Forest plots for medical specialty

**Section S2: Supplementary Tables**

 Appendix Table S1: PROBAST modifications

 Appendix Table S2: Detailed study characteristics

**Section S1: Supplementary Figures**

Appendix Figure S1: Forest plot and funnel plot for overall accuracy of generative AI models



The left panel (A) presents a forest plot depicting the pooled accuracy of generative AI, with an overall estimate at 57% and a 95% confidence interval ranging from 51% to 63%. Each line represents a different study or subset within the meta-analysis. The highlighted area indicates the overall pooled accuracy estimate with its confidence interval. The right panel (B) shows a funnel plot. The scatter of points suggests variations in standard errors across studies of different effect sizes. The horizontal axis represents the effect size, measured as the accuracy of AI models, and the vertical axis denotes the standard error, inversely related to study size. Each dot represents an individual study, with its position determined by the study's accuracy and precision. The dashed diagonal lines, known as 'funnel lines', are drawn at 95% confidence intervals around the mean effect size, indicating the expected distribution of studies in the absence of heterogeneity or bias. The Egger test z-value is 0.565 and p-value is 0.572, illustrating no potential publication bias.

Appendix Figure S2: Forest plots for model accuracy



This figure presents forest plots for model accuracy across eight generative AI models. Panel (A) displays GPT-3.5 with a pooled accuracy of 60% and a 95% confidence interval (CI) ranging from 51–69%. Panel (B) shows GPT-4 with a pooled accuracy of 60% and a 95% CI of 51–68%. Panel (C) illustrates GPT-4V with a pooled accuracy of 53% and a 95% CI from 24–82%. Panel (D) features Llama 2 with a pooled accuracy of 22% and a 95% CI spanning 0–48%. Panel (E) depicts PaLM2 with a pooled accuracy of 43% and a 95% CI between 35–50%. Panel (F) represents Prometheus with a pooled accuracy of 73% and a 95% CI of 58–89%. Panel (G) visualizes Glass with a pooled accuracy of 33% and a 95% CI extending from 4–78%. Lastly, Panel (H) reveals Med-42 with a pooled accuracy of 58% and a 95% CI from 52–63%.

Appendix Figure S3: Forest plots for medical specialty

****

This figure delineates forest plots for the diagnostic accuracy of generative AI models across various medical specialties. Panel (A) shows General medicine with a pooled accuracy of 62% and a 95% confidence interval (CI) of 55–69%. Panel (B) depicts Gastroenterology with a pooled accuracy of 80% and a 95% CI of 56–94%. Cardiology is presented in Panel (C) with a pooled accuracy of 53% and a 95% CI of 45–60%. Panel (D) illustrates Neurology with a pooled accuracy of 46% and a 95% CI of 38–53%. Emergency medicine is represented in Panel (E) with a pooled accuracy of 53% and a 95% CI of 35–71%. Panel (F) reflects Pediatrics with a pooled accuracy of 93% and a 95% CI of 78–99%. Gynecology, in Panel (G), shows a pooled accuracy of 90% and a 95% CI of 73–98%. Urology, in Panel (H), exhibits a pooled accuracy of 92% and a 95% CI of 89–95%. Otolaryngology is visualized in Panel (I) with a pooled accuracy of 86% and a 95% CI of 42–100%. Panel (J) conveys Endocrinology with a pooled accuracy of 8% and a 95% CI of 0–18%. Orthopedic surgery is shown in Panel (K) with a pooled accuracy of 93% and a 95% CI of 77–99%. Rheumatology is detailed in Panel (L) with a pooled accuracy of 35% and a 95% CI of 27–44%. Panel (M) presents Psychiatry with a pooled accuracy of 53% and a 95% CI of 40–66%. Ophthalmology is displayed in Panel (N) with a pooled accuracy of 74% and a 95% CI of 65–83%. Dermatology, in Panel (O), reports a pooled accuracy of 54% and a 95% CI of 37–70%. Plastic surgery is depicted in Panel (P) with a pooled accuracy of 88% and a 95% CI of 62–98%. Lastly, Panel (Q) shows Radiology with a pooled accuracy of 44% and a 95% CI of 29–59%.

