npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01543-z

A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians

Hirotaka Takita[1], Daijiro Kabata[2], Shannon L. Walston[1,3], Hiroyuki Tatekawa[1], Kenichi Saito[4], Yasushi Tsujimoto[5,6,7], Yukio Miki[1] & Daiju Ueda[1,3,8]

While generative artificial intelligence (AI) has shown potential in medical diagnostics, comprehensive evaluation of its diagnostic performance and comparison with physicians has not been extensively explored. We conducted a systematic review and meta-analysis of studies validating generative AI models for diagnostic tasks published between June 2018 and June 2024. Analysis of 83 studies revealed an overall diagnostic accuracy of 52.1%. No significant performance difference was found between AI models and physicians overall ($p = 0.10$) or non-expert physicians ($p = 0.93$). However, AI models performed significantly worse than expert physicians ($p = 0.007$). Several models demonstrated slightly higher performance compared to non-experts, although the differences were not significant. Generative AI demonstrates promising diagnostic capabilities with accuracy varying by model. Although it has not yet achieved expert-level reliability, these findings suggest potential for enhancing healthcare delivery and medical education when implemented with appropriate understanding of its limitations.

In recent years, the advent of generative artificial intelligence (AI) has marked a transformative era in our society[1–8]. These advanced computational systems have demonstrated exceptional proficiency in interpreting and generating human language, thereby setting new benchmarks in AI's capabilities. Generative AI, with its deep learning architectures, has rapidly evolved, showcasing a remarkable understanding of complex language structures, contexts, and even images. This evolution has not only expanded the horizons of AI but also opened new possibilities in various fields, including healthcare[9].

The integration of generative AI models in the medical domain has spurred a growing body of research focusing on their diagnostic capabilities[10]. Studies have extensively examined the performance of these models in interpreting clinical data, understanding patient histories, and even suggesting possible diagnoses[11,12]. In medical diagnosis, the accuracy, speed, and efficiency of generative AI models in processing vast amounts of medical literature and patient information have been highlighted, positioning them as valuable tools. This research has begun to outline the strengths and limitations of generative AI models in diagnostic tasks in healthcare.

Despite the growing research on generative AI models in medical diagnostics, there remains a significant gap in the literature: a comprehensive meta-analysis of the diagnostic capabilities of the models, followed by a comparison of their performance with that of physicians. Such a comparison is crucial for understanding the practical implications and effectiveness of generative AI models in real-world medical settings. While individual studies have provided insights into the capabilities of generative AI models[13,14], a systematic review and meta-analysis are necessary to aggregate these findings and draw more robust conclusions about their comparative effectiveness against traditional diagnostic practices by physicians.

This paper aims to bridge the existing gap in the literature by conducting a meticulous meta-analysis of the diagnostic capabilities of generative AI models in healthcare. Our focus is to provide a comprehensive diagnostic performance evaluation of generative AI models and compare their diagnostic performance with that of physicians. By synthesizing the findings from various studies, we endeavor to offer a nuanced understanding of the effectiveness, potential, and limitations of generative AI

[1]Department of Diagnostic and Interventional Radiology, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan. [2]Center for Mathematical and Data Science, Kobe University, Kobe, Japan. [3]Department of Artificial Intelligence, Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan. [4]Center for Digital Transformation of Health Care, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [5]Oku Medical Clinic, Osaka, Japan. [6]Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto University, Kyoto, Japan. [7]Scientific Research WorkS Peer Support Group (SRWS-PSG), Osaka, Japan. [8]Center for Health Science Innovation, Osaka Metropolitan University, Osaka, Japan. e-mail: ai.labo.ocu@gmail.com

models in medical diagnostics. This analysis is intended to serve as a foundational reference for future research and practical applications in the field, ultimately contributing to the advancement of AI-assisted diagnostics in healthcare.

Results

Study selection and characteristics

We identified 18,371 studies, of which 10,357 were duplicates. After screening, 83 studies were included in the meta-analysis[11–93] (Fig. 1 and Table 1). The most evaluated models were GPT-4[3] (54 articles) and GPT-3.5[2] (40), while models such as GPT-4V[94] (9), PaLM2[7] (9), Llama 2[5] (5), Prometheus (4), Claude 3 Opus (4)[95], Gemini 1.5 Pro (3)[96], GPT-4o (2)[97], Llama 3 70B (2)[98], Claude 3 Sonnet (2)[95], and Perplexity (2)[99] had less representation. Each other model, including Aya[100], Claude 2[101], Claude 3.5 Sonnet[102], Clinical Camel[103], Gemini 1.0[104], Gemini 1.0 Pro[104], Gemini 1.5 Flash[96], Gemini Pro, Glass[105], GPT-3[2], Llama 3 8B[98], Med-42[106], MedAlpaca[107], Meditron[108], Mistral 7B[109], Mistral Large, Mixtral8x22B[110], Mixtral8x7B[110], Nemotron[111], Open Assistant[112], WizardLM[113], was used in only one article. Details of each model are in Supplementary Table 1 (online). The review spanned a wide range of medical specialties, with General medicine being the most common (27 articles). Other specialties like Radiology (16), Ophthalmology (11), Emergency medicine (8), Neurology (4), Dermatology (4), Otolaryngology (2), and Psychiatry (2) were represented, as well as Gastroenterology, Cardiology, Pediatrics, Urology, Endocrinology, Gynecology, Orthopedic surgery, Rheumatology, and Plastic surgery with one article each. Regarding model tasks, free text tasks were the most common, with 73 articles, followed by choice tasks at 15. For test dataset types, 59 articles involved external testing, while 25 were unknown because the training data for the generative AI models was unknown. Of the included studies, 71 were peer-reviewed, while 12 were preprints. Study characteristics are shown in Table 1 and Supplementary Table 2 (online). Seventeen studies compared the performance of generative AI models with that of physicians[36,39,43,55,72,74,75,77,78,80,81,90,93]. GPT-4 (11 articles) and GPT- 3.5 (11) were the most frequently compared with physicians, followed by GPT-4V (3), Llama 2 (2), Claude 3 Opus (1), Claude 3 Sonnet (1), Claude 3.5 Sonnet (1), Clinical Camel (1), Gemini 1.0 (1), Gemini 1.5 Flash (1), Gemini 1.5 Pro (1), Gemini Pro (1), GPT-4o (1), Llama 3 70B (1), Meditron (1), Mistral Large (1), Open Assistant (1), PaLM2 (1), Perplexity (1), Prometheus (1), and WizardLM (1).

Quality assessment

Prediction Model Study Risk of Bias Assessment Tool (PROBAST) assessment led to an overall rating of 63/83 (76%) studies at high risk of bias, 20/83 (24%) studies at low risk of bias, 18/83 (22%) studies at high concern for generalizability, and 65/83 (78%) studies at low concern for generalizability[114] (Fig. 2). The main factors of this evaluation were studies that evaluated models with a small test set and studies that cannot prove external evaluation due to the unknown training data of generative AI models. Detailed results are shown in Supplementary Table 2 (online).

Meta-analysis

The overall accuracy for generative AI models was found to be 52.1% with a 95% CI of 47.0–57.1%. The meta-analysis demonstrated no significant performance difference between generative AI models overall and physicians (physicians' accuracy was 9.9% higher [95% CI: −2.3 to 22.0%], $p = 0.10$) and non-expert physicians (non-expert physicians' accuracy was 0.6% higher [95% CI: −14.5 to 15.7%], $p = 0.93$), whereas generative AI models overall were significantly inferior to expert physicians (difference in accuracy: 15.8% [95% CI: 4.4–27.1%], $p = 0.007$, Fig. 3). Interestingly, several models, including GPT-4, GPT-4o, Llama3 70B, Gemini 1.0 Pro, Gemini 1.5 Pro, Claude 3 Sonnet, Claude 3 Opus, and Perplexity, demonstrated slightly higher performance compared to non-experts, although the differences were not significant. GPT-3.5, GPT-4, Llama2, Llama3 8B, PaLM2, Mistral 7B, Mixtral8x7B, Mixtral8x22B, and Med-42 were significantly inferior when compared to expert physicians, whereas GPT-4V, GPT-4o, Prometheus, Llama 3 70B, Gemini 1.0 Pro, Gemini 1.5 Pro, Claude 3 Sonnet, Claude 3 Opus, and Perplexity demonstrated no significant difference against experts.

