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

Running title: Diagnostic performance review of generative AI and physicians

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

1) Department of Diagnostic and Interventional Radiology, Graduate School of Medicine, Osaka Metropolitan University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan
2) Department of Medical Statistics, Graduate School of Medicine, Osaka Metropolitan University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan
3) Center for Digital Transformation of Health Care, Graduate School of Medicine, Kyoto University, Shogoin-Kawahara-cho 53, Sakyo-ku, Kyoto 606-8507, Japan
4) Oku medical clinic, Shimmori 7-1-4, Asahi-ku, Osaka 535-0022, Japan
5) Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
6) Scientific Research WorkS Peer Support Group (SRWS-PSG), Koraibashi 1-7-7-2302, Chuo-ku, Osaka 541-0043, Japan
7) Center for Health Science Innovation, Osaka Metropolitan University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan
8) Department of Artificial Intelligence, Graduate School of Medicine, Osaka Metropolitan University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan

Corresponding author: Daiju Ueda, MD, PhD

Department of Artificial Intelligence, Graduate School of Medicine, Osaka Metropolitan University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan

Phone: +81-6-6645-3831; Fax: +81-6-6646-6655

E-mail: ai.labo.ocu@gmail.com

Manuscript word count: 2998

Key Points

Question: What is the diagnostic accuracy of generative AI models and how does this accuracy compare to that of physicians?

Findings: This meta-analysis found that generative AI models have a pooled accuracy of 56.9% (95% confidence interval: 51.0–62.7%). The accuracy of expert physicians exceeds that of AI in all specialties, however, some generative AI models are comparable to non-expert physicians.

Meaning: The diagnostic performance of generative AI models suggests that they do not match the level of experienced physicians but that they may have potential applications in healthcare delivery and medical education.

Abstract

Background: The rapid advancement of generative artificial intelligence (AI) has led to the wide dissemination of models with exceptional understanding and generation of human language. Their integration into healthcare has shown potential for improving medical diagnostics, yet a comprehensive diagnostic performance evaluation of generative AI models and the comparison of their diagnostic performance with that of physicians has not been extensively explored.

Methods: In this systematic review and meta-analysis, a comprehensive search of Medline, Scopus, Web of Science, Cochrane Central, and MedRxiv was conducted for studies published from June 2018 through December 2023, focusing on those that validate generative AI models for diagnostic tasks. The risk of bias was assessed using the Prediction Model Study Risk of Bias Assessment Tool. Meta-regression was performed to summarize the performance of the models and to compare the accuracy of the models with that of physicians.

Results: The search resulted in 54 studies being included in the meta-analysis. Nine generative AI models were evaluated across 17 medical specialties. The quality assessment indicated a high risk of bias in the majority of studies, primarily due to small sample sizes. The overall accuracy for generative AI models across 54 studies was 56.9% (95% confidence interval [CI]: 51.0–62.7%). The meta-analysis demonstrated that, on average, physicians exceeded the accuracy of the models (difference in accuracy: 14.4% [95% CI: 4.9–23.8%], p-value =0.004). However, both Prometheus (Bing) and GPT-4 showed slightly better performance compared to non-experts (-2.3% [95% CI: -27.0–22.4%], p-value = 0.848 and -0.32% [95% CI: -14.4–13.7%], p-value = 0.962), but slightly underperformed when compared to experts (10.9% [95% CI: -13.1–35.0%], p-value = 0.356 and 12.9% [95% CI: 0.15–25.7%], p-value = 0.048). The sub-analysis revealed significantly improved accuracy in the fields of Gynecology, Pediatrics, Orthopedic surgery, Plastic surgery, and Otolaryngology, while showing reduced accuracy for Neurology, Psychiatry, Rheumatology, and Endocrinology compared to that of General Medicine. No significant heterogeneity was observed based on the risk of bias.

Conclusions: Generative AI exhibits promising diagnostic capabilities, with accuracy varying significantly by model and medical specialty. Although they have not reached the reliability of expert physicians, the findings suggest that generative AI models have the potential to enhance healthcare delivery and medical education, provided they are integrated with caution and their limitations are well-understood.

Introduction

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 their deep learning architectures, have 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 is 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 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.