**Section S2: Supplementary Tables**

Appendix Table S1: PROBAST modifications

|  |  |
| --- | --- |
| PROBAST Items | Modifications |
| Domain 1: Participants | No changes (Refer to participant data for diagnosis) |
| Domain 2: Predictors | N/A–removed from scoring |
| Domain 3: Outcome | Items 3.3, 3.5 and 3.6 N/A |
| Domain 4: Analysis  | Items 4.5, 4.6, and 4.9 N/A |
| Domain 5: Overall | No changes |

Appendix Table S2: Detailed study characteristics

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **DOI** | **Publication year** | **First author** | **Model** | **Version** | **Database explanation** | **Reference standard** | **ROB for participants** | **Applicability for participants** | **ROB for outcome** | **Applicability for outcome** | **ROB for analysis** |
| [12](https://paperpile.com/c/Dvd6J9/soXA) | 2023 | Ueda | GPT-4 | Not written | The quiz “Diagnosis Please” in Radiology | Answer | Low | High | Low | Low | Low |
| [13](https://paperpile.com/c/Dvd6J9/UD2s) | 2023 | Kanjee | GPT-4 | Not written | Clinical vignettes from case reports in NEJM  | Expert consensus | Low | Low | Low | Low | High |
| [21](https://paperpile.com/c/Dvd6J9/5GWu) | 2023 | Chee | GPT-3.5 | Not written | Hypothetical clinical vignettes representing vertigo | Expert consensus | Unclear | Low | High | Low | High |
| [22](https://paperpile.com/c/Dvd6J9/hhYv) | 2023 | Lyons | Prometheus, GPT-4 | Not written | Hypothetical clinical vignettes representing common ophthalmic complaints | Expert consensus | Low | Low | Low | Low | High |
| [23](https://paperpile.com/c/Dvd6J9/rW69) | 2023 | Hirosawa | GPT-3.5, GPT-4 | March 14, 2023, March 25, 2023 | Clinical vignettes about general internal medicine | Expert consensus | Low | Low | Low | Low | High |
| [24](https://paperpile.com/c/Dvd6J9/zE3R) | 2023 | Benoit | GPT-3.5 | January 9, 2023 | Clinical vignettes representing various common complaints | Expert consensus | Low | Low | Unclear | Low | High |
| [25](https://paperpile.com/c/Dvd6J9/Wj6i) | 2023 | Hirosawa | GPT-3.5 | December 15, 2023 | Clinical vignettes about general internal medicine | Expert consensus | Low | Low | Low | Low | High |
| [26](https://paperpile.com/c/Dvd6J9/2MoR) | 2023 | Wei | GPT-4 | Not written | Clinical vignettes about pediatrics | Expert consensus | Low | Low | Low | Low | High |
| [27](https://paperpile.com/c/Dvd6J9/6WNk) | 2023 | Ueda | GPT-4 | March 23, 2023 | The quiz “Image Challenge” in NEJM  | Answer | Low | High | Low | Low | High |
| [28](https://paperpile.com/c/Dvd6J9/Nx05) | 2023 | Allahqoli | GPT-3.5 | Not written | Clinical vignettes about obstetrics and gynecology | Expert consensus | Low | Low | Low | Low | High |
| [29](https://paperpile.com/c/Dvd6J9/qej9) | 2023 | Levartovsky | GPT-4 | Not written | Clinical vignettes about patients with ulcerative colitis | Expert consensus | Low | Low | Low | Low | High |
| [30](https://paperpile.com/c/Dvd6J9/7JId) | 2023 | Bushuven | GPT-3.5, GPT-4 | March 23, 2023 | Clinical vignettes about emergency medicine | Expert consensus | Low | Low | Low | Low | High |
| [31](https://paperpile.com/c/Dvd6J9/FP1X) | 2023 | Knebel | GPT-3.5 | March 14, 2023 | Hypothetical clinical vignettes representing acute ocular symptoms | Expert consensus | Low | Low | Low | Low | High |
| [32](https://paperpile.com/c/Dvd6J9/B2SE) | 2023 | Mitsuyama | GPT-4 | May 24, 2023 | Clinical vignettes about imaging findings of brain tumors | Expert consensus | Low | Low | Low | Low | High |
| [33](https://paperpile.com/c/Dvd6J9/Da9W) | 2023 | Pillai | GPT-3.5, GPT-4, Llama 2 | Not written | Clinical vignettes about FMF and DIRA | Expert consensus | Low | Low | Low | Low | High |