In our meta-regression, we also found no significant difference in performance between general medicine and various specialties, except for

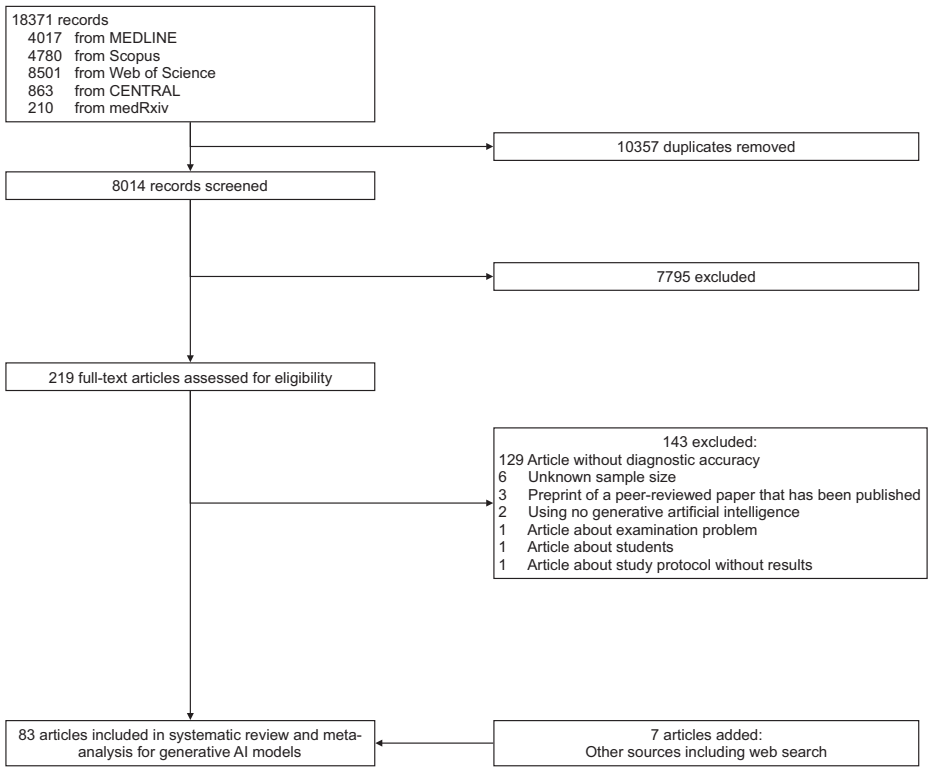

Fig. 1 | Eligibility criteria. The flow diagram illustrates the systematic review process, starting with 18,371 initial records identified from multiple databases: 4017 from MEDLINE, 4780 from Scopus, 8501 from Web of Science, 863 from CENTRAL, and 210 from medRxiv. After removing 10,357 duplicates, 8014 records were screened. Of these, 7795 were excluded as they did not align with the objectives of this systematic review, leaving 219 full-text articles for eligibility assessment. Further evaluation resulted in 143 exclusions due to various reasons: 129 articles without diagnostic accuracy, 6 with unknown sample size, 3 preprints of already published peer-reviewed papers, 2 not using generative artificial intelligence, 1 article about examination problems, 1 about students, and 1 about study protocol without results. Seven additional articles were identified through other sources including web search, resulting in a final total of 83 articles included in the systematic review and meta-analysis focusing on generative AI models.

Table 1 | Study characteristics

Citation	First author	Model	Model task	Test type	Specialty	Comparison group	Eligible	Preprint	Overall ROB	Overall applicability
11	Ueda	GPT-4	Free text	External	Radiology	NA	313	Peer-reviewed	Low	High
12	Kanjee	GPT-4	Free text	External	General medicine	NA	70	Peer-reviewed	High	Low
13	Hirosawa	PaLM2 (Bard)	Free text	External	General medicine	NA	82	Peer-reviewed	High	Low
14	Shea	GPT-4	Free text	External	General medicine	NA	6	Peer-reviewed	High	Low
15	Chee	GPT-3.5	Free text	External	Otolaryngology	NA	7	Peer-reviewed	High	Low
16	Lyons	Prometheus (Bing), GPT-4	Free text, Choice	External	Ophthalmology	NA	44	Peer-reviewed	High	Low
17	Benoit	GPT-3.5	Free text, Choice	Unknown	General medicine	NA	45	Preprint	High	Low
18	Hirosawa	GPT-3.5, GPT-4	Free text	Unknown	General medicine	NA	52	Peer-reviewed	High	Low
19	Hirosawa	GPT-3.5	Free text	External	General medicine	NA	30	Peer-reviewed	High	Low
20	Wei	GPT-4	Choice	External	Psychiatry	NA	60	Peer-reviewed	High	Low
21	Allahqoli	GPT-3.5	Free text	Unknown	Gynecology	NA	30	Peer-reviewed	High	Low
22	Levartovsky	GPT-4	Choice	External	Gastroenterology	NA	20	Peer-reviewed	High	Low
23	Bushuven	GPT-3.5, GPT-4	Free text, Choice	External	Emergency medicine	NA	22	Peer-reviewed	High	Low
24	Knebel	GPT-3.5	Free text, Choice	External	Ophthalmology	NA	10	Peer-reviewed	High	Low
25	Pillai	GPT-3.5, GPT-4, Llama 2	Free text	Unknown	Endocrinology	Expert	20	Peer-reviewed	High	Low
26	Ito	GPT-4	Free text, Choice	Unknown	General medicine	Expert	45	Peer-reviewed	High	Low
27	Sorin	GPT-4V	Free text	External	Ophthalmology	Non-expert	40	Preprint	High	Low
28	Madadi	GPT-3.5, GPT-4	Free text	Unknown	Ophthalmology	Expert	22	Preprint	High	Low
29	Schubert	GPT-4V	Free text	External	General medicine	NA	93	Preprint	High	High
30	Kiyohara	PaLM2 (Bard), GPT-3.5, GPT-4	Choice	Unknown	Cardiology	NA	66	Preprint	High	Low
31	Sultan	GPT-3.5	Free text	External	Pediatrics	NA	30	Peer-reviewed	High	Low
32	Horiuchi	GPT-4	Free text	External	Radiology	NA	100	Peer-reviewed	Low	High
33	Stoneham	GPT-4	Free text	External	Dermatology	NA	36	Peer-reviewed	High	Low
34	Rundle	GPT-3.5	Free text	External	Dermatology	NA	39	Peer-reviewed	High	Low

