**Diagnostic Performance of 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], Shannon L Walston, MS[1], Hiroyuki Tatekawa, MD, PhD[1], Kenichi Saito, MD[2], Yasushi Tsujimoto, MD, MPH,[3,4,5] Yukio Miki, MD, PhD[1], Daiju Ueda, MD, PhD[1,6]

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) Center for Digital Transformation of Health Care, Graduate School of Medicine, Kyoto University, Shogoin-Kawahara-cho 53, Sakyo-ku, Kyoto 606-8507, Japan
3) Oku medical clinic, Shimmori 7-1-4, Asahi-ku, Osaka 535-0022, Japan
4) 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
5) Scientific Research WorkS Peer Support Group (SRWS-PSG), Koraibashi 1-7-7-2302, Chuo-ku, Osaka 541-0043, Japan
6) Center for Health Science Innovation, Osaka Metropolitan University, 1-4-3 Asahi-machi, Abeno-ku, Osaka 545-8585, Japan

**Corresponding author:** Daiju Ueda, MD, PhD

Department of Diagnostic and Interventional Radiology, 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

**Abstract**

**Background:** The rapid advancement of generative artificial intelligence (AI) has revolutionized 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. Meta-analysis was performed to summarize the performance of the models and to compare the accuracy of the models with that of physicians. The quality of studies was assessed using the Prediction Model Study Risk of Bias Assessment Tool.

**Results:** The search resulted in 54 studies being included in the meta-analysis, with 13 of these also used in the comparative analysis. Eight models were evaluated across 17 medical specialties. The overall accuracy for generative AI models across 54 studies was 57% (95% confidence interval [CI]: 51–63%). The I-squared statistic of 96% signifies a high degree of heterogeneity among the study results. Meta-regression analysis of generative AI models revealed significantly improved accuracy for GPT-4, and reduced accuracy for some specialties such as Neurology, Endocrinology, Rheumatology, and Radiology. The comparison meta-analysis demonstrated that, on average, physicians exceeded the accuracy of the models (difference in accuracy: 14% [95% CI: 8–19%], p-value <0.001). However, in the performance comparison between GPT-4 and physicians, GPT-4 performed slightly higher than non-experts (-4% [95% CI: -10–2%], p-value = 0.173), and slightly underperformed compared to experts (6% [95% CI: -1–13%], p-value = 0.091). The quality assessment indicated a high risk of bias in the majority of studies, primarily due to small sample sizes.

**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. This study also highlights the need for more rigorous research standards and a larger number of cases in the future.

## 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.[9,10]

The integration of generative AI models in the medical domain has spurred a growing body of research focusing on their diagnostic capabilities.[11] Studies have extensively examined the performance of these models in interpreting clinical data, understanding patient histories, and even suggesting possible diagnoses.[12,13] In medical diagnoses, 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 AIs 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.[14] 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,[12,13] 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 the comparison of 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"[1], "Bing", "Bard", "PaLM"[7,8], "Pathways Language Model", "LaMDA"[17], "Language Model for Dialogue Applications", "Llama"[5,6], 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, retracted articles, and those not related to diagnostic performance.

### 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),[18] 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 was used for training. Therefore, when generative AI models are tested with data outside of the training period, the test dataset type is classified as external testing, and when

tested with data that was publicly available during the training period, it is classified as unknown. In addition to this, when both the model and the physician's diagnostic performance are presented in the same paper, we extracted both for comparative 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.[19] 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**

Initially, we conducted a meta-analysis of generative AI studies reporting accuracy data to estimate the pooled accuracy of the diagnostic performance. Subsequently, a meta-regression analysis was performed on the accuracy of these models to identify sources of heterogeneity across studies, incorporating covariates such as model type, medical specialty, task of the model, type of test dataset, level of bias, and publication status. Secondly, we compared the diagnostic performance of generative AI models with that of physicians. For this analysis, we used the difference in accuracy, calculated by subtracting the physicians' accuracy from that of the models. An inverse-variance-weighted random-effects model with the DerSimonian–Laird estimator was utilized to estimate the between-study variance and normal approximation intervals based on summary measures to calculate confidence intervals (CI) for individual study results. The random-effects model (DerSimonian–Laird method) rather than a fixed-effects model was selected at the time of the study protocol because of the expected heterogeneity of the included studies. To assess publication bias, we used a funnel plot, evaluating effect size and standard error as described by Egger et al.[20] Statistical significance was set at a P value of 0.05. All calculations were performed using R (version 4.0.0), utilizing the 'metafor' package.