| [34](https://paperpile.com/c/Dvd6J9/8jAp) | 2023 | Brin | GPT-4V | Not written | Clinical vignettes about imaging findings | Expert consensus | Low | Low | Low | Low | High |
| [35](https://paperpile.com/c/Dvd6J9/UFqY) | 2023 | Horiuchi | GPT-4 | May 24, 2023 | The quiz "Freiburg Neuropathology Case Conference" in Clinical Neuroradiology | Answer | Low | High | Low | Low | High |
| [36](https://paperpile.com/c/Dvd6J9/rmXI) | 2023 | Ito | GPT-4 | March 14, 2023 | Clinical vignettes representing various common complaints | Expert consensus | Low | Low | Low | Low | High |
| [37](https://paperpile.com/c/Dvd6J9/65RC) | 2023 | Horiuchi | GPT-4, GPT-4V | September 25, 2023 | The quiz "Test yourself" in Skeletal Radiology | Answer | Low | High | Low | Low | Low |
| [38](https://paperpile.com/c/Dvd6J9/sFW1) | 2023 | Madadi | GPT-3.5, GPT-4 | Not written | Clinical vignettes about ophthalmology | Expert consensus | Low | Low | Low | Low | High |
| [39](https://paperpile.com/c/Dvd6J9/I20r) | 2023 | Sorin | GPT-4V | Not written | Clinical vignettes about patients with ocular symptoms | Expert consensus | Low | Low | Low | Low | High |
| [40](https://paperpile.com/c/Dvd6J9/esRb) | 2023 | Delsoz | GPT-3.5, GPT-4 | Not written | Clinical vignettes about ophthalmology | Expert consensus | Low | Low | Low | Low | High |
| [41](https://paperpile.com/c/Dvd6J9/6opV) | 2023 | Levine | GPT-3 | Not written | Hypothetical clinical vignettes representing various common complaints | Expert consensus | Low | Low | Low | Low | High |
| [42](https://paperpile.com/c/Dvd6J9/vuS0) | 2023 | Schubert | GPT-4V | Not written | The quiz “Image Challenge” in NEJM  | Answer | Low | High | Low | Low | High |
| [43](https://paperpile.com/c/Dvd6J9/ZVRC) | 2023 | Sultan | GPT-3.5 | Not written | Clinical vignettes about patients with cancer predisposition syndromes | Expert consensus | Low | Low | Low | Low | High |
| [44](https://paperpile.com/c/Dvd6J9/qUEP) | 2023 | Kiyohara | PaLM2, GPT-3.5, GPT-4 | Not written | Clinical vignettes about vasospastic angina and acute coronary syndrome | Expert consensus | Low | Low | Low | Low | High |
| [45](https://paperpile.com/c/Dvd6J9/vW9r) | 2023 | Horiuchi | GPT-4 | August 3, 2023 | The quiz “Case of the Week” in AJNR | Answer | Low | High | Low | Low | Low |
| [46](https://paperpile.com/c/Dvd6J9/DEUe) | 2023 | Stoneham | GPT-4 | Not written | Clinical vignettes about patients with cutaneous symptoms | Expert consensus | Low | Low | Low | Low | High |
| [47](https://paperpile.com/c/Dvd6J9/7RBC) | 2023 | Rundle | GPT-3.5 | May 24, 2023 | Hypothetical clinical vignettes representing cutaneous tumors | Expert consensus | Low | Low | Low | Low | High |
| [48](https://paperpile.com/c/Dvd6J9/ZW9l) | 2023 | Rojas-Carabali | GPT-3.5, GPT-4, Glass | Not written | Clinical vignettes about patients with uveitis | Expert consensus | Low | Low | Low | Low | High |
| [49](https://paperpile.com/c/Dvd6J9/DKdd) | 2023 | Fraser | GPT-3.5, GPT-4 | Not written | Clinical vignettes about emergency medicine | Expert consensus | Low | Low | Low | Low | High |
| [50](https://paperpile.com/c/Dvd6J9/v01g) | 2023 | Krusche | GPT-4 | Not written | Clinical vignettes about rheumatology | Expert consensus | Low | Low | Low | Low | Low |
| [51](https://paperpile.com/c/Dvd6J9/xwhk) | 2023 | Galetta | GPT-4 | Not written | Clinical vignettes about neurology | Expert consensus | Low | Low | Low | Low | High |
| [52](https://paperpile.com/c/Dvd6J9/e80Z) | 2023 | Delsoz | GPT-3.5 | Not written | Clinical vignettes about ophthalmology | Expert consensus | Low | Low | Low | Low | High |
| [53](https://paperpile.com/c/Dvd6J9/4Iep) | 2023 | Hu | GPT-4 | Not written | Clinical vignettes about ophthalmology | Expert consensus | Low | Low | Low | Low | High |