Table 1 (continued) | Study characteristics

Citation	First author	Model	Model task	Test type	Specialty	Comparison group	Eligible	Preprint	Overall ROB	Overall applicability
35	Rojas-Carabali	GPT-3.5, GPT-4, Glass	Free text	External	Ophthalmology	NA	6	Peer-reviewed	High	Low
36	Fraser	GPT-3.5, GPT-4	Free text	Unknown	Emergency medicine	Expert	30	Peer-reviewed	High	Low
37	Krusche	GPT-4	Free text	External	Rheumatology	NA	132	Peer-reviewed	Low	Low
38	Galetta	GPT-4	Free text	External	Neurology	NA	24	Peer-reviewed	High	Low
39	Delsoz	GPT-3.5	Free text	Unknown	Ophthalmology	Non-expert	11	Peer-reviewed	High	Low
40	Hu	GPT-4	Free text	Unknown	Ophthalmology	NA	10	Peer-reviewed	High	Low
41	Abi-Rafeh	GPT-3.5	Free text	External	Plastic surgery	NA	16	Peer-reviewed	High	Low
42	Koga	PaLM2 (Bard), GPT-3.5, GPT-4	Free text	External	Neurology	NA	25	Peer-reviewed	High	Low
43	Xv	GPT-3.5	Free text	External	Urology	Non-expert	306	Peer-reviewed	Low	Low
44	Senthujan	GPT-4V	Free text	Unknown	Radiology	NA	69	Preprint	High	Low
45	Mori	GPT-4	Choice	External	Radiology	NA	151	Peer-reviewed	Low	Low
46	Mykhalko	GPT-3.5	Free text	External	General medicine	NA	50	Peer-reviewed	High	High
47	Andrade-Castellanos	GPT-3.5	Free text	External	General medicine	NA	10	Peer-reviewed	High	High
48	Daher	GPT-3.5	Free text	External	Orthopedic surgery	NA	29	Peer-reviewed	High	Low
49	Suthar	GPT-4	Free text	External	Radiology	NA	140	Peer-reviewed	Low	High
50	Nakaura	Prometheus (Bing), GPT-3.5	Free text	External	Radiology	NA	28	Peer-reviewed	High	Low
51	Berg	GPT-3.5, GPT-4	Free text	External	Emergency medicine	NA	30	Peer-reviewed	High	Low
52	Gebrael	GPT-4	Choice	External	Emergency medicine	NA	56	Peer-reviewed	High	Low
53	Ravipati	GPT-3.5	Free text	Unknown	Dermatology	NA	32	Peer-reviewed	High	Low
54	Shikino	GPT-4	Free text	External	General medicine	NA	25	Peer-reviewed	High	Low
55	Horiuchi	GPT-4, GPT-4V	Free text	External	Radiology	Non-expert, Expert	32	Peer-reviewed	High	High
56	Kumar	PaLM2 (Bard), GPT-3.5, GPT-4, Perplexity	Free text	External	Neurology	NA	20	Peer-reviewed	High	Low
57	Chiu	PaLM2 (Bard), Claude 2, GPT-4	Free text	External	General medicine	NA	104	Peer-reviewed	Low	Low

https://doi.org/10.1038/s41746-025-01543-z

Table 1 (continued) | Study characteristics

Citation	First author	Model	Model task	Test type	Specialty	Comparison group	Eligible	Preprint	Overall ROB	Overall applicability
58	Kikuchi	GPT-3.5, GPT-4	Free text	External	Radiology	NA	115	Peer-reviewed	Low	High
59	Bridges	GPT-4	Free text	External	General medicine	NA	201	Peer-reviewed	High	Low
60	Shieh	GPT-4	Free text	External	General medicine	NA	63	Peer-reviewed	High	Low
61	Warrier	PaLM2 (Bard), Prometheus (Bing), GPT-3.5, GPT-4	Free text	Unknown	Otolaryngology	NA	100	Peer-reviewed	High	Low
62	Han	GPT-3.5, GPT-4, GPT-4V, Gemini 1.0 Pro, Llama 2, Med-42	Choice	External	General medicine	NA	140, 348	Peer-reviewed	Low	High
63	Milad	GPT-4	Choice	Unknown	Ophthalmology	NA	422	Peer-reviewed	High	High
64	Abdullahi	PaLM2 (Bard), GPT-3.5, GPT-4, MedAlpaca	Free text	Unknown	General medicine	NA	30	Peer-reviewed	High	High
65	Tenner	GPT-3.5	Free text	External	General medicine	NA	40	Peer-reviewed	High	Low
66	Luk	GPT-4	Free text	External	General medicine	NA	81	Peer-reviewed	High	Low
67	Savage	GPT-4	Free text	External	General medicine	NA	300	Peer-reviewed	Low	Low
68	Franc	GPT-3.5	Choice	Unknown	Emergency medicine	NA	61	Peer-reviewed	High	Low
69	Yang	GPT-3.5, GPT-4	Free text	Unknown	General medicine	NA	238	Peer-reviewed	High	Low
70	Reese	GPT-4	Free text	External	General medicine	NA	75	Preprint	High	Low
71	Olmo	Claude 3 Opus, Claude 3 Sonnet, GPT-4, Gemini 1.5 Pro, Llama 2, Llama 3 70B, Llama 3 8B, Mistral 7B, Mixtral8x22B, Mixtral8x7B	Free text	Unknown, External	General medicine	NA	200, 75	Preprint	Low	Low
72	Cesur	Claude 3 Opus, Claude 3 Sonnet, Claude 3.5 Sonnet, GPT-3.5, GPT-4, GPT-4o, Gemini 1.0, Gemini 1.5 Flash, Gemini 1.5 Pro, Llama 3 70B, Mistral Large, Perplexity	Free text	External	Radiology	Expert	80	Preprint	High	High
73	Schramm	GPT-4V	Free text	External	Neurology	NA	30	Preprint	High	Low
74	Gunes	PaLM2 (Bard), Prometheus (Bing), GPT-3.5	Free text	External	Radiology	Expert	124	Preprint	Low	High
75	Olshaker	GPT-3.5, GPT-4, Gemini Pro	Free text	Unknown	Radiology	Non-expert	60	Preprint	High	Low
76	Hirosawa	PaLM2 (Bard), GPT-4, Llama 2	Free text	External	General medicine	NA	392	Peer-reviewed	Low	Low
77	Mitsuyama	GPT-4	Free text	External	Radiology	Expert, Non-expert	150	Peer-reviewed	Low	Low
78	Yazaki	GPT-3.5, GPT-4	Choice	External	Emergency medicine	Non-expert	100	Peer-reviewed	Low	Low
79	Ghalibafan	GPT-4V	Free text	External	Ophthalmology	NA	143	Peer-reviewed	Low	Low

Table 1 (continued) | Study characteristics

Citation	First author	Model	Model task	Test type	Specialty	Comparison group	Eligible	Preprint	Overall ROB	Overall applicability
80	Hager	Clinical Camel, Llama 2, Meditron, Open Assistant, WizardLM	Free text	Unknown	Emergency medicine	Expert	80	Peer-reviewed	High	Low
81	Horiuchi	GPT-4, GPT-4V	Free text	External	Radiology	Expert, Non-expert	106	Peer-reviewed	Low	High
82	Ríos-Hoyo	GPT-3.5, GPT-4	Free text	External	General medicine	NA	75	Peer-reviewed	High	Low
83	Liu	Claude 3 Opus, GPT-4	Free text	External	Dermatology	NA	100	Peer-reviewed	High	Low
84	Sonoda	Claude 3 Opus, GPT-4o, Gemini 1.5 Pro	Free text	External	Radiology	NA	324	Peer-reviewed	Low	High
85	Wada	GPT-4	Free text	Unknown	Radiology	NA	751	Peer-reviewed	High	High
86	Gargari	Aya, GPT-3.5, GPT-4, Nemotron	Free text	Unknown	Psychiatry	NA	20	Peer-reviewed	High	Low
87	Mihalache	GPT-4	Free text	Unknown	Ophthalmology	NA	69	Peer-reviewed	High	High
88	Rutledge	GPT-4	Free text	Unknown	General medicine	NA	45	Peer-reviewed	High	Low
89	Ueda	GPT-4	Free text	External	General medicine	NA	62	Peer-reviewed	High	High
90	Delsoz	GPT-3.5, GPT-4	Free text	Unknown	Ophthalmology	Expert	20	Peer-reviewed	High	Low
91	Brin	GPT-4V	Free text	External	Radiology	NA	216	Peer-reviewed	Low	Low
92	Levine	GPT-3	Free text	External	General medicine	NA	48	Peer-reviewed	High	Low
93	Williams	GPT-3.5	Choice	External	Emergency medicine	Non-expert	500	Peer-reviewed	Low	Low

ROB risk of bias.