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.[15,16] 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 December 2023: 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.[16] We extracted information from each study including the first author, model with its version, model task, test dataset type (internal, external, or unknown),[17] medical specialty, accuracy, sample size, and publication status (pre-print or peer-reviewed) for the meta-analysis of generative AI performance. Most generative AI models only presented their training period without any information on which data were used for training. Therefore, when generative AI models were tested with data outside of the training period, the test dataset type was classified as

external testing, and when tested with data that were publicly available during the training period, it was classified as unknown. 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 the Prediction Model Study Risk of Bias Assessment Tool (PROBAST) to assess papers for bias and applicability.[17] 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 Appendix Table S1 (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 overall physicians using the multivariable random-effect meta-regression model with adjustment for medical speciality, task of models, type of test dataset, level of bias, and publication status. In addition to the comparison of all AI models and overall physicians, we compared each AI model with overall physicians and each AI model with each physician experience level (expert or non-expert). Furthermore, we assessed the variation of generative AI model accuracy across specialities. For fitting the meta-regression models, a restricted maximum likelihood estimator was utilized with the “metafor” package in R. 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. All statistical analyses were conducted using R version 4.3.0.

Results

Study Selection and Characteristics

We identified 13,966 studies, of which 7,940 were duplicates. After screening, 54 studies were included in the meta-analysis[11–14,19–68] (Figure 1 and Table 1). The most evaluated models were GPT-4[3] (31 articles) and GPT-3.5[2] (28), while models such as GPT-4V[69] (6), PaLM2[7] (3), Llama 2[5] (2), Prometheus (2), GPT-3[2] (1), Glass AI[70] (1), and Med-42[56] (1) had less representation. GPT-3, GPT-3.5, GPT-4, and GPT-4V are available in ChatGPT or its application programming interface (Open AI, San Francisco, CA). PaLM2 is implemented in Bard (Google, Menlo Park, CA). Bing (Microsoft, Redmond, WA) incorporates Prometheus, which is based on OpenAI's GPT technology. Med-42 is a fine-tuned version of the open-source large language model, Llama 2 (Meta, Menlo Park, CA). Lastly, Glass AI is implemented in Glass (Glass Health, San Francisco, CA). The review spanned a wide range of medical specialties, with General medicine being the most common (14 articles). Other specialties like Radiology (10), Ophthalmology (8), Emergency medicine (5), Neurology (3), and Dermatology (3) were represented, as well as Gastroenterology, Cardiology, Pediatrics, Otolaryngology, Urology, Endocrinology, Gynecology, Orthopedic surgery, Rheumatology, Psychiatry, and Plastic surgery with one article each. Regarding model tasks, free text tasks were the most common, with 47 articles, followed by choice tasks at 13. For test dataset types, 40 articles involved external testing, while 14 were unknown because the training data for the generative AI models was unknown. Of the included studies, 37 were peer-reviewed, while 17 were preprints. Study characteristics are shown in Table 1 and Appendix Table S2 (online). Thirteen studies compared the performance of generative AI models with that of physicians.[30,31,33–39,47,50,54,58] GPT-4 (8 articles) was the most frequently compared with physicians, followed by GPT-3.5 (7), GPT-4V (2), Llama 2 (1), and GPT-3 (1). While comparisons between both expert and non-expert physicians were found for GPT-4, GPT-3.5, GPT-4V, and GPT-3, only comparisons with experts were found for Llama 2, with no comparisons involving non-experts.

Quality Assessment

PROBAST assessment led to an overall rating of 45/54 (83%) studies at high risk of bias, 8/54 (15%) studies at low risk of bias, 10/54 (19%) studies at high concern for generalizability, and 44/54 (81%) studies at low concern for generalizability (Figure 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 Appendix Table S2 (online).

Meta-analysis

The overall accuracy for generative AI models was found to be 56.9% with a 95% CI of 51.0–62.7%. In the meta-regression, we observed that physicians generally outperformed generative AI models in various scenarios (Figure 3). This superiority was evident when comparing AI models to overall physician performance, where physicians demonstrated a significant 14.4% higher performance on average (95% CI: 4.9–23.8%, p =0.004). Interestingly, when comparing the performance of the Prometheus and GPT-4 models against non-experts, both models demonstrated a slight but not statistically significant superiority, with differences of -2% (95% CI: -27.0 to 22.4%, p = 0.848) and -0.3% (95% CI: -14.4 to 13.7%, p = 0.962), respectively. However, both models underperformed in comparison to experts, showing a 10.9% difference [95% CI: -13.1 to 35.0%, p-value = 0.356 for Prometheus] and 12.9% [95% CI: 0.15 to 25.6%, p-value = 0.048 for GPT-4]. The performance of all models but Prometheus and GPT-4 was inferior to both experts and non-experts in all comparisons. GPT-3, GPT-3.5, and PaLM2 were significantly inferior only when compared to expert physicians, whereas Llama 2, Glass, and Med-42 demonstrated substantial inferiority against both expert and non-expert physicians (p-values < 0.05).

