Society of Medical Education & Research

2022, Volume 2, Page No: 176-190
ISSN: 3108-4850

Annals of Pharmacy Education, Safety, and Public Health Advocacy

Artificial Intelligence in Health Professional Licensing: Performance of ChatGPT-3.5 and GPT-4; Systematic Review and Meta-Analysis

Andrew K. Foster[1], Natalie J. Price[1]*, Victoria L. Brown[1], Samuel T. Reed[1]

[1]Department of Pharmacy Practice and Patient Safety, School of Pharmacy, University of Bath, Bath, United Kingdom.

***E-mail** ✉ natalie.price@outlook.com

Abstract

ChatGPT, a recently launched AI chatbot, has shown notable performance in medical exams. However, there has been no comprehensive evaluation of the ChatGPT models (ChatGPT-3.5 and GPT-4) across multiple national health licensing exams. This study sought to systematically assess how ChatGPT performs in licensing examinations for medicine, pharmacy, dentistry, and nursing via a meta-analysis. Following the PRISMA guidelines, relevant full-text studies were retrieved from MEDLINE/PubMed, EMBASE, ERIC, Cochrane Library, Web of Science, and key journals, covering the period from ChatGPT's release up to February 27, 2024. Studies were included if they investigated ChatGPT-3.5 or GPT-4 performance in national licensing exams in medicine, pharmacy, dentistry, or nursing, used multiple-choice questions, and provided data suitable for effect size calculation. Data extraction, coding, and quality assessment were independently performed by two reviewers. The quality of included studies was assessed using the JBI Critical Appraisal Tools. A random-effects model was applied to compute pooled effect sizes with 95% confidence intervals (CIs).

A total of 23 studies met the inclusion criteria, covering four types of national licensing exams. Study distribution included medicine (n = 17), pharmacy (n = 3), nursing (n = 2), and dentistry (n = 1). Accuracy rates ranged from 36–77% for ChatGPT-3.5 and 64.4–100% for GPT-4. The overall pooled accuracy was 70.1% (95% CI, 65–74.8%), statistically significant ($p < 0.001$). Subgroup analyses showed that GPT-4 consistently outperformed ChatGPT-3.5. Across disciplines, the models demonstrated the highest performance in pharmacy, followed by medicine, dentistry, and nursing. Limitations included the narrow range of question types (absence of open-ended and scenario-based items) and considerable heterogeneity among studies. This analysis provides insight into ChatGPT's accuracy across four national health licensing exams and offers both practical guidance and theoretical support for future research. Further work should investigate the use of ChatGPT in healthcare education with more diverse question types and advanced AI versions.

Keywords: ChatGPT-3.5, GPT-4, National licensing examination, Healthcare professionals, Meta-analysis

Introduction

Artificial intelligence (AI) has undergone rapid development over the last decade, producing substantial advances in multiple domains [1, 2]. Among the most notable innovations is ChatGPT (OpenAI, San Francisco, CA), a language model capable of generating human-like text [3]. Since its release, ChatGPT has been applied in healthcare, education, research, business, and industry [4–6]. The system represents a sophisticated, evolving AI tool with potential utility for clinicians, educators, and students. For example, in education, advanced ChatGPT versions can provide personalized tutoring and learning guidance [7]. In healthcare, it can assist with diagnosis, treatment planning, and patient education by integrating knowledge with interactive conversation [1, 8–10].

With continuous improvements, ChatGPT's reliability is expected to increase, enhancing its applicability in health education. Gilson *et al*. reported that ChatGPT could

Access this article online https://smerpub.com/

Received: 28 February 2022; **Accepted:** 16 May 2022

How to cite this article: Foster AK, Price NJ, Brown VL, Reed ST. Artificial Intelligence in Health Professional Licensing: Performance of ChatGPT-3.5 and GPT-4; Systematic Review and Meta-Analysis. Ann Pharm Educ Saf Public Health Advocacy. 2022;2:176-90. https://doi.org/10.51847/bMVKZquCSZ

perform at the level of third-year US medical students, providing logical explanations and feedback [7]. The model has also shown effectiveness in the US Medical Licensing Examination (USMLE), with accuracy ranging from 40–60% [7, 11]. GPT-4 surpasses ChatGPT-3.5 in reliability and precision, delivering expert-level responses [12]. Yang *et al*. found GPT-4 achieved 90.7% accuracy on USMLE items requiring image interpretation, exceeding the approximate 60% passing mark [13]. GPT-4 additionally handled complex ethical, interpersonal, and professional medical tasks [14].