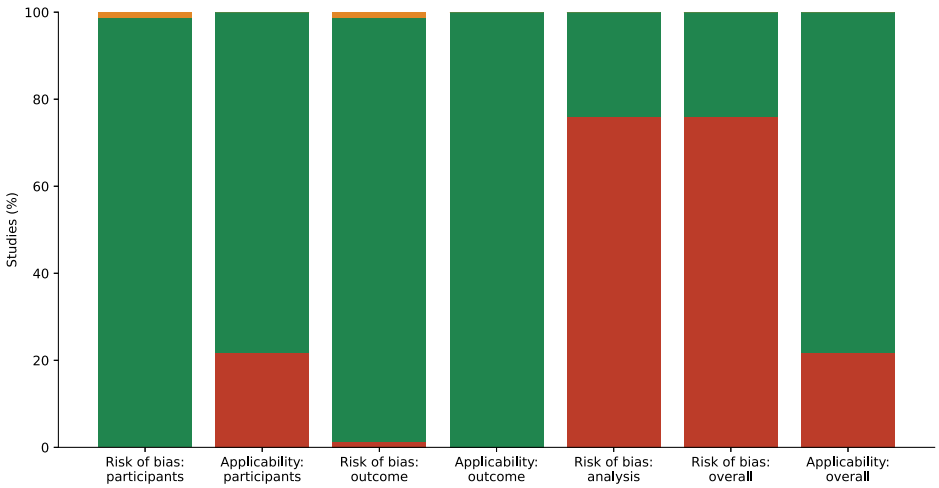

Fig. 2 | Summary of Prediction Model Study Risk of Bias Assessment Tool (PROBAST) risk of bias. Assessment for generative AI model studies included in the meta-analysis ($N = 83$). The participants and the outcome determination were predominantly at low risk of bias, but there was a high risk of bias for analysis (76%) and the overall evaluation (76%). Overall applicability and applicability for participants and outcomes are predominantly of low concern, with 22% at high concern.

Urology and Dermatology, where significant differences were observed (p-values < 0.001, Fig. 4). While medical-domain models demonstrated a slightly higher accuracy (mean difference = 2.1%, 95% CI: −28.6 to 24.3%), this difference was not statistically significant ($p = 0.87$). In the analysis of the low risk of bias subgroup, generative AI models overall demonstrated no significant performance difference compared to physicians overall ($p = 0.069$). Evaluating only studies with a low overall risk of bias showed little change compared to the full dataset results. No significant difference was observed based on the risk of bias ($p = 0.92$) or based on publication status ($p = 0.28$). We assessed publication bias by using a regression analysis to quantify funnel plot asymmetry (Supplementary Fig. 1 [online]), and it suggested a risk of publication bias ($p = 0.045$). In the heterogeneity analysis, the R^2 values (amount of heterogeneity accounted for) were 45.2% for all studies and 57.1% for the studies with a low overall risk of bias, indicating moderate levels of explained variability.

Discussion

In this systematic review and meta-analysis, we analyzed the diagnostic performance of generative AI and physicians. We initially identified 18,371 studies, ultimately including 83 in the meta-analysis. The study spanned various AI models and medical specialties, with GPT-4 being the most evaluated. Quality assessment revealed a majority of studies at high risk of bias. The meta-analysis showed a pooled accuracy of 52.1% (95% CI: 47.0–57.1%) for generative AI models. Some generative AI models showed comparable performance to non-expert physicians although no significant performance difference was found (difference in accuracy: 0.6% [95% CI: −14.5 to 15.7%], $p = 0.93$). In contrast, AI models overall were significantly inferior to expert physicians (difference in accuracy: 15.8% [95% CI: 4.4–27.1%], $p = 0.007$). Our analysis also highlighted no significant differences in effectiveness across most medical fields. To the best of our knowledge, this is the first meta-analysis of generative AI models in diagnostic tasks. This comprehensive study highlights the varied capabilities and limitations of generative AI in medical diagnostics.

The meta-analysis of generative AI models in healthcare reveals crucial insights for clinical practice. Despite the overall modest accuracy of 52% for generative AI models in medical applications, this suggests its potential utility in certain clinical scenarios. Importantly, the similar performance of generative AI models such as GPT-4, GPT-4o, Llama3 70B, Gemini 1.0 Pro, Gemini 1.5 Pro, Claude 3 Sonnet, Claude 3 Opus, and Perplexity to physicians in non-expert scenarios highlights the possibility of AI augmenting healthcare delivery in resource-limited settings or as a preliminary diagnostic tool, thereby potentially increasing accessibility and efficiency in patient care[115]. Further analysis revealed no significant difference in performance between general medicine and various specialties, except for Urology and Dermatology. These findings suggest that while there was some variation in performance across different specialties, generative AI has a wide range of capabilities. However, caution is needed when interpreting the results for Urology and Dermatology. The Urology result is based on a large single study, which may limit its generalizability. Regarding Dermatology, the superior performance may be attributed to the visual nature of the specialty, which aligns well with AI's strengths in pattern recognition. However, it's important to note that Dermatology involves complex clinical reasoning and patient-specific factors that go beyond visual pattern recognition. Further research is needed to elucidate the factors contributing to these specialty-specific differences in generative AI performance.

The studies comparing generative AI and physician performance also offer intriguing perspectives in the context of medical education[116]. At present, the significantly higher accuracy of expert physicians compared to AI models overall emphasizes the irreplaceable value of human judgment and experience in medical decision-making. However, the comparable performance of current generative AI models to physicians in non-expert settings reveals an opportunity for integrating AI into medical training. This could include using AI as a teaching aid for medical students and residents, especially in simulating non-expert scenarios where AI's performance is nearly equivalent to that of healthcare professionals[117]. Such integration could enhance learning experiences, offering diverse clinical case studies and facilitating self-assessment and feedback. Additionally, the narrower performance gap between some generative AI models and physicians, even in expert settings, suggests that AI could be used to supplement advanced medical education, helping to identify areas for improvement and providing supporting information. This approach could foster a more dynamic and adaptive learning environment, preparing future medical professionals for an increasingly digital healthcare landscape.

To examine the impact of the overall risk of bias, we conducted a subgroup analysis of studies with a low overall risk of bias. The result of studies with a low overall risk of bias showed little change compared to that of the full dataset. This result suggests that the high proportion of studies with high overall risk of bias does not substantially affect our study's findings or generalizability. While many generative AI models do not disclose details of their training data, the transparency of training data and its collection period is paramount. Without this transparency, it is impossible to determine whether the test dataset is an external dataset or not and this can lead to bias. Transparency ensures an understanding of the model's knowledge, context, and limitations, aids in identifying potential biases, and facilitates independent replication and validation, which are fundamental to scientific integrity. As generative AI continues to evolve, fostering a culture of rigorous transparency is essential to ensure its safe, effective, and equitable application in clinical settings[118], ultimately enhancing the quality of healthcare delivery and medical education.