## Results

### Study Selection and Characteristics

We identified 13966 studies, of which 7940 were duplicates. After screening, 54 studies were included in the meta-analysis of generative AI diagnostic performance[12,13,21–72] and 13 studies in the comparative analysis between generative AI models and physicians (Figure 1 and Table 1).[32,33,35–41,49,52,58,62] The most evaluated models were GPT-4[4] (31 articles) and GPT-3.5 (28), while models such as GPT-4V (6), PaLM2[8] (3), Llama 2[6] (2), Prometheus (2), GPT-3[3] (1), Glass (1), and Med-42 (1) had less representation. OpenAI developed GPT-3, GPT-3.5, GPT-4, and GPT-4V, several of which are accessible through ChatGPT. Google's PaLM2 is implemented in its Bard system. Meta created Llama 2, and Med-42 is a fine-tuned version of Llama 2. Microsoft's Bing incorporates Prometheus, which is based on OpenAI's GPT technology. Lastly, Glass Health developed a model named Glass. 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 due to the training data for the generative AI models being 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 physicians.[32,33,35–41,49,52,58,62] GPT-4 (8 articles) was the most frequently evaluated, 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. The studies covered a variety of medical specialties. Ophthalmology was the most frequently studied specialty with 4 articles, followed by Radiology with 3 articles. General medicine and Emergency medicine were evaluated in 2 articles each. Endocrinology and Urology were each represented once. For model tasks, free text tasks were more prevalent with 12 articles, whereas choice tasks were represented in 3 articles. Regarding test types, external testing was more common with 7 articles, compared to 6 articles of unspecified or unknown test types.

### 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 for generative AI models**

The pooled accuracy of generative AI models showed varied performance across different models and medical specialties (Figure 3 and Appendix Figures S1–3 [online]). The overall accuracy for generative AI models was found to be 57% with a 95% CI of 51–63%. The I-squared statistic of 96% signifies a high degree of heterogeneity among the study results. In the meta-regression analysis examining the performance of various generative AI models across different specialties, the results revealed differences in effectiveness (Table 2). For the models, GPT-4 showed statistically significant performance with a coefficient of 26.1 (95% CI: 6.6–45.6, p = 0.009) while other models, such as GPT-3.5, GPT-4V, Llama 2, PaLM2 (Bard), and Prometheus (Bing), did not demonstrate significant results. Regarding the performance across specialties, wide variations were observed. The fields of Neurology, Endocrinology, Rheumatology, and Radiology displayed significant negative coefficients, with Neurology at -21.7 (95% CI: -41.2–-2.1, p = 0.03), Endocrinology at -42.0 (95% CI: -61.3–-22.6, p = 0.002), Rheumatology at -41.4 (95% CI: -78.5–-4.3, p = 0.029) and Radiology at -24.9 (95% CI: -40.7–-9.1, p <0.001). Other specialties such as Pediatrics, Gynecology, Urology, Otolaryngology, Orthopedic surgery, Ophthalmology, and Plastic surgery showed positive coefficients but not significant differences. No significant heterogeneity was observed based on the risk of bias, or based on publication status. Overall, the meta-regression analysis indicates that among various generative AI models, GPT-4 significantly outperforms others in effectiveness, though performance varies considerably across medical specialties, with some showing negative impacts.

We assessed publication bias by using a regression analysis to quantify funnel plot asymmetry (Appendix Figures S1 [online]) and it suggested a low risk of publication bias (p = 0.572).