In our meta-regression, we also found a remarkable difference in accuracy, with significant improvements in several specialties compared with General medicine. Specifically, AI performance in Gynecology, Pediatrics, Orthopedic surgery, Plastic surgery, and Otolaryngology outpace General medicine significantly, exhibiting differences of 34.4% (95% CI: 16.7–52.0%, p < 0.001), 34.3% (95% CI: 17.3–51.4%, p < 0.001), 34.1% (95% CI: 17.0–51.1%, p < 0.001), 28.5% (95% CI: 11.4–45.5%, p = 0.002), and 26.7% (95% CI: 9.6–43.7%, p = 0.004) respectively. Conversely, General medicine outperformed some areas such as Neurology, Psychiatry, Rheumatology, and Endocrinology. These areas witnessed a decline in accuracy with differences of -21.7% (95% CI: -41.0 to -2.3%, p = 0.030) in Neurology, -25.1% (95% CI: -44.4 to -5.9%, p = 0.012) in Psychiatry, -41.4% (95% CI: -73.2 to -9.6%, p = 0.013) in Rheumatology, and the most notable decrease in Endocrinology with -42.0% (95% CI: -60.5 to -23.4%, p < 0.001). These findings suggest that generative AI's performance is not uniform across all medical specialties, highlighting the necessity for specialty-specific optimization to harness its full potential effectively. No significant difference was observed based on the risk of bias (p = 0.77) or based on publication status (p = 0.58). We assessed publication bias by using a regression analysis to quantify funnel plot asymmetry (Appendix Figures S1 [online]), and it suggested a risk of publication bias (p = 0.027).

Discussion

In this systematic review and meta-analysis, we analyzed the diagnostic performance of generative AI and physicians. We initially identified 13,966 studies, ultimately including 54 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 57% (95% CI: 51–63%) for generative AI models. Physicians generally outperformed AI models, although in non-expert settings, some AI models showed comparable performance. Our analysis also highlighted significant differences in effectiveness across 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 57% for generative AI models in medical applications, this suggests its potential utility in certain clinical scenarios. The variation in effectiveness across specialties, particularly the lower effectiveness in some fields underscores the need for cautious implementation and further refinement of AI models in these areas. The data indicates that generative AI models possess a propensity towards knowledge in some medical specialties, and by understanding and utilizing their characteristics, they have the potential to function as a valuable support tool in medical settings. Importantly, the similar performance of Prometheus and GPT-4 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.[71]

The studies comparing generative AI and physician performance, particularly in the context of medical education, offer intriguing perspectives.[72] The overall higher accuracy of physicians compared to AI models emphasizes the irreplaceable value of human judgement and experience in medical decision-making. However, the comparable performance of Prometheus and GPT-4 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.[73] 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.

Although there are no statistically significant differences in diagnostic performance among the risks of bias, the PROBAST quality assessment reveals a high risk of bias in 83% of studies.[17] This raises significant concerns about the reliability of current generative AI research in healthcare. This highlights the crucial need for rigorous and transparent methodologies, including the necessity of large amounts of external evaluation to assess real-world performance accurately.[74] Moreover, 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. 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 their safe, effective, and equitable application in clinical settings,[75] ultimately enhancing the quality of healthcare delivery and medical education.

The methodology of this study, while comprehensive, has limitations. The performance of generative AI models might vary significantly in real-world scenarios, which are often more complex than research settings. There were not many studies that compared generative AI and physicians using the same sample. Future research should focus on addressing limitations by conducting studies with more diverse datasets, exploring the performance of generative AI models in varied clinical environments, and examining their impact on different patient demographics. 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. While generative AI models, particularly advanced iterations like Prometheus and GPT-4, have shown progressive improvements and hold promise for assisting in diagnosis, their effectiveness remains highly variable across different models and medical specialties. With an overall moderate accuracy of 57%, 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.

Funding

There was no funding provided for this study.

Role of the Sponsor

There was no funding provided for this study. The corresponding author had full access to all data in the study and final responsibility for the decision to submit the report for publication.

Acknowledgement

We utilized ChatGPT for assistance with parts of the English proofing.