| [54](https://paperpile.com/c/Dvd6J9/fiVX) | 2023 | Hirosawa | PaLM2 | Not written | Clinical vignettes representing various common complaints | Expert consensus | Low | Low | Low | Low | High |
| [55](https://paperpile.com/c/Dvd6J9/aMa0) | 2023 | Abi-Rafeh | GPT-3.5 | Not written | Hypothetical clinical vignettes about plastic surgery | Expert consensus | Low | Low | Low | Low | High |
| [56](https://paperpile.com/c/Dvd6J9/j210) | 2023 | Shea | GPT-4 | Not written | Clinical vignettes about patients who had delay of definitive diagnosis longer than 1 month | Expert consensus | Low | Low | Low | Low | High |
| [57](https://paperpile.com/c/Dvd6J9/Wdiz) | 2023 | Koga | PaLM2, GPT-3.5, GPT-4 | Not written | Clinical vignettes about neuropathology | Expert consensus | Low | Low | Low | Low | High |
| [58](https://paperpile.com/c/Dvd6J9/Zz4y) | 2023 | Xv | GPT-3.5 | Not written | Clinical vignettes about urology | Expert consensus | Low | Low | Low | Low | Low |
| [59](https://paperpile.com/c/Dvd6J9/rnDl) | 2023 | Reese | GPT-4 | Not written | Clinical vignettes from case reports in NEJM  | Expert consensus | Low | Low | Low | Low | High |
| [60](https://paperpile.com/c/Dvd6J9/v42h) | 2023 | Han | GPT-3.5, GPT-4, GPT-4V, Llama 2,Med-42 | Not written | The quiz “Image Challenge” in NEJM and the quiz "Image of the Month" in JAMA | Answer | Low | High | Low | Low | Unclear |
| [61](https://paperpile.com/c/Dvd6J9/q3Kq) | 2023 | Senkaiahliyan | GPT-4V | Not written | Clinical vignettes about various medical images | Expert consensus | Low | Low | Low | Low | High |
| [62](https://paperpile.com/c/Dvd6J9/cw0I) | 2023 | Williams | GPT-3.5 | March 1, 2023 | Clinical vignettes about emergency medicine | Expert consensus | Low | Low | Low | Low | Low |
| [63](https://paperpile.com/c/Dvd6J9/uYQG) | 2023 | Tenner | GPT-3.5 | Not written | Clinical vignettes from case reports in NEJM  | Expert consensus | Low | Low | Low | Low | High |
| [64](https://paperpile.com/c/Dvd6J9/blAl) | 2023 | Mori | GPT-4 | Not written | Clinical vignettes about imaging findings of SAPHO syndrome | Expert consensus | Low | Low | Low | Low | Low |
| [65](https://paperpile.com/c/Dvd6J9/ZC70) | 2023 | Mykhalko | GPT-3.5 | Not written | The quiz “Case Challenges” from the Medscape website | Answer | Low | High | Low | Low | High |
| [66](https://paperpile.com/c/Dvd6J9/BVpO) | 2023 | Andrade-Castellanos | GPT-3.5 | Not written | The quiz “Test yourself” from the American College of Physicians website | Answer | Low | High | Low | Low | High |
| [67](https://paperpile.com/c/Dvd6J9/aRQ5) | 2023 | Daher | GPT-3.5 | Not written | Clinical vignettes about patients with joint symptoms | Expert consensus | Low | Low | Low | Low | High |
| [68](https://paperpile.com/c/Dvd6J9/bMa9) | 2023 | Suthar | GPT-4 | July 20, 2023 | The quiz “Case of the Month” in AJNR | Answer | Low | High | Low | Low | Low |
| [69](https://paperpile.com/c/Dvd6J9/4vUT) | 2023 | Nakaura | Prometheus, GPT-3.5 | Not written, June 13, 2023 | Clinical vignettes about various imaging findings | Expert consensus | Low | Low | Low | Low | High |
| [70](https://paperpile.com/c/Dvd6J9/3wc2) | 2023 | Berg | GPT-3.5, GPT-4 | Not written | Clinical vignettes about emergency medicine | Expert consensus | Low | Low | Low | Low | High |
| [71](https://paperpile.com/c/Dvd6J9/Im2C) | 2023 | Gebrael | GPT-4 | Not written | Clinical vignettes about patients with prostate cancer | Expert consensus | Low | Low | Low | Low | High |
| [72](https://paperpile.com/c/Dvd6J9/LEaR) | 2023 | Ravipati | GPT-3.5 | Not written | Clinical vignettes about dermatology | Expert consensus | Low | Low | Low | Low | High |