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

Hirotaka Takita, MD, PhD, Daijiro Kabata, MPH, PhD, Shannon L Walston, MS, Hiroyuki Tatekawa, MD, PhD, 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: Funnel plot

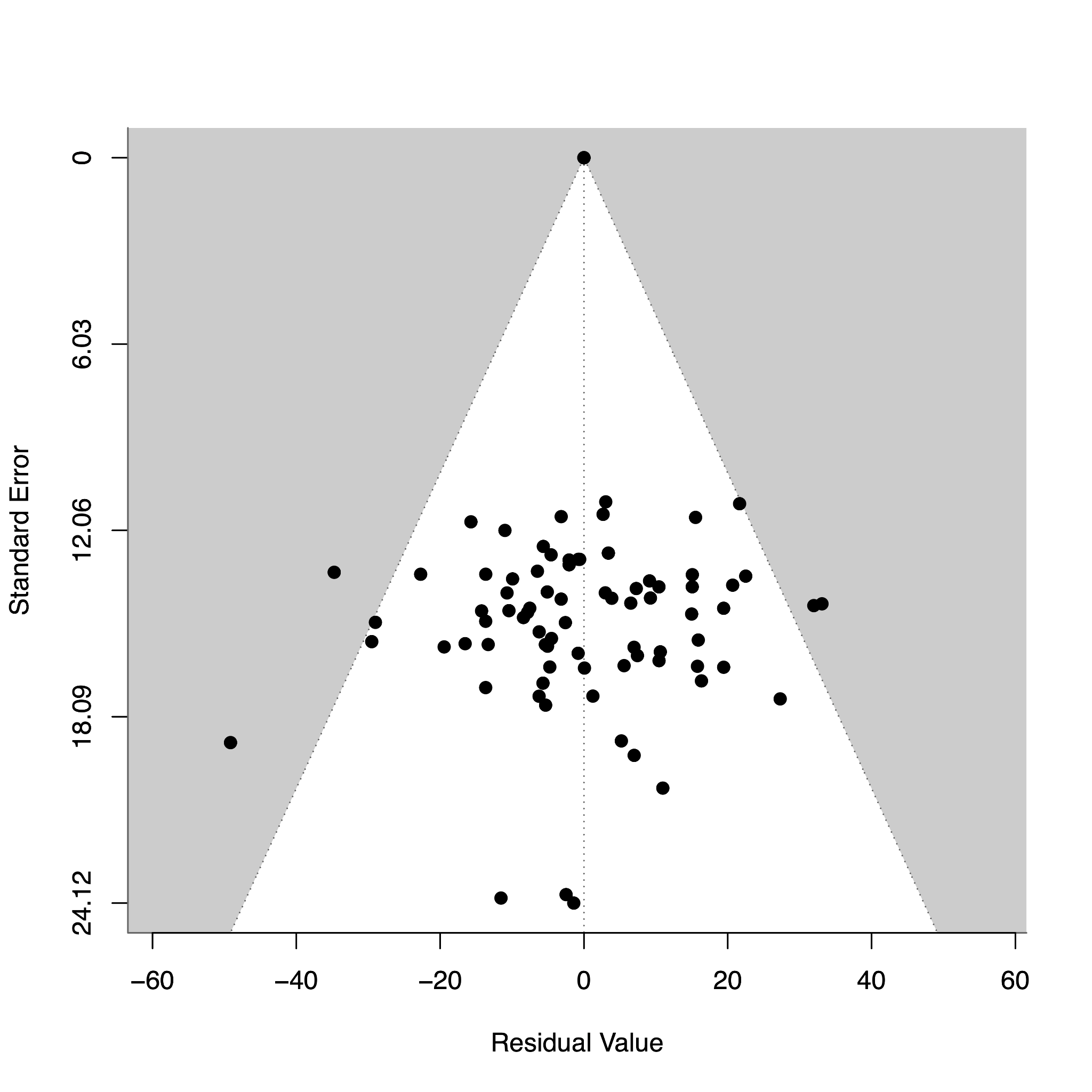
**Section S2: Supplementary Tables**

Appendix Table S1: PROBAST modifications

Appendix Table S2: Detailed study characteristics

**Section S1: Supplementary Figures**

Appendix Figure S1: Funnel plot



The funnel plot illustrates the distribution of the residuals of the fitted values corresponding to their standard errors in the meta-regression. The Egger test results z-value = -2.2094 and p = 0.0271, which indicates the possible presence of publication bias.