IRB Approval

Not applicable.

Disclosures:

The authors have nothing to disclose.

Reproducible Research Statement:

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

Tables

Table 1: Study characteristics

Citation	First author	Model	Model task	Test type	Specialty	Comparison group	Cases	Publication status	Overall risk of bias	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	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
[19]	Chee	GPT-3.5	Free text	External	Otolaryngology	NA	7	Peer-reviewed	High	Low
[20]	Lyons	Prometheus, GPT-4	Free text, Choice	External	Ophthalmology	NA	44	Peer-reviewed	High	Low
[21]	Hirosawa	GPT-3.5, GPT-4	Free text	Unknown	General medicine	NA	52	Peer-reviewed	High	Low
[22]	Benoit	GPT-3.5	Free text, Choice	Unknown	General medicine	NA	45	Preprint	High	Low
[23]	Hirosawa	GPT-3.5	Free text	External	General medicine	NA	30	Peer-reviewed	High	Low
[24]	Wei	GPT-4	Choice	External	Psychiatry	NA	60	Peer-reviewed	High	Low
[25]	Ueda	GPT-4	Free text	External	General medicine	NA	62	Preprint	High	High
[26]	Allahqoli	GPT-3.5	Free text	Unknown	Gynecology	NA	30	Peer-reviewed	High	Low
[27]	Levartovsky	GPT-4	Choice	External	Gastroenterology	NA	20	Peer-reviewed	High	Low
[28]	Bushuven	GPT-3.5, GPT-4	Free text, Choice	External	Emergency medicine	NA	22	Peer-reviewed	High	Low
[29]	Knebel	GPT-3.5	Free text, Choice	External	Ophthalmology	NA	10	Peer-reviewed	High	Low
[30]	Mitsuyama	GPT-4	Free text	External	Radiology	Expert, Non-expert	99	Preprint	High	Low
[31]	Pillai	GPT-3.5, GPT-4, Llama 2	Free text	Unknown	Endocrinology	Expert	20	Peer-reviewed	High	Low
[32]	Brin	GPT-4V	Free text	External	Radiology	NA	36	Preprint	High	Low
[33]	Horiuchi	GPT-4	Free text	External	Radiology	Expert, Non-expert	30	Preprint	High	High

[34]	Ito	GPT-4	Free text, Choice	Unknown	General medicine	Expert	45	Peer-reviewed	High	Low
[35]	Horiuchi	GPT-4, GPT-4V	Free text	External	Radiology	Expert, Non-expert	106	Preprint	Low	High
[36]	Madadi	GPT-3.5, GPT-4	Free text	Unknown	Ophthalmology	Expert	22	Preprint	High	Low
[37]	Sorin	GPT-4V	Free text	External	Ophthalmology	Non-expert	40	Preprint	High	Low
[38]	Delsoz	GPT-3.5, GPT-4	Free text	Unknown	Ophthalmology	Expert	20	Preprint	High	Low
[39]	Levine	GPT-3	Free text, Choice	External	General medicine	Expert	48	Preprint	High	Low
[40]	Schubert	GPT-4V	Free text	External	General medicine	NA	93	Preprint	High	High
[41]	Sultan	GPT-3.5	Free text	External	Pediatrics	NA	30	Peer-reviewed	High	Low
[42]	Kiyohara	PaLM2, GPT-3.5, GPT-4	Choice	Unknown	Cardiology	NA	66	Preprint	High	Low
[43]	Horiuchi	GPT-4	Free text	External	Neurology	NA	100	Peer-reviewed	Low	High
[44]	Stoneham	GPT-4	Free text	External	Dermatology	NA	36	Peer-reviewed	High	Low
[45]	Rundle	GPT-3.5	Free text	External	Dermatology	NA	39	Peer-reviewed	High	Low
[46]	Rojas-Carabali	GPT-3.5, GPT-4, Glass AI	Free text	External	Ophthalmology	NA	6	Peer-reviewed	High	Low
[47]	Fraser	GPT-3.5, GPT-4	Free text	Unknown	Emergency medicine	Expert	30	Peer-reviewed	High	Low
[48]	Krusche	GPT-4	Free text	External	Rheumatology	NA	132	Peer-reviewed	Low	Low
[49]	Galetta	GPT-4	Free text	External	Neurology	NA	24	Peer-reviewed	High	Low
[50]	Delsoz	GPT-3.5	Free text	Unknown	Ophthalmology	Non-expert	11	Peer-reviewed	High	Low
[51]	Hu	GPT-4	Free text	Unknown	Ophthalmology	NA	10	Peer-reviewed	High	Low
[52]	Abi-Rafeh	GPT-3.5	Free text	External	Plastic surgery	NA	16	Peer-reviewed	High	Low
[53]	Koga	PaLM2, GPT-3.5, GPT-4	Free text	External	Neurology	NA	25	Peer-reviewed	High	Low
[54]	Xv	GPT-3.5	Free text	External	Urology	Non-expert	306	Peer-reviewed	Low	Low