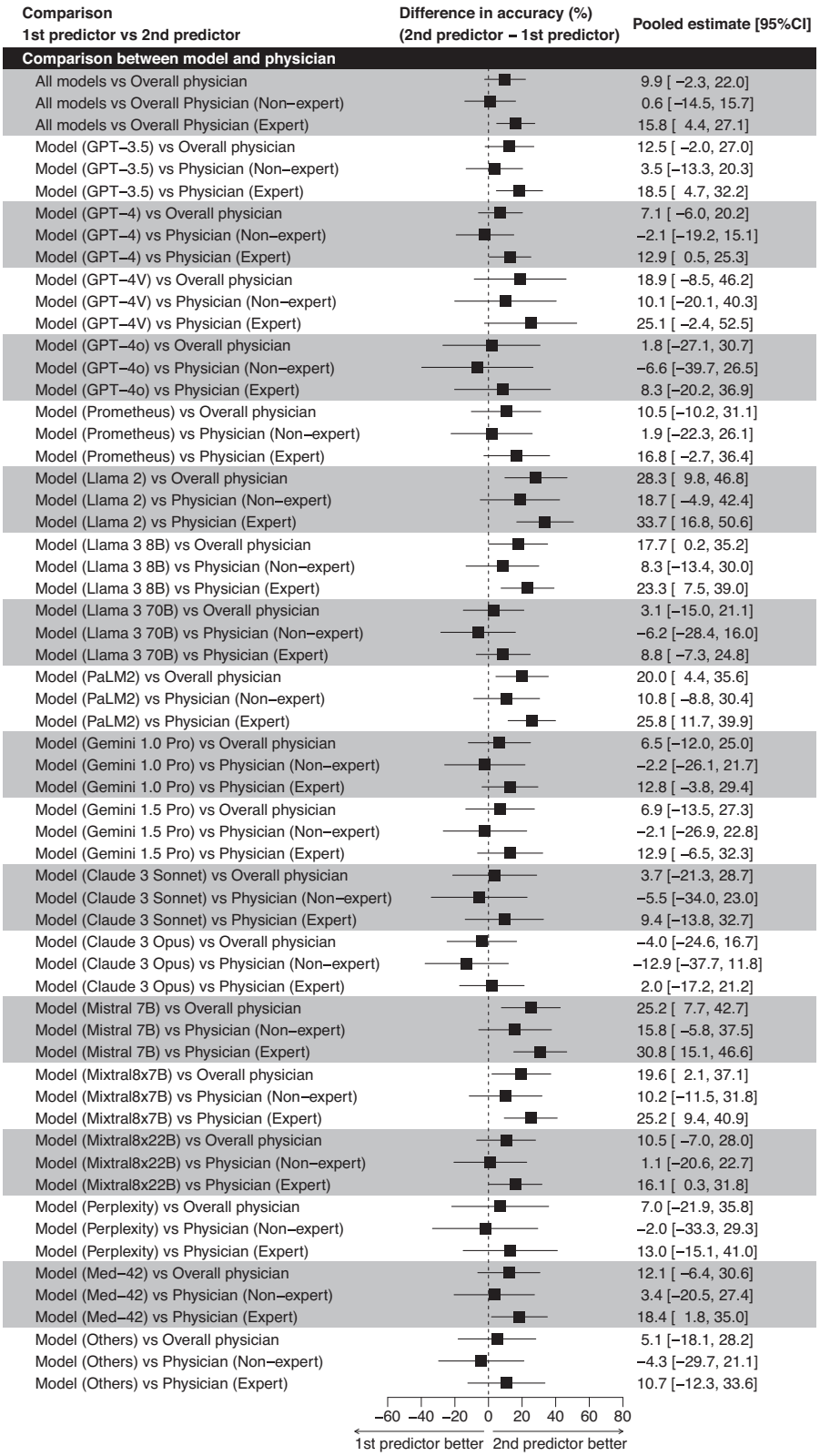

Fig. 3 | Comparison results between models and physicians. This figure demonstrates the differences in accuracy between various AI models and physicians. It specifically compares the performance of AI models against the overall accuracy of physicians, as well as against non-experts and experts separately. Each horizontal line represents the range of accuracy differences for the model compared to the physician category. The percentage values displayed on the right-hand side correspond to these mean differences, with the values in parentheses providing the 95% confidence intervals for these estimates. The dotted vertical line marks the 0% difference threshold, indicating where the model's accuracy is exactly the same as that of the physicians. Positive values (to the right of the dotted line) suggest that the physicians outperformed the model, whereas negative values (to the left) indicate that the model was more accurate than the physicians.

The methodology of this study, while comprehensive, has limitations. The generalizability of our findings warrants careful consideration. Our heterogeneity analysis revealed moderate levels of explained variability, suggesting that while our meta-regression model accounts for a substantial portion of the differences between studies, other factors not captured in our analysis may influence generative AI performance. Furthermore, the lack of demographic information in many included studies limits our ability to assess the generalizability of these findings across diverse populations and geographic regions. The performance of generative AI may vary considerably depending on the demographic characteristics and healthcare contexts represented in their training data. Although our meta-analysis shows no significant difference between the pooled accuracy of models and

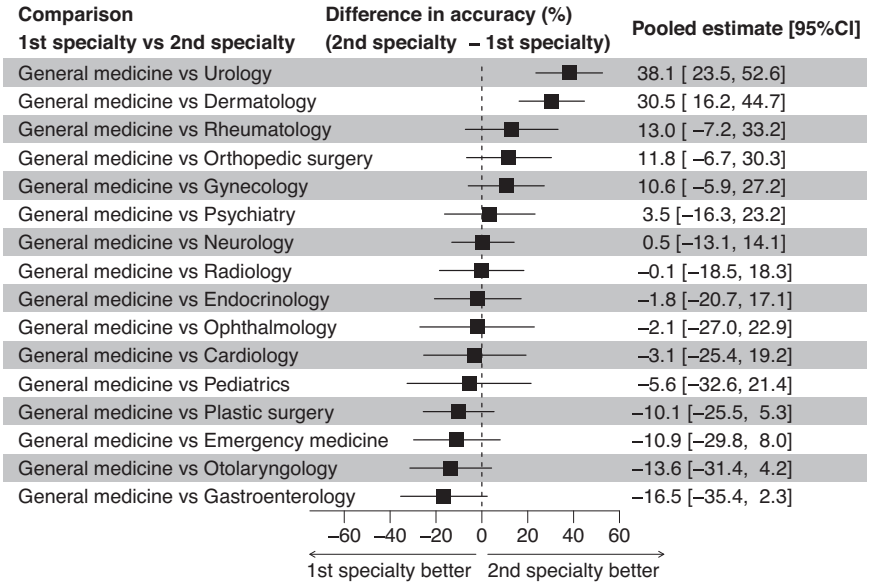

Fig. 4 | Generative AI performance among specialties. This figure demonstrates the differences in accuracy of generative AI models for specialties. Each horizontal line represents the range of accuracy differences between the specialty and General medicine. The percentage values displayed on the right-hand side correspond to these mean differences, with the values in parentheses providing the 95% confidence intervals for these estimates. The dotted vertical line marks the 0% difference threshold, indicating where the performance of generative AI models in the specialty is exactly the same as that of General medicine. Positive values (to the right of the dotted line) suggest that the model performance for the specialty was greater than that for General medicine, whereas negative values (to the left) indicate that the model performance for the specialty was less than that for General medicine.

that of physicians, recent research demonstrates that generative AI may perform significantly worse than physicians in more complex scenarios where models are provided with detailed information from electronic health records[80]. This suggests that generative AI model performance may degrade in more complex, real-world scenarios. Future research should prioritize the inclusion of diverse patient populations and cases that reflect more complex, real-world scenarios to better understand the generalizability of generative AI performance across different populations and clinical settings. Additionally, investigating the intersecting impact of physicians using generative AI models clinically, such as changes in performance, would be valuable.

In conclusion, this meta-analysis provides a nuanced understanding of the capabilities and limitations of generative AI in medical diagnostics. Generative AI models, particularly advanced iterations like GPT-4, GPT-4o, Llama3 70B, Gemini 1.0 Pro, Gemini 1.5 Pro, Claude 3 Sonnet, Claude 3 Opus, and Perplexity, have shown progressive improvements and hold promise for assisting in diagnosis, though their effectiveness varies by model. With an overall moderate accuracy of 52%, generative AI models are not yet reliable substitutes for expert physicians but may serve as valuable aids in non-expert scenarios and as educational tools for medical trainees. The findings also underscore the need for continued advancements and specialization in model development, as well as rigorous, externally validated research to overcome the prevalent high risk of bias and ensure generative AIs' effective integration into clinical practice. As the field evolves, continuous learning and adaptation for both generative AI models and medical professionals are imperative, alongside a commitment to transparency and stringent research standards. This approach will be crucial in harnessing the potential of generative AI models to enhance healthcare delivery and medical education while safeguarding against their limitations and biases.

Methods

Protocol and registration

This systematic review was prospectively registered with PROSPERO (CRD42023494733). Our study adhered to the relevant sections of guidelines from the Preferred Reporting Items for a Systematic Review and Meta-analysis (PRISMA) of Diagnostic Test Accuracy Studies (Supplementary Table 4)[119,120]. All stages of the review (title and abstract screening, full-text screening, data extraction, and assessment of bias) were performed in duplicate by two independent reviewers (H.Takita and D.U.), and disagreements were resolved by discussion with a third independent reviewer (H.Tatekawa).