Despite these capabilities, adoption in healthcare education remains limited [15]. ChatGPT can generate inaccurate outputs [16], and reported accuracy varies across medical fields and exam formats [17, 18]. Calculations are a particular challenge for large language models (LLMs) [19], affecting performance in non-English National Medical (NMLE), Pharmacist (NPLE), and Nurse (NNLE) Licensing Examinations [20, 21]. Differences in curricula and exam content may explain performance variability [22, 23]. Although accuracy has improved over time, ChatGPT still scores below medical students [24, 25], leading to inconsistent and inconclusive findings.

To address this, a meta-analysis is needed to summarize evidence on ChatGPT's performance in licensing exams for medicine, pharmacy, dentistry, and nursing across various countries. This study aimed to evaluate ChatGPT-3.5 and GPT-4 performance without country-specific restrictions. Research questions included: (1) What is the overall effect size of ChatGPT models in these licensing exams? (2) Which model has the greatest impact on effect sizes? (3) Which discipline exams contribute most to ChatGPT performance outcomes?

Materials and Methods

Reporting framework

This meta-analysis and review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework [26].

Data sources

No protocol was formally registered for this review. Studies were retrieved from MEDLINE/PubMed, EMBASE, Cochrane Library, ERIC, and Web of Science. To ensure completeness, we also hand-searched key journals in healthcare education, examined reference lists of included studies [11, 13, 14, 21, 27–45], and reviewed prior systematic reviews [17, 46, 47]. All references were managed using EndNote™. The search was restricted to English-language articles published between November 30, 2022 (the launch date of OpenAI) and February 27, 2024.

Search strategy

A structured search was performed using terms related to ChatGPT, including "Artificial intelligence" [MeSH], "ChatGPT" [Title/Abstract], "GPT-4" [Title/Abstract], "Chatbot" [Title/Abstract], "Natural language processing" [Title/Abstract], and "Large language models" [Title/Abstract]. These were combined with terms targeting licensing examinations: "educational measurement" [MeSH], "licensure, medical" [MeSH], "licensure, pharmacy" [MeSH], "licensure, dental" [MeSH], "licensure, nursing" [MeSH], and "medical licensure exam*" [Title/Abstract]. Controlled vocabulary (MeSH for MEDLINE, ERIC thesaurus) and advanced search tools such as truncation were used to maximize retrieval. The search was executed by HK.

Eligibility criteria

Two independent reviewers (HK and HE) screened studies, resolving disagreements through discussion or consultation with a third reviewer (EY). Inclusion required that studies: (1) evaluated ChatGPT-3.5 or GPT-4 performance; (2) pertained to national licensing exams in medicine, pharmacy, dentistry, or nursing; (3) used multiple-choice questions; and (4) provided sufficient data to calculate effect sizes. Exclusion criteria included: (1) review articles or meta-analyses; (2) use of non-ChatGPT AI platforms; (3) unrelated to national licensing exams; (4) insufficient data for effect size computation; (5) open-ended question formats; (6) inaccessible full texts; and (7) non-English language publications.

Outcome measure

The primary metric was the accuracy of ChatGPT-3.5 and GPT-4 on licensing exams in medicine, pharmacy, dentistry, and nursing.

Risk of bias assessment

The Joanna Briggs Institute (JBI) critical appraisal tool for analytical cross-sectional studies [48] was employed to assess bias risk. The checklist includes eight items scored 1 (yes) or 0 (no/unclear/not applicable), yielding

a maximum of 8. Discrepancies were resolved via discussion. Risk was categorized as high if ≥2 domains were high risk, moderate if one domain was high or ≥2 were unclear, and low if all domains were low risk.

Data extraction and coding

Two reviewers (HK and HE) independently extracted data using a predesigned form; differences were resolved through discussion or adjudication by EY. Extracted details included descriptive study information (first author, year, journal, country) and study-specific data (ChatGPT version, discipline, number of questions, accuracy, and key findings). HK coded all 23 studies and trained additional coders, who then independently completed coding.

Handling dependent effect sizes

Some studies contributed multiple effect sizes, violating independence assumptions [49]. This review included 144 effect sizes from 23 studies, generating multivariate data. To address this, a "shifting unit of analysis" method was applied [50]. Overall effect sizes were calculated at the study level, while subgroup analyses used individual effect sizes to estimate separate effects.

Statistical analysis

Given the high heterogeneity among the included studies, all effect size calculations and moderator analyses were performed using a random-effects model. Analyses were executed with Comprehensive Meta-Analysis, version 4 (Biostat, Englewood, NJ, USA). For each study, the proportion of correctly answered questions relative to the total number of items was computed to represent effect sizes. Contributions of individual studies to the overall meta-analysis were displayed using forest plots. Between-study variability was quantified with Q statistics, and I^2 statistics were calculated to indicate the extent of heterogeneity. Sensitivity analyses were conducted by excluding studies deemed high risk of bias to assess the stability of the pooled results. Subgroup analyses were applied to explore potential sources of heterogeneity based on the type of ChatGPT model and the discipline of the health licensing exam. Meta-regression was not performed due to the absence of suitable continuous variables. Publication bias was evaluated using funnel plots, Egger's regression test, Duval and Tweedie's trim-and-fill procedure, and the classic fail-safe N test. All statistical evaluations were performed at 95% confidence intervals.