[55]	Reese	GPT-4	Free text	External	General medicine	NA	75	Preprint	High	Low
[56]	Han	GPT-3.5, GPT-4, GPT-4V, Llama 2, Med-42	Choice	Unknown	General medicine	NA	140, 348	Preprint	Unclear	High
[57]	Senkaiahliyan	GPT-4V	Free text	Unknown	Radiology	NA	69	Preprint	High	Low
[58]	Williams	GPT-3.5	Choice	External	Emergency medicine	Non-expert	500	Preprint	Low	Low
[59]	Tenner	GPT-3.5	Free text	External	Radiology	NA	40	Preprint	High	Low
[60]	Mori	GPT-4	Choice	External	Radiology	NA	151	Peer-reviewed	Low	Low
[61]	Mykhalko	GPT-3.5	Free text	External	General medicine	NA	50	Peer-reviewed	High	High
[62]	Andrade-Castellanos	GPT-3.5	Free text	External	General medicine	NA	10	Peer-reviewed	High	High
[63]	Daher	GPT-3.5	Free text	External	Orthopedic surgery	NA	29	Peer-reviewed	High	Low
[64]	Suthar	GPT-4	Free text	External	Radiology	NA	140	Peer-reviewed	Low	High
[65]	Nakaura	Prometheus, GPT-3.5	Free text	External	Radiology	NA	28	Peer-reviewed	High	Low
[66]	Berg	GPT-3.5, GPT-4	Free text	External	Emergency medicine	NA	30	Peer-reviewed	High	Low
[67]	Gebrael	GPT-4	Choice	External	Emergency medicine	NA	56	Peer-reviewed	High	Low
[68]	Ravipati	GPT-3.5	Free text	Unknown	Dermatology	NA	32	Peer-reviewed	High	Low

Figure legends

Figure 1: Eligibility criteria

Figure 2: Summary of Prediction Model Study Risk of Bias Assessment Tool (PROBAST) risk of bias

Assessment of risk of biases using the PROBAST tool for generative AI model studies included in the meta-analysis (N = 54). The participants and the outcome determination were predominantly at low risk of bias, but there was a high risk of bias for analysis (83%) and the overall evaluation (83%). Applicability for participants and outcomes shows a predominantly low concern, whereas overall applicability has 19% high concern.

Figure 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.

Figure 4: Generative AI performance among specialities

This figure demonstrates the differences in accuracy of generative AI models for specialties. Each horizontal line represents the range of accuracy differences between the speciality 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 speciality 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 speciality was greater than that for General medicine, whereas negative values (to the left) indicate that the model performance for the speciality was less than that for General medicine.

References

1. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training [Internet]. [cited 2023 Dec 26];Available from: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

2. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst [Internet] 2020;33:1877–901. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction

3. OpenAI, :, Achiam J, et al. GPT-4 Technical Report [Internet]. arXiv [cs.CL]. 2023;Available from: http://arxiv.org/abs/2303.08774

4. Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and Efficient Foundation Language Models [Internet]. arXiv [cs.CL]. 2023;Available from: http://arxiv.org/abs/2302.13971

5. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models [Internet]. arXiv [cs.CL]. 2023;Available from: http://arxiv.org/abs/2307.09288

6. Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling Language Modeling with Pathways. J Mach Learn Res [Internet] 2023 [cited 2023 Dec 26];24(240):1–113. Available from: https://www.jmlr.org/papers/v24/22-1144.html

7. Anil R, Dai AM, Firat O, et al. PaLM 2 Technical Report [Internet]. arXiv [cs.CL]. 2023;Available from: http://arxiv.org/abs/2305.10403

8. Thoppilan R, De Freitas D, Hall J, et al. LaMDA: Language Models for Dialog Applications [Internet]. arXiv [cs.CL]. 2022;Available from: http://arxiv.org/abs/2201.08239

9. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med [Internet] 2023;29(8):1930–40. Available from: http://dx.doi.org/10.1038/s41591-023-02448-8

10. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature [Internet] 2023;620(7972):172–80. Available from: http://dx.doi.org/10.1038/s41586-023-06291-2

11. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. Radiology [Internet] 2023;308(1):e231040. Available from: http://dx.doi.org/10.1148/radiol.231040

12. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA [Internet] 2023;330(1):78–80. Available from: http://dx.doi.org/10.1001/jama.2023.8288

13. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative Evaluation of Diagnostic Accuracy Between Google Bard and Physicians. Am J Med [Internet] 2023;136(11):1119–23.e18. Available from: http://dx.doi.org/10.1016/j.amjmed.2023.08.003

14. Shea Y-F, Lee CMY, Ip WCT, Luk DWA, Wong SSW. Use of GPT-4 to Analyze Medical Records of Patients With Extensive Investigations and Delayed Diagnosis. JAMA Netw Open [Internet] 2023;6(8):e2325000. Available from: http://dx.doi.org/10.1001/jamanetworkopen.2023.25000

15. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA [Internet] 2018;319(4):388–96. Available from: http://dx.doi.org/10.1001/jama.2017.19163

16. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews

and meta-analyses: the PRISMA statement. BMJ [Internet] 2009;339:b2535. Available from: http://dx.doi.org/10.1136/bmj.b2535

17. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med [Internet] 2019;170(1):51–8. Available from: http://dx.doi.org/10.7326/M18-1376

18. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ [Internet] 1997;315(7109):629–34. Available from: http://dx.doi.org/10.1136/bmj.315.7109.629

19. Chee J, Kwa ED, Goh X. "Vertigo, likely peripheral": the dizzying rise of ChatGPT. Eur Arch Otorhinolaryngol [Internet] 2023;280(10):4687–9. Available from: https://doi.org/10.1007/s00405-023-08135-1

20. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol [Internet] 2023;Available from: http://dx.doi.org/10.1016/j.jcjo.2023.07.016

21. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-Generated Differential Diagnosis Lists for Complex Case-Derived Clinical Vignettes: Diagnostic Accuracy Evaluation. JMIR Med Inform [Internet] 2023;11:e48808. Available from: http://dx.doi.org/10.2196/48808

22. Benoit JRA. ChatGPT for clinical vignette generation, revision, and evaluation [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.02.04.23285478v1

23. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. 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 [Internet] 2023;20(4):3378. Available from: http://dx.doi.org/10.3390/ijerph20043378

24. 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 [Internet] 2023;327:115351. Available from: http://dx.doi.org/10.1016/j.psychres.2023.115351

25. Ueda D, Walston SL, Matsumoto T, Deguchi R, Tatekawa H, Miki Y. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.05.04.23289493v1

26. Allahqoli L, Ghiasvand MM, Mazidimoradi A, Salehiniya H, Alkatout I. Diagnostic and Management Performance of ChatGPT in Obstetrics and Gynecology. Gynecol Obstet Invest [Internet] 2023;88(5):310–3. Available from: http://dx.doi.org/10.1159/000533177

27. 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 [Internet] 2023;118(12):2283–9. Available from: http://dx.doi.org/10.14309/ajg.0000000000002483

28. Bushuven S, Bentele M, Bentele 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 [Internet] 2023;47(1):123. Available from: http://dx.doi.org/10.1007/s10916-023-02019-x

29. Knebel D, Priglinger S, Scherer N, Klaas J, Siedlecki J, Schworm B. Assessment of ChatGPT in the Prehospital Management of Ophthalmological Emergencies - An Analysis of 10 Fictional Case Vignettes. Klin Monbl Augenheilkd [Internet] 2023;Available from: http://dx.doi.org/10.1055/a-2149-0447

30. Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.10.27.23297585v1

31. 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 [Internet] 2023;7:100213. Available from: http://dx.doi.org/10.1016/j.jtauto.2023.100213

32. Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 Multimodal Performance in Radiological Image Analysis [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.11.15.23298583v1

33. Horiuchi D, Tatekawa H, Oura T, et al. Comparison of the diagnostic performance from patient's medical history and imaging findings between GPT-4 based ChatGPT and radiologists in challenging neuroradiology cases [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.08.28.23294607v1

34. Ito N, Kadomatsu S, Fujisawa M, 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 [Internet] 2023;9:e47532. Available from: http://dx.doi.org/10.2196/47532