**Meta-analysis comparing between generative AI models and physicians**

In our comparison meta-analysis, we observed that physicians generally outperformed generative AI models in various scenarios (Figure 4). This superiority was particularly evident when comparing AI models to overall physician performance, where physicians demonstrated a significant 14% higher performance on average (95% CI:

8–19%, p < 0.001). Though physicians overall and experts specifically both outperformed GPT-4, the differences were not statistically significant (4% difference, 95% CI: -2–10%, p = 0.192 against physicians overall, and 6% difference, 95% CI: -1–13%, p = 0.091 against experts). Interestingly, in the scenario of GPT-4 versus non-experts, GPT-4 showed a slight, yet not statistically significant, superiority (difference of -4%, 95% CI: -10–2%, p = 0.173). GPT-3.5 was also consistently outperformed by physicians, with performance 16% lower than that of physicians overall (95% CI: 7–24%, p < 0.001), 4% lower than that of non-experts (95% CI: 2–6%, p < 0.001), and a more pronounced 26% lower performance than that of experts (95% CI: 16–36%, p < 0.001). GPT-4V followed a similar pattern as GPT-3.5. GPT-4V had 22% lower performance (95% CI: 1–43%, p = 0.039) against physicians overall. Specifically, 14% lower performance against non-experts (95% CI: -7–35%, p = 0.188) and 44% lower performance than expert physicians (95% CI: 33–56%, p < 0.001). Similarly, Llama 2 also showed 47% lower performance than experts (95% CI: 33–61%, p < 0.001).

**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 and 13 in the comparative analysis with physicians. 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. Meta-regression analysis highlighted significant differences in effectiveness of different AI models across medical fields. The comparative analysis revealed that physicians generally outperformed AI models, although in non-expert settings, some AI models showed comparable performance. 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, the significant performance of GPT-4, suggests its potential utility in certain clinical scenarios. The variation in effectiveness across specialties, particularly the lower effectiveness in fields like Neurology, Endocrinology, Rheumatology, and Radiology 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 its characteristics, it has the potential to function as a valuable support tool in medical settings. Importantly, the close performance of 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.[73,74]

The comparison between generative AI and physician performances, particularly in the context of medical education, offers intriguing perspectives.[75] 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 GPT-4 and 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.[76] Such integration could enhance learning experiences, offering diverse clinical case studies and facilitating self-

assessment and feedback. Additionally, the narrower performance gap between GPT-4 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 among the risks of bias, the PROBAST quality assessment reveals a high risk of bias in 80% of studies.[19] 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.[77] 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. It 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,[78] ultimately enhancing the quality of healthcare delivery and medical education.