Search strategy and study selection

A search was performed to identify studies that validate a generative AI model for diagnostic tasks. A search strategy was developed, including variations of the terms *generative AI* and *diagnosis*. The search strategy was as follows: articles in English that included the words "large language model," "LLM," "generative artificial intelligence," "generative AI," "generative pre-trained transformers,"[1] "GPT," "Bing," "Prometheus," "Bard," "PaLM,"[6,7] "Pathways Language Model," "LaMDA,"[8] "Language Model for Dialogue Applications," "Llama,"[4,5] or "Large Language Model Meta AI" and also "diagnosis," "diagnostic," "quiz," "examination," or "vignette" were included. We searched the following electronic databases for literature from June 2018 through June 2024: Medline, Scopus, Web of Science, Cochrane Central, and MedRxiv. June 2018 represents when the first generative AI model was published[1]. We included all articles that fulfilled the following inclusion criteria: primary research studies that validate a generative AI for diagnosis. We applied the following exclusion criteria to our search: review articles, case reports, comments, editorials, and retracted articles.

Data extraction

Titles and abstracts were screened before full-text screening. Data was extracted using a predefined data extraction sheet. A count of excluded studies, including the reason for exclusion, was recorded in a PRISMA flow diagram[120]. We extracted information from each study including the first author, model with its version, model task, test dataset type (internal, external, or unknown)[114], medical specialty, accuracy, sample size, and publication status (preprint or peer-reviewed) for the meta-analysis of generative AI performance. We defined three test types based on the relationship between the model's training and test data[121]. We defined internal testing as when the test data was derived from the same source or distribution as the training data but was properly separated from the training set through standard practices such as cross-validation or random splitting. We defined external testing as when the test data was collected after the training data cutoff, or the model was tested on private data. We defined unknown testing as when the test data was collected before the training data cutoff, and the data was publicly available. This is because companies developing these models have not disclosed their complete training datasets. In addition to this, when both the model and the physician's diagnostic performance were presented in the same paper, we extracted both for meta-analysis. We also considered the type of physician involved in relevant studies. We classified physicians as non-experts if they were trainees or residents. In contrast, those beyond this stage in their career were

categorized as experts. When a single model used multiple prompts and individual performances were available in one article, we took the average of them.

Quality assessment

We used PROBAST to assess papers for bias and applicability[114]. This tool uses signaling questions in four domains (participants, predictors, outcomes, and analysis) to provide both an overall and a granular assessment. We did not include some PROBAST signaling questions because they are not relevant to generative AI models. Details of modifications made to PROBAST are in Supplementary Table 3 (online).

Statistical analysis

We calculated the pooled accuracy of diagnosis brought by generative AI models and physicians based on the previously reported studies. The pooled diagnosis accuracies were compared between all AI models and physicians overall using the multivariable random-effect meta-regression model with adjustment for medical specialty, task of models, type of test dataset, level of bias, and publication status. We compared AI models overall with physicians overall, expert physicians, and non-expert physicians. Additionally, we compared each AI model with physicians overall and each AI model with each physician's experience level (expert or non-expert). Furthermore, we assessed the variation of generative AI model accuracy across specialties. For fitting the meta-regression models, a restricted maximum likelihood estimator was utilized with the "metafor" package in R. To explore variation between knowledge domains, we performed a subgroup analysis comparing medical-domain models with non-medical-domain models. To assess the impact of the overall risk of bias, we conducted a subgroup analysis limited to studies with a low overall risk of bias. To assess the impact of publication bias on the comparison of the diagnosis performance between the AI models and the physicians, we used a funnel plot and Egger's regression test. Additionally, to assess the impact of heterogeneity, we conducted heterogeneity analyses in both the full dataset and the subgroup that had a low overall risk of bias. All statistical analyses were conducted using R version 4.4.0.

Data availability

The corresponding author had full access to all data in the study and final responsibility for the decision to submit the report for publication. The data used and analyzed during the current study are available from the corresponding author upon reasonable request.

Code availability

Study protocol and metadata are available from Dr. Ueda (e-mail, ai.labo.ocu@gmail.com).

Received: 26 July 2024; Accepted: 26 February 2025;
Published online: 22 March 2025

References

1. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. *Improving Language Understanding by Generative Pre-training*. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (2018).
2. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
3. OpenAI et al. GPT-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
4. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.13971 (2023).
5. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2307.09288 (2023).
6. Chowdhery, A. et al. PaLM: scaling language modeling with pathways. *J. Mach. Learn. Res.* **24**, 1–113 (2023).
7. Anil, R. et al. PaLM 2 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.10403 (2023).
8. Thoppilan, R. et al. LaMDA: language models for dialog applications. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2201.08239 (2022).
9. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
10. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
11. Ueda, D. et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. *Radiology* **308**, e231040 (2023).
12. Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* **330**, 78–80 (2023).
13. Hirosawa, T., Mizuta, K., Harada, Y. & Shimizu, T. Comparative evaluation of diagnostic accuracy between Google Bard and physicians. *Am. J. Med.* **136**, 1119–1123.e18 (2023).
14. Shea, Y.-F., Lee, C. M. Y., Ip, W. C. T., Luk, D. W. A. & Wong, S. S. W. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. *JAMA Netw. Open* **6**, e2325000 (2023).
15. Chee, J., Kwa, E. D. & Goh, X. Vertigo, likely peripheral': the dizzying rise of ChatGPT. *Eur. Arch. Otorhinolaryngol.* **280**, 4687–4689 (2023).
16. Lyons, R. J., Arepalli, S. R., Fromal, O., Choi, J. D. & Jain, N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Can. J. Ophthalmol*. https://doi.org/10.1016/j.jcjo.2023.07.016 (2023).
17. Benoit, J. R. A. ChatGPT for clinical vignette generation, revision, and evaluation. Preprint at *medRxiv* https://doi.org/10.1101/2023.02.04.23285478 (2023).
18. Hirosawa, T. et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. *JMIR Med. Inf.* **11**, e48808 (2023).
19. Hirosawa, T. et al. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int. J. Environ. Res. Public Health* **20**, 3378 (2023).
20. Wei, Q., Cui, Y., Wei, B., Cheng, Q. & Xu, X. Evaluating the performance of ChatGPT in differential diagnosis of neurodevelopmental disorders: a pediatricians-machine comparison. *Psychiatry Res.* **327**, 115351 (2023).
21. Allahqoli, L., Ghiasvand, M. M., Mazidimoradi, A., Salehiniya, H. & Alkatout, I. Diagnostic and management performance of ChatGPT in obstetrics and gynecology. *Gynecol. Obstet. Invest.* **88**, 310–313 (2023).
22. Levartovsky, A., Ben-Horin, S., Kopylov, U., Klang, E. & Barash, Y. Towards AI-augmented clinical decision-making: an examination of ChatGPT's utility in acute ulcerative colitis presentations. *Am. J. Gastroenterol.* **118**, 2283–2289 (2023).
23. Bushuven, S. et al. 'ChatGPT, can you help me save my child's life?'—diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital basic life support and paediatric advanced life support cases—an in-silico analysis. *J. Med. Syst.* **47**, 123 (2023).
24. Knebel, D. et al. Assessment of ChatGPT in the prehospital management of ophthalmological emergencies—an analysis of 10 fictional case vignettes. *Klin. Monbl. Augenheilkd*. https://doi.org/10.1055/a-2149-0447 (2023).
25. Pillai, J. & Pillai, K. Accuracy of generative artificial intelligence models in differential diagnoses of familial Mediterranean fever and deficiency of Interleukin-1 receptor antagonist. *J. Transl. Autoimmun.* **7**, 100213 (2023).