35. Horiuchi D, Tatekawa H, Oura T, et al. Comparison of the diagnostic accuracy among GPT-4 based ChatGPT, GPT-4V based ChatGPT, and radiologists in musculoskeletal radiology [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.12.07.23299707v1

36. Madadi Y, Delsoz M, Lao PA, et al. ChatGPT Assisting Diagnosis of Neuro-ophthalmology Diseases Based on Case Reports [Internet]. medRxiv. 2023;Available from: http://dx.doi.org/10.1101/2023.09.13.23295508

37. Sorin V, Kapelushnik N, Hecht I, et al. GPT-4 multimodal analysis on ophthalmology clinical cases including text and images [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.11.24.23298953v1

38. Delsoz M, Madadi Y, Munir WM, et al. Performance of ChatGPT in Diagnosis of Corneal Eye Diseases [Internet]. medRxiv. 2023;Available from: http://dx.doi.org/10.1101/2023.08.25.23294635

39. Levine DM, Tuwani R, Kompa B, et al. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model [Internet]. medRxiv. 2023;Available from: http://dx.doi.org/10.1101/2023.01.30.23285067

40. Schubert MC, Lasotta M, Sahm F, Wick W, Venkataramani V. Evaluating the multimodal capabilities of generative AI in complex clinical diagnostics [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.11.01.23297938v1

41. Sultan I, Al-Abdallat H, Alnajjar Z, et al. Using ChatGPT to Predict Cancer Predisposition Genes: A Promising Tool for Pediatric Oncologists. Cureus [Internet] 2023;15(10):e47594. Available from: http://dx.doi.org/10.7759/cureus.47594

42. Kiyohara Y, Kodera S, Sato M, et al. Large language models to differentiate vasospastic angina using patient information [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.06.26.23291913v1

43. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology [Internet] 2023;66(1):73–9. Available from: http://dx.doi.org/10.1007/s00234-023-03252-4

44. Stoneham S, Livesey A, Cooper H, Mitchell C. Chat GPT vs Clinician: challenging the diagnostic capabilities of A.I. in dermatology. Clin Exp Dermatol [Internet] 2023;Available from: http://dx.doi.org/10.1093/ced/llad402

45. Rundle CW, Szeto MD, Presley CL, Shahwan KT, Carr DR. Analysis of ChatGPT generated differential diagnoses in response to physical exam findings for benign and malignant cutaneous neoplasms. J Am Acad Dermatol [Internet] 2023 [cited 2023 Dec 29];Available from: http://dx.doi.org/10.1016/j.jaad.2023.10.040

46. Rojas-Carabali W, Sen A, Agarwal A, et al. Chatbots Vs. Human Experts: Evaluating Diagnostic Performance

of Chatbots in Uveitis and the Perspectives on AI Adoption in Ophthalmology. Ocul Immunol Inflamm [Internet] 2023;1–8. Available from: http://dx.doi.org/10.1080/09273948.2023.2266730

47. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. 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 [Internet] 2023;11:e49995. Available from: http://dx.doi.org/10.2196/49995

48. 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 [Internet] 2023;Available from: http://dx.doi.org/10.1007/s00296-023-05464-6

49. 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 [Internet] 2023;453:120804. Available from: http://dx.doi.org/10.1016/j.jns.2023.120804

50. Delsoz M, Raja H, Madadi Y, et al. The Use of ChatGPT to Assist in Diagnosing Glaucoma Based on Clinical Case Reports. Ophthalmol Ther [Internet] 2023;12(6):3121–32. Available from: http://dx.doi.org/10.1007/s40123-023-00805-x

51. Hu X, Ran AR, Nguyen TX, et al. What can GPT-4 do for Diagnosing Rare Eye Diseases? A Pilot Study. Ophthalmol Ther [Internet] 2023;12(6):3395–402. Available from: http://dx.doi.org/10.1007/s40123-023-00789-8

52. 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 [Internet] 2023;47:2407–14. Available from: http://dx.doi.org/10.1007/s00266-023-03538-1

53. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol [Internet] 2023;e13207. Available from: http://dx.doi.org/10.1111/bpa.13207

54. 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 [Internet] 2023;41(9):2569–71. Available from: http://dx.doi.org/10.1007/s00345-023-04539-0

55. Reese JT, Danis D, Caulfied JH, et al. On the limitations of large language models in clinical diagnosis [Internet]. medRxiv. 2023;Available from: http://dx.doi.org/10.1101/2023.07.13.23292613