The methodology of this study, while comprehensive, has limitations. This meta-analysis involved primary studies with considerable heterogeneity. 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 the identified limitations. This includes 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, longitudinal studies assessing the long-term efficacy and impact of generative AI models in clinical practice 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 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 |
|---|---|---|---|---|---|---|---|---|---|---|
| 12 | Ueda | GPT-4 | Free text | External | Radiology | NA | 313 | Peer-reviewed | Low | High |
| 13 | Kanjee | GPT-4 | Free text | External | General medicine | NA | 70 | Peer-reviewed | High | Low |
| 21 | Chee | GPT-3.5 | Free text | External | Otolaryngology | NA | 7 | Peer-reviewed | High | Low |
| 22 | Lyons | Prometheus, GPT-4 | Free text, Choice | External | Ophthalmology | NA | 44 | Peer-reviewed | High | Low |
| 23 | Hirosawa | GPT-3.5, GPT-4 | Free text | Unknown | General medicine | NA | 52 | Peer-reviewed | High | Low |
| 24 | Benoit | GPT-3.5 | Free text, Choice | Unknown | General medicine | NA | 45 | Preprint | High | Low |
| 25 | Hirosawa | GPT-3.5 | Free text | External | General medicine | NA | 30 | Peer-reviewed | High | Low |
| 26 | Wei | GPT-4 | Choice | External | Psychiatry | NA | 60 | Peer-reviewed | High | Low |
| 27 | Ueda | GPT-4 | Free text | External | General medicine | NA | 62 | Preprint | High | High |
| 28 | Allahqoli | GPT-3.5 | Free text | Unknown | Gynecology | NA | 30 | Peer-reviewed | High | Low |
| 29 | Levartovsky | GPT-4 | Choice | External | Gastroenterology | NA | 20 | Peer-reviewed | High | Low |
| 30 | Bushuven | GPT-3.5, GPT-4 | Free text, Choice | External | Emergency medicine | NA | 22 | Peer-reviewed | High | Low |
| 31 | Knebel | GPT-3.5 | Free text, Choice | External | Ophthalmology | NA | 10 | Peer-reviewed | High | Low |
| 32 | Mitsuyama | GPT-4 | Free text | External | Radiology | Expert, Non-expert | 99 | Preprint | High | Low |
| 33 | Pillai | GPT-3.5, GPT-4, Llama 2 | Free text | Unknown | Endocrinology | Expert | 20 | Peer-reviewed | High | Low |
| 34 | Brin | GPT-4V | Free text | External | Radiology | NA | 36 | Preprint | High | Low |
| 35 | Horiuchi | GPT-4 | Free text | External | Radiology | Expert, Non-expert | 30 | Preprint | High | High |
| 36 | Ito | GPT-4 | Free text, Choice | Unknown | General medicine | Expert | 45 | Peer-reviewed | High | Low |
| 37 | Horiuchi | GPT-4, GPT-4V | Free text | External | Radiology | Expert, Non-expert | 106 | Preprint | Low | High |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 38 | Madadi | GPT-3.5, GPT-4 | Free text | Unknown | Ophthalmology | Expert | 22 | Preprint | High | Low |
| 39 | Sorin | GPT-4V | Free text | External | Ophthalmology | Non-expert | 40 | Preprint | High | Low |
| 40 | Delsoz | GPT-3.5, GPT-4 | Free text | Unknown | Ophthalmology | Expert | 20 | Preprint | High | Low |
| 41 | Levine | GPT-3 | Free text, Choice | External | General medicine | Expert | 48 | Preprint | High | Low |
| 42 | Schubert | GPT-4V | Free text | External | General medicine | NA | 93 | Preprint | High | High |