**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

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Citation** | **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** |
| [11](https://paperpile.com/c/Dvd6J9/soXA) | 2023 | Ueda | GPT-4 | Not written | The quiz “Diagnosis Please” in Radiology | Answer | Low | High | Low | Low | Low |
| [12](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 |
| [13](https://paperpile.com/c/Dvd6J9/fiVX) | 2023 | Hirosawa | PaLM2 | Not written | Clinical vignettes representing various common complaints | Expert consensus | Low | Low | Low | Low | High |
| [14](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 |
| [19](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 |
| [20](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 |
| [21](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 |
| [22](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 |
| [23](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 |
| [24](https://paperpile.com/c/Dvd6J9/2MoR) | 2023 | Wei | GPT-4 | Not written | Clinical vignettes about pediatrics | Expert consensus | Low | Low | Low | Low | High |
| [25](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 |
| [26](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 |
| [27](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 |
| [28](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 |
| [29](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 |
| [30](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 |
| [31](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 |
| [32](https://paperpile.com/c/Dvd6J9/8jAp) | 2023 | Brin | GPT-4V | Not written | Clinical vignettes about imaging findings | Expert consensus | Low | Low | Low | Low | High |
| [33](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 |
| [34](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 |
| [35](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 |
| [36](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 |
| [37](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 |
| [38](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 |
| [39](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 |
| [40](https://paperpile.com/c/Dvd6J9/vuS0) | 2023 | Schubert | GPT-4V | Not written | The quiz “Image Challenge” in NEJM | Answer | Low | High | Low | Low | High |
| [41](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 |
| [42](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 |
| [43](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 |
| [44](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 |
| [45](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 |
| [46](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 |
| [47](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 |
| [48](https://paperpile.com/c/Dvd6J9/v01g) | 2023 | Krusche | GPT-4 | Not written | Clinical vignettes about rheumatology | Expert consensus | Low | Low | Low | Low | Low |
| [49](https://paperpile.com/c/Dvd6J9/xwhk) | 2023 | Galetta | GPT-4 | Not written | Clinical vignettes about neurology | Expert consensus | Low | Low | Low | Low | High |
| [50](https://paperpile.com/c/Dvd6J9/e80Z) | 2023 | Delsoz | GPT-3.5 | Not written | Clinical vignettes about ophthalmology | Expert consensus | Low | Low | Low | Low | High |
| [51](https://paperpile.com/c/Dvd6J9/4Iep) | 2023 | Hu | GPT-4 | Not written | Clinical vignettes about ophthalmology | Expert consensus | Low | Low | Low | Low | High |
| [52](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 |
| [53](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 |
| [54](https://paperpile.com/c/Dvd6J9/Zz4y) | 2023 | Xv | GPT-3.5 | Not written | Clinical vignettes about urology | Expert consensus | Low | Low | Low | Low | Low |
| [55](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 |
| [56](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 |
| [57](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 |
| [58](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 |
| [59](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 |
| [60](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 |
| [61](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 |
| [62](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 |
| [63](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 |
| [64](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 |
| [65](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 |
| [66](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 |
| [67](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 |
| [68](https://paperpile.com/c/Dvd6J9/LEaR) | 2023 | Ravipati | GPT-3.5 | Not written | Clinical vignettes about dermatology | Expert consensus | Low | Low | Low | Low | High |