26. Ito, N. et al. The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: evaluation study. *JMIR Med. Educ.* **9**, e47532 (2023).
27. Sorin, V. et al. GPT-4 multimodal analysis on ophthalmology clinical cases including text and images. Preprint at *medRxiv* https://doi.org/10.1101/2023.11.24.23298953 (2023).
28. Madadi, Y. et al. ChatGPT assisting diagnosis of neuro-ophthalmology diseases based on case reports. Preprint at *medRxiv* https://doi.org/10.1101/2023.09.13.23295508 (2023).
29. Schubert, M. C., Lasotta, M., Sahm, F., Wick, W. & Venkataramani, V. Evaluating the multimodal capabilities of generative AI in complex clinical diagnostics. Preprint at *medRxiv* https://doi.org/10.1101/2023.11.01.23297938 (2023).
30. Kiyohara, Y. et al. Large language models to differentiate vasospastic angina using patient information. Preprint at *medRxiv* https://doi.org/10.1101/2023.06.26.23291913 (2023).
31. Sultan, I. et al. Using ChatGPT to predict cancer predisposition genes: a promising tool for pediatric oncologists. *Cureus* **15**, e47594 (2023).
32. Horiuchi, D. et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology* **66**, 73–79 (2023).
33. Stoneham, S., Livesey, A., Cooper, H. & Mitchell, C. Chat GPT vs clinician: challenging the diagnostic capabilities of A.I. In dermatology. *Clin. Exp. Dermatol*. https://doi.org/10.1093/ced/llad402 (2023).
34. Rundle, C. W., Szeto, M. D., Presley, C. L., Shahwan, K. T. & Carr, D. R. Analysis of ChatGPT generated differential diagnoses in response to physical exam findings for benign and malignant cutaneous neoplasms. *J. Am. Acad. Dermatol*. https://doi.org/10.1016/j.jaad.2023.10.040 (2023).
35. Rojas-Carabali, W. et al. Chatbots Vs. human experts: evaluating diagnostic performance of chatbots in uveitis and the perspectives on AI adoption in ophthalmology. *Ocul. Immunol. Inflamm*. **32**, 1591–1598 (2024).
36. Fraser, H. et al. Comparison of diagnostic and triage accuracy of Ada Health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. *JMIR Mhealth Uhealth* **11**, e49995 (2023).
37. Krusche, M., Callhoff, J., Knitza, J. & Ruffer, N. Diagnostic accuracy of a large language model in rheumatology: comparison of physician and ChatGPT-4. *Rheumatol. Int*. https://doi.org/10.1007/s00296-023-05464-6 (2023).
38. Galetta, K. & Meltzer, E. Does GPT-4 have neurophobia? Localization and diagnostic accuracy of an artificial intelligence-powered chatbot in clinical vignettes. *J. Neurol. Sci.* **453**, 120804 (2023).
39. Delsoz, M. et al. The use of ChatGPT to assist in diagnosing glaucoma based on clinical case reports. *Ophthalmol. Ther.* **12**, 3121–3132 (2023).
40. Hu, X. et al. What can GPT-4 do for diagnosing rare eye diseases? A pilot study. *Ophthalmol. Ther.* **12**, 3395–3402 (2023).
41. Abi-Rafeh, J., Hanna, S., Bassiri-Tehrani, B., Kazan, R. & Nahai, F. Complications following facelift and neck lift: implementation and assessment of large language model and artificial intelligence (ChatGPT) performance across 16 simulated patient presentations. *Aesthetic Plast. Surg.* **47**, 2407–2414 (2023).
42. Koga, S., Martin, N. B. & Dickson, D. W. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathol*. **34**, e13207 (2023).
43. Xv, Y., Peng, C., Wei, Z., Liao, F. & Xiao, M. Can Chat-GPT a substitute for urological resident physician in diagnosing diseases?: a preliminary conclusion from an exploratory investigation. *World J. Urol.* **41**, 2569–2571 (2023).
44. Senthujan, S. M. et al. GPT-4V(ision) unsuitable for clinical care and education: a clinician-evaluated assessment. Preprint at *medRxiv* https://doi.org/10.1101/2023.11.15.23298575 (2023).
45. Mori, Y., Izumiyama, T., Kanabuchi, R., Mori, N. & Aizawa, T. Large language model may assist diagnosis of SAPHO syndrome by bone scintigraphy. *Mod. Rheumatol*. **34**, 1043–1046 (2024).
46. Mykhalko, Y., Kish, P., Rubtsova, Y., Kutsyn, O. & Koval, V. From text to diagnose: ChatGPT'S efficacy in medical decision-making. *Wiad. Lek.* **76**, 2345–2350 (2023).
47. Andrade-Castellanos, C. A., Paz, M. T. T.I.a. & Farfán-Flores, P. E. Accuracy of ChatGPT for the diagnosis of clinical entities in the field of internal medicine. *Gac. Med. Mex.* **159**, 439–442 (2023).
48. Daher, M. et al. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? *JSES Int.* **7**, 2534–2541 (2023).
49. Suthar, P. P., Kounsal, A., Chhetri, L., Saini, D. & Dua, S. G. Artificial intelligence (AI) in radiology: a deep dive into ChatGPT 4.0's accuracy with the American Journal of Neuroradiology's (AJNR) 'Case of the Month'. *Cureus* **15**, e43958 (2023).
50. Nakaura, T. et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Jpn. J. Radiol.* **15**, 1–11 (2023).
51. Berg, H. T. et al. ChatGPT and generating a differential diagnosis early in an emergency department presentation. *Ann. Emerg. Med.* **83**, 83–86 (2024).
52. Gebrael, G. et al. Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. *Cancers* **15**, 3717 (2023).
53. Ravipati, A., Pradeep, T. & Elman, S. A. The role of artificial intelligence in dermatology: the promising but limited accuracy of ChatGPT in diagnosing clinical scenarios. *Int. J. Dermatol.* **62**, e547–e548 (2023).
54. Shikino, K. et al. Evaluation of ChatGPT-generated differential diagnosis for common diseases with atypical presentation: descriptive research. *JMIR Med. Educ.* **10**, e58758 (2024).
55. Horiuchi, D. et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. *Clin. Neuroradiol*. **34**, 779–787 (2024).
56. Kumar, R. P. et al. Can artificial intelligence mitigate missed diagnoses by generating differential diagnoses for neurosurgeons? *World Neurosurg.* **187**, e1083–e1088 (2024).
57. Chiu, W. H. K. et al. Evaluating the diagnostic performance of large language models on complex multimodal medical cases. *J. Med. Internet Res.* **26**, e53724 (2024).
58. Kikuchi, T. et al. Toward improved radiologic diagnostics: investigating the utility and limitations of GPT-3.5 Turbo and GPT-4 with quiz cases. *AJNR Am. J. Neuroradiol.* **45**, 1506–1511 (2024).
59. Bridges, J. M. Computerized diagnostic decision support systems—a comparative performance study of Isabel Pro vs. ChatGPT4. *Acta Radiol. Diagn.* **11**, 250–258 (2024).
60. Shieh, A. et al. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. *Sci. Rep.* **14**, 1–8 (2024).
61. Warrier, A., Singh, R., Haleem, A., Zaki, H. & Eloy, J. A. The comparative diagnostic capability of large language models in otolaryngology. *Laryngoscope* **134**, 3997–4002 (2024).
62. Han, T. et al. Comparative analysis of multimodal large language model performance on clinical vignette questions. *JAMA* **331**, 1320–1321 (2024).
63. Milad, D. et al. Assessing the medical reasoning skills of GPT-4 in complex ophthalmology cases. *Br. J. Ophthalmol.* **108**, 1398–1405 (2024).