| 43 | Sultan | GPT-3.5 | Free text | External | Pediatrics | NA | 30 | Peer-reviewed | High | Low |
| 44 | Kiyohara | PaLM2, GPT-3.5, GPT-4 | Choice | Unknown | Cardiology | NA | 66 | Preprint | High | Low |
| 45 | Horiuchi | GPT-4 | Free text | External | Neurology | NA | 100 | Peer-reviewed | Low | High |
| 46 | Stoneham | GPT-4 | Free text | External | Dermatology | NA | 36 | Peer-reviewed | High | Low |
| 47 | Rundle | GPT-3.5 | Free text | External | Dermatology | NA | 39 | Peer-reviewed | High | Low |
| 48 | Rojas-Carabali | GPT-3.5, GPT-4, Glass | Free text | External | Ophthalmology | NA | 6 | Peer-reviewed | High | Low |
| 49 | Fraser | GPT-3.5, GPT-4 | Free text | Unknown | Emergency medicine | Expert | 30 | Peer-reviewed | High | Low |
| 50 | Krusche | GPT-4 | Free text | External | Rheumatology | NA | 132 | Peer-reviewed | Low | Low |
| 51 | Galetta | GPT-4 | Free text | External | Neurology | NA | 24 | Peer-reviewed | High | Low |
| 52 | Delsoz | GPT-3.5 | Free text | Unknown | Ophthalmology | Non-expert | 11 | Peer-reviewed | High | Low |
| 53 | Hu | GPT-4 | Free text | Unknown | Ophthalmology | NA | 10 | Peer-reviewed | High | Low |
| 54 | Hirosawa | PaLM2 | Free text | External | General medicine | NA | 82 | Peer-reviewed | High | Low |
| 55 | Abi-Rafeh | GPT-3.5 | Free text | External | Plastic surgery | NA | 16 | Peer-reviewed | High | Low |
| 56 | Shea | GPT-4 | Free text | External | General medicine | NA | 6 | Peer-reviewed | High | Low |
| 57 | Koga | PaLM2, GPT-3.5, GPT-4 | Free text | External | Neurology | NA | 25 | Peer-reviewed | High | Low |
| 58 | Xv | GPT-3.5 | Free text | External | Urology | Non-expert | 306 | Peer-reviewed | Low | Low |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 59 | Reese | GPT-4 | Free text | External | General medicine | NA | 75 | Preprint | High | Low |
| 60 | Han | GPT-3.5, GPT-4, GPT-4V, Llama 2, Med-42 | Choice | Unknown | General medicine | NA | 140, 348 | Preprint | Unclear | High |
| 61 | Senkaiahliyan | GPT-4V | Free text | Unknown | Radiology | NA | 69 | Preprint | High | Low |
| 62 | Williams | GPT-3.5 | Choice | External | Emergency medicine | Non-expert | 500 | Preprint | Low | Low |
| 63 | Tenner | GPT-3.5 | Free text | External | Radiology | NA | 40 | Preprint | High | Low |
| 64 | Mori | GPT-4 | Choice | External | Radiology | NA | 151 | Peer-reviewed | Low | Low |
| 65 | Mykhalko | GPT-3.5 | Free text | External | General medicine | NA | 50 | Peer-reviewed | High | High |
| 66 | Andrade-Castellanos | GPT-3.5 | Free text | External | General medicine | NA | 10 | Peer-reviewed | High | High |
| 67 | Daher | GPT-3.5 | Free text | External | Orthopedic surgery | NA | 29 | Peer-reviewed | High | Low |
| 68 | Suthar | GPT-4 | Free text | External | Radiology | NA | 140 | Peer-reviewed | Low | High |
| 69 | Nakaura | Prometheus, GPT-3.5 | Free text | External | Radiology | NA | 28 | Peer-reviewed | High | Low |
| 70 | Berg | GPT-3.5, GPT-4 | Free text | External | Emergency medicine | NA | 30 | Peer-reviewed | High | Low |
| 71 | Gebrael | GPT-4 | Choice | External | Emergency medicine | NA | 56 | Peer-reviewed | High | Low |
| 72 | Ravipati | GPT-3.5 | Free text | Unknown | Dermatology | NA | 32 | Peer-reviewed | High | Low |