56. Han T, Adams LC, Bressem K, et al. Comparative Analysis of GPT-4Vision, GPT-4 and Open Source LLMs in Clinical Diagnostic Accuracy: A Benchmark Against Human Expertise [Internet]. medRxiv. 2023 [cited 2023 Dec 29];2023.11.03.23297957. Available from: https://www.medrxiv.org/content/10.1101/2023.11.03.23297957v2

57. Senthujan SM, Toma A, Ma J, et al. GPT-4V(ision) Unsuitable for Clinical Care and Education: A Clinician-Evaluated Assessment [Internet]. medRxiv. 2023 [cited 2023 Dec 29];2023.11.15.23298575. Available from: https://www.medrxiv.org/content/10.1101/2023.11.15.23298575v1

58. Williams CYK, Zack T, Miao BY, Sushil M, Wang M, Butte AJ. Assessing clinical acuity in the Emergency Department using the GPT-3.5 Artificial Intelligence Model [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.08.09.23293795v1

59. Tenner ZM, Cottone M, Chavez M. Harnessing the open access version of ChatGPT for enhanced clinical opinions [Internet]. medRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.08.23.23294478v1

60. Mori Y, Izumiyama T, Kanabuchi R, Mori N, Aizawa T. Large language model may assist diagnosis of

SAPHO syndrome by bone scintigraphy. Mod Rheumatol [Internet] 2023;road115. Available from: http://dx.doi.org/10.1093/mr/road115

61. Mykhalko Y, Kish P, Rubtsova Y, Kutsyn O, Koval V. From Text To Diagnose: ChatGPT'S Efficacy In Medical Decision-Making. Wiad Lek [Internet] 2023;76(11):2345–50. Available from: http://dx.doi.org/10.36740/WLek202311101

62. Andrade-Castellanos CA, Paz MTT la, Farfán-Flores PE. Accuracy of ChatGPT for the diagnosis of clinical entities in the field of internal medicine. Gac Med Mex [Internet] 2023;159(5):439–42. Available from: http://dx.doi.org/10.24875/GMM.M23000824

63. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES Int [Internet] 2023;7(6):2534–41. Available from: http://dx.doi.org/10.1016/j.jseint.2023.07.018

64. Suthar PP, Kounsal A, Chhetri L, Saini D, Dua SG. 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 [Internet] 2023;15(8):e43958. Available from: http://dx.doi.org/10.7759/cureus.43958

65. Nakaura T, Yoshida N, Kobayashi N, et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. Jpn J Radiol [Internet] 2023;15:1–11. Available from: http://dx.doi.org/10.1007/s11604-023-01487-y

66. Berg HT, van Bakel B, van de Wouw L, et al. ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation. Ann Emerg Med [Internet] 2024;83(1):83–6. Available from: http://dx.doi.org/10.1016/j.annemergmed.2023.08.003

67. Gebrael G, Sahu KK, Chigarira B, 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 [Internet] 2023;15(14):3717. Available from: http://dx.doi.org/10.3390/cancers15143717

68. Ravipati A, Pradeep T, Elman SA. The role of artificial intelligence in dermatology: the promising but limited accuracy of ChatGPT in diagnosing clinical scenarios. Int J Dermatol [Internet] 2023;62(10):e547–8. Available from: http://dx.doi.org/10.1111/ijd.16746

69. GPT-4V(ision) System Card [Internet]. Open AI; 2023. Available from: https://cdn.openai.com/papers/GPTV_System_Card.pdf

70. Glass version 2.0 [Internet]. GLASS. [cited 2024 Jan 23];Available from: https://glass.health/ai

71. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health [Internet] 2018;3(4):e000798. Available from: http://dx.doi.org/10.1136/bmjgh-2018-000798

72. Preiksaitis C, Rose C. Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review. JMIR Med Educ [Internet] 2023;9:e48785. Available from: http://dx.doi.org/10.2196/48785

73. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med [Internet] 2023;3(1):141. Available from: http://dx.doi.org/10.1038/s43856-023-00370-1

74. The Lancet Digital Health. Large language models: a new chapter in digital health. Lancet Digit Health [Internet] 2024;6(1):e1. Available from: http://dx.doi.org/10.1016/S2589-7500(23)00254-6

75. Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. Jpn J Radiol [Internet] 2023;42:3–15. Available from: http://dx.doi.org/10.1007/s11604-023-01474-3