64. Abdullahi, T., Singh, R. & Eickhoff, C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. *JMIR Med. Educ.* **10**, e51391 (2024).
65. Tenner, Z. M., Cottone, M. & Chavez, M. Harnessing the open access version of ChatGPT for enhanced clinical opinions. Preprint at *medRxiv* https://doi.org/10.1101/2023.08.23.23294478 (2023).
66. Luk, D. W. A., Ip, W. C. T. & Shea, Y.-F. Performance of GPT-4 and GPT-3.5 in generating accurate and comprehensive diagnoses across medical subspecialties. *J. Chin. Med. Assoc.* **87**, 259 (2024).
67. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit. Med.* **7**, 1–7 (2024).
68. Franc, J. M., Cheng, L., Hart, A., Hata, R. & Hertelendy, A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. *CJEM* **26**, 40–46 (2024).
69. Yang, J. et al. RDmaster: a novel phenotype-oriented dialogue system supporting differential diagnosis of rare disease. *Comput. Biol. Med.* **169**, 107924 (2024).
70. Reese, J. T. et al. On the limitations of large language models in clinical diagnosis. Preprint at *medRxiv* https://doi.org/10.1101/2023.07.13.23292613 (2023).
71. do Olmo, J., Logroño, J., Mascías, C., Martínez, M. & Isla, J. Assessing DxGPT: diagnosing rare diseases with various large language models. Preprint at *medRxiv* https://doi.org/10.1101/2024.05.08.24307062 (2024).
72. Cesur, T., Gunes, Y. C., Camur, E. & Dağlı, M. Empowering radiologists with ChatGPT-4o: comparative evaluation of large language models and radiologists in cardiac cases. Preprint at *medRxiv* https://doi.org/10.1101/2024.06.25.24309247 (2024).
73. Schramm, S. et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4(V) in challenging brain MRI cases. Preprint at *medRxiv* https://doi.org/10.1101/2024.03.05.24303767 (2024).
74. Gunes, Y. C. & Cesur, T. A comparative study: diagnostic performance of ChatGPT 3.5, Google Bard, Microsoft Bing, and radiologists in thoracic radiology cases. Preprint at *medRxiv* https://doi.org/10.1101/2024.01.18.24301495 (2024).
75. Olshaker, H. et al. Evaluating the diagnostic performance of large language models in identifying complex multisystemic syndromes: a comparative study with radiology residents. Preprint at *medRxiv* https://doi.org/10.1101/2024.06.05.24308335 (2024).
76. Hirosawa, T. et al. Diagnostic performance of generative artificial intelligences for a series of complex case reports. *Digit. Health* **10**, 20552076241265215 (2024).
77. Mitsuyama, Y. et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. *Eur. Radiol*. https://doi.org/10.1007/s00330-024-11032-8 (2024).
78. Yazaki, M. et al. Emergency patient triage improvement through a retrieval-augmented generation enhanced large-scale language model. *Prehosp. Emerg. Care* https://doi.org/10.1080/10903127.2024.2374400 (2024).
79. Ghalibafan, S. et al. Applications of multimodal generative artificial intelligence in a real-world retina clinic setting. *Retina* **44**, 1732–1740 (2024).
80. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
81. Horiuchi, D. et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur. Radiol.* https://doi.org/10.1007/s00330-024-10902-5 (2024).
82. Ríos-Hoyo, A. et al. Evaluation of large language models as a diagnostic aid for complex medical cases. *Front. Med.* **11**, 1380148 (2024).
83. Liu, X. et al. Claude 3 Opus and ChatGPT With GPT-4 in dermoscopic image analysis for melanoma diagnosis: comparative performance analysis. *JMIR Med. Inform.* **12**, e59273 (2024).
84. Sonoda, Y. et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in 'Diagnosis Please' cases. *Jpn. J. Radiol*. https://doi.org/10.1007/s11604-024-01619-y (2024).
85. Wada, A. et al. Optimizing GPT-4 Turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds. *Diagnostics* **14**, 1541 (2024).
86. Gargari, O. K. et al. Diagnostic accuracy of large language models in psychiatry. *Asian J. Psychiatr.* **100**, 104168 (2024).
87. Mihalache, A. et al. Interpretation of clinical retinal images using an artificial intelligence Chatbot. *Ophthalmol. Sci.* **4**, 100556 (2024).
88. Rutledge, G. W. Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. *Learn Health Syst.* **8**, e10438 (2024).
89. Ueda, D. et al. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. *BMC Digit. Health* **2**, 4 (2024).
90. Delsoz, M. et al. Performance of ChatGPT in diagnosis of corneal eye diseases. Preprint at *medRxiv* https://doi.org/10.1101/2023.08.25.23294635 (2023).
91. Brin, D. et al. Assessing GPT-4 multimodal performance in radiological image analysis. Preprint at *medRxiv* https://doi.org/10.1101/2023.11.15.23298583 (2023).
92. Levine, D. M. et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: an observational study. *Lancet Digit. Health* **6**, e555–e561 (2024).
93. Williams, C. Y. K. et al. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA Netw. Open* **7**, e248895 (2024).
94. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf (2023).
95. Model Card Claude 3. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf (2024).
96. Gemini Team et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2403.05530 (2024).
97. GPT-4o(mni) System card. https://cdn.openai.com/gpt-4o-system-card.pdf (2024).
98. Dubey, A. et al. The Llama 3 herd of models. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2407.21783 (2024).
99. Perplexity Model Card https://docs.perplexity.ai/guides/model-cards. Perplexity (2024).
100. Üstün, A. et al. Aya model: an instruction finetuned open-access multilingual language model. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2402.07827 (2024).
101. Model Card Claude 2. https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf (2023).
102. Model Card Claude 3.5. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf (2024).
103. Toma, A. et al. Clinical Camel: an open expert-level medical language model with dialogue-based knowledge encoding. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2305.12031 (2023).
104. Gemini Team et al. Gemini: a family of highly capable multimodal models. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2312.11805 (2023).
105. Glass version 2.0. *GLASS* https://glass.health/ai (2024).
106. Han, T. et al. Comparative analysis of GPT-4Vision, GPT-4 and Open Source LLMs in clinical diagnostic accuracy: a benchmark against human expertise. Preprint at *medRxiv* https://doi.org/10.1101/2023.11.03.23297957 (2023).

107. Han, T. et al. MedAlpaca—an open-source collection of medical conversational AI models and training data. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2304.08247 (2023).
108. Chen, Z. et al. MEDITRON-70B: scaling medical pretraining for large language models. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2311.16079 (2023).
109. Jiang, A. Q. et al. Mistral 7B. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2310.06825 (2023).
110. Jiang, A. Q. et al. Mixtral of experts. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2401.04088 (2024).
111. Zhang, V. NVIDIA AI foundation models: build custom enterprise Chatbots and co-pilots with production-ready LLMs. *NVIDIA Technical Blog* https://developer.nvidia.com/blog/nvidia-ai-foundation-models-build-custom-enterprise-chatbots-and-co-pilots-with-production-ready-llms/ (2023).
112. Köpf, A. et al. OpenAssistant Conversations—democratizing large language model alignment. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2304.07327 (2023).
113. Xu, C. et al. WizardLM: empowering large language models to follow complex instructions. Preprint at *arXiv* https://doi.org/10.48550/ARXIV.2304.12244 (2023).
114. Wolff, R. F. et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **170**, 51–58 (2019).
115. Wahl, B., Cossy-Gantner, A., Germann, S. & Schwalbe, N. R. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob. Health* **3**, e000798 (2018).
116. Preiksaitis, C. & Rose, C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR Med. Educ.* **9**, e48785 (2023).
117. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
118. Ueda, D. et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn. J. Radiol.* **42**, 3–15 (2023).
119. McInnes, M. D. F. et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. *JAMA* **319**, 388–396 (2018).
120. Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* **339**, b2535 (2009).
121. Walston, S. L. et al. Data set terminology of deep learning in medicine: a historical review and recommendation. *Jpn. J. Radiol.* **42**, 1100–1109 (2024).

Acknowledgements

There was no funding provided for this study. We utilized ChatGPT for assistance with parts of the English proofing.

Author contributions

H.Tak., K.S., and D.U. contributed to data acquisition. H.Tak., H.Tat., and D.U. contributed to the data quality check. D.K., Y.T. and D.U. analyzed the data. H.Tak. and D.U. drafted the manuscript. S.L.W. and H.Tat. edited and revised the manuscript. Y.M. supervised the study. All authors had access to all the data reported in the study and accept responsibility for the decision to submit for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01543-z.

Correspondence and requests for materials should be addressed to Daiju Ueda.