Table 2: Meta regression results

| | Coefficient | Standard Error | Z-Value | P-Value |
|---|---|---|---|---|
| Model | | | | |
| GPT-3 | 7.5 (-25.7–40.7) | 16.9 | 0.4 | 0.658 |
| GPT-3.5 | 17.1 (-2.5–36.6) | 10.0 | 1.7 | 0.088 |
| GPT-4 | 26.1 (6.6–45.6) | 10.0 | 2.6 | 0.009 |
| GPT-4V | 18.4 (-4.0–40.8) | 11.4 | 1.6 | 0.107 |
| PaLM2 | 9.3 (-19.8–38.3) | 14.8 | 0.6 | 0.533 |
| Prometheus | 27.2 (-1.9–56.4) | 14.9 | 1.8 | 0.067 |
| Glass | -18.8 (-72.4–34.9) | 27.4 | -0.7 | 0.493 |
| Med-42 | 8.8 (-19.8–37.5) | 14.6 | 0.6 | 0.546 |
| Specialty | | | | |
| Gastroenterology | 1.5 (-37.1–40.2) | 19.7 | 0.1 | 0.939 |
| Cardiology | -13.5 (-40.3–13.2) | 13.6 | -1.0 | 0.322 |
| Neurology | -21.7 (-41.2–-2.1) | 10.0 | -2.2 | 0.03 |
| Emergency medicine | -14.8 (-30.7–1.1) | 8.1 | -1.8 | 0.068 |
| Pediatrics | 34.3 (-0.1–68.7) | 17.5 | 2.0 | 0.05 |
| Gynecology | 34.4 (-0.8–69.5) | 17.9 | 1.9 | 0.055 |
| Urology | 25.3 (-11.7–62.3) | 18.9 | 1.3 | 0.18 |
| Otolaryngology | 26.7 (-15.4–68.8) | 21.5 | 1.2 | 0.214 |
| Endocrinology | -42.0 (-61.3–-22.6) | 9.9 | -4.3 | <0.001 |
| Orthopedic surgery | 34.1 (-0.4–68.5) | 17.6 | 1.9 | 0.053 |
| Rheumatology | -41.4 (-78.5–-4.3) | 18.9 | -2.2 | 0.029 |
| Psychiatry | -25.2 (-61.8–11.5) | 18.7 | -1.3 | 0.179 |
| Ophthalmology | 10.1 (-3.3–23.6) | 6.9 | 1.5 | 0.14 |
| Dermatology | -7.4 (-30.0–15.1) | 11.5 | -0.6 | 0.518 |
| Plastic surgery | 28.5 (-8.5–65.4) | 18.8 | 1.5 | 0.131 |
| Radiology | -24.9 (-40.7–-9.1) | 8.1 | -3.1 | 0.002 |
| Task | | | | |
| Free text | -10.4 (-23.3–2.5) | 6.6 | -1.6 | 0.114 |
| Test dataset type | | | | |
| External | 3.4 (-8.0–14.8) | 5.8 | 0.6 | 0.557 |
| Risk of bias | | | | |
| High | -8.2 (-24.1–7.8) | 8.1 | -1.0 | 0.315 |
| Unclear | -8.8 (-34.3–16.6) | 13.0 | -0.7 | 0.497 |
| Publication status | | | | |
| Preprint | -0.4 (-12.0–11.2) | 5.9 | -0.1 | 0.944 |

**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 risk of bias regarding the participants (98%) and the outcome (96%) determination is predominantly low, but high at 83% for analysis and the overall evaluation. Applicability for participants and outcomes shows a predominantly low concern, whereas overall applicability has 19% high concern.

Figure 3: Pooled accuracy

This figure presents a comparative analysis of pooled accuracy across different studies. Panel A (left) illustrates the pooled accuracy for a range of models. Panel B (right) displays the pooled accuracy for various medical specialties. The dotted vertical line (57%) indicates the average pooled accuracy across all models or specialties, and the gray shaded area (51–63%) represents the 95% confidence interval (CI) for this average. Each data point is accompanied by a horizontal error bar which denotes the 95% CI. The numeric values in parentheses next to each entity represent the lower and upper bounds of the 95% CI for that specific model or specialty.

Figure 4: Comparison analysis results

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.

**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. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners [Internet]. [cited 2023 Dec 26];Available from: https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

3. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.

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

5. 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

6. 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

7. Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling Language Modeling with Pathways. J Mach Learn Res 2023;24(240):1–113.

8. 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

9. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29(8):1930–40.

10. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med 2023;6(1):158.

11. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80.

12. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. Radiology 2023;308(1):e231040.

13. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA 2023;330(1):78–80.

14. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature 2018;555(7695):175–82.

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 2018;319(4):388–96.

16. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535.

17. 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

18. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. Radiol Artif Intell 2020;2(2):e200029.

19. 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 2019;170(1):51–8.

20. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315(7109):629–34.

21. Chee J, Kwa ED, Goh X. "Vertigo, likely peripheral": the dizzying rise of ChatGPT. Eur Arch Otorhinolaryngol 2023;280(10):4687–9.

22. 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

23. 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 2023;11:e48808.

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

25. 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). Available from: http://dx.doi.org/10.3390/ijerph20043378

26. 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 2023;327:115351.

27. 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]. bioRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.05.04.23289493v1

28. Allahqoli L, Ghiasvand MM, Mazidimoradi A, Salehiniya H, Alkatout I. Diagnostic and Management Performance of ChatGPT in Obstetrics and Gynecology. Gynecol Obstet Invest 2023;88(5):310–3.

29. 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 2023;118(12):2283–9.

30. 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 2023;47(1):123.

31. 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

32. 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]. bioRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.10.27.23297585v1

33. 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 2023;7:100213.

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

35. 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]. bioRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.08.28.23294607v1

36. 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 2023;9:e47532.

37. 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]. bioRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.12.07.23299707v1

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

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

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

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

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

43. Sultan I, Al-Abdallat H, Alnajjar Z, et al. Using ChatGPT to Predict Cancer Predisposition Genes: A Promising Tool for Pediatric Oncologists. Cureus 2023;15(10):e47594.

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

45. 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;Available from: http://dx.doi.org/10.1007/s00234-023-03252-4

46. 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

47. 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;Available from: http://dx.doi.org/10.1016/j.jaad.2023.10.040

48. 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 2023;1–8.

49. 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 2023;11:e49995.

50. 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

51. 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 2023;453:120804.

52. Delsoz M, Raja H, Madadi Y, et al. The Use of ChatGPT to Assist in Diagnosing Glaucoma Based on Clinical Case Reports. Ophthalmol Ther 2023;12(6):3121–32.

53. Hu X, Ran AR, Nguyen TX, et al. What can GPT-4 do for Diagnosing Rare Eye Diseases? A Pilot Study. Ophthalmol Ther 2023;12(6):3395–402.

54. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative Evaluation of Diagnostic Accuracy Between Google Bard and Physicians. Am J Med 2023;136(11):1119–23.e18.

55. 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;Available from: http://dx.doi.org/10.1007/s00266-023-03538-1

56. 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 2023;6(8):e2325000.

57. 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 2023;e13207.

58. 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 2023;41(9):2569–71.

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

60. 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

61. 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

62. 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]. bioRxiv. 2023;Available from: https://www.medrxiv.org/content/10.1101/2023.08.09.23293795v1

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

64. 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;Available from: http://dx.doi.org/10.1093/mr/road115

65. Mykhalko Y, Kish P, Rubtsova Y, Kutsyn O, Koval V. FROM TEXT TO DIAGNOSE: CHATGPT'S EFFICACY IN MEDICAL DECISION-MAKING. Wiad Lek 2023;76(11):2345–50.

66. 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 2023;159(5):439–42.

67. 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 2023;7(6):2534–41.

68. 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 2023;15(8):e43958.

69. 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;Available from: http://dx.doi.org/10.1007/s11604-023-01487-y

70. 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 2024;83(1):83–6.

71. 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). Available from: http://dx.doi.org/10.3390/cancers15143717

72. 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 2023;62(10):e547–8.

73. Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? BMC Health Serv Res 2018;18(1):545.

74. 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 2018;3(4):e000798.

75. Preiksaitis C, Rose C. Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review. JMIR Med Educ 2023;9:e48785.

76. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med 2023;3(1):141.

77. The Lancet Digital Health. Large language models: a new chapter in digital health. Lancet Digit Health 2024;6(1):e1.

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

13966 records
3188 from MEDLINE
3139 from Scopus
6706 from Web of Science
794 from CENTRAL
139 from medRxiv

7940 duplicates removed

6026 records screened

5884 excluded

142 full-text articles assessed for eligibility

95 excluded:
81 Article without diagnostic accuracy
6 Unknown number of questions
3 Preprint of a peer-reviewed paper that has been published
2 Using no generative artificial intelligence
1 Article about examination problem
1 Article about students
1 Article about study protocol without results

54 articles included in systematic review and meta-analysis for generative AI models

7 articles added:
Citation search

41 excluded:
No data of diagnostic accuracy for physicians

13 articles included in meta-analysis for comparison between generative AI models and physicians