Ethical considerations

This meta-analysis relied exclusively on previously published data; therefore, institutional review board approval was not required.

Results and Discussion

Study selection

The initial database search identified 433 records. After duplicate removal and title/abstract screening, 31 articles were assessed at the full-text level. Exclusions included one non-English publication, three studies using non–multiple-choice formats, and eight studies lacking sufficient data for effect size calculation. Consequently, 23 studies were retained for inclusion in the meta-analysis **(Figure 1)**.

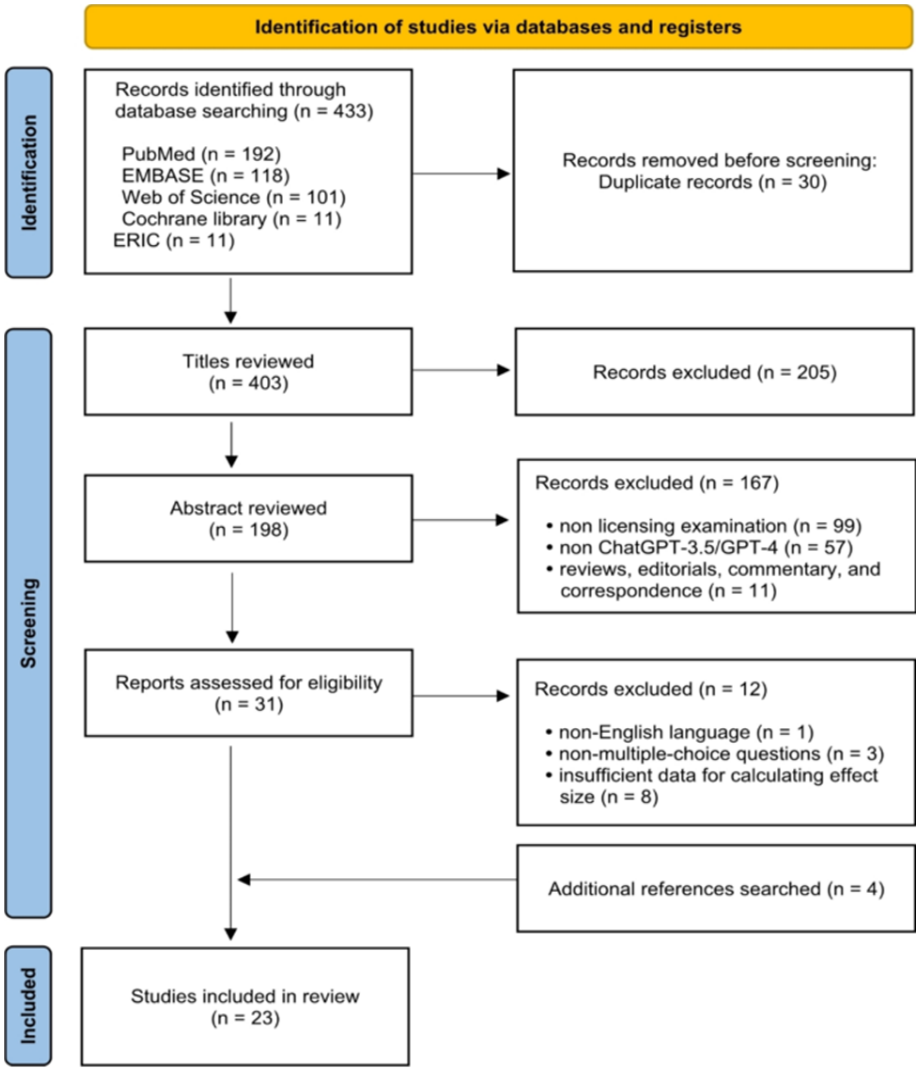

Figure 1. PRISMA flow diagram of study inclusion

Risk of bias assessment

Table 1 presents the results of the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for analytical cross-sectional studies. Two studies were classified as high risk [33, 34], 18 as low risk [11, 13, 21, 27, 29–31, 35–45], with common limitations related to identifying and addressing confounding variables. The remaining studies were rated as moderate risk [14, 28, 32].

Table 1. Risk of bias evaluation using JBI checklist (score range 0–8)

Study ID	Q8	Q7	Q6	Q5	Q4	Q3	Q2	Q1	Total score[a]	Overall Risk of Bias
Aljindan *et al.* [27]	1	1	1	1	1	1	1	1	8	Low
Angel *et al.* [28]	1	1	0	1	1	1	1	1	7	Moderate
Brin *et al.* [14]	1	1	0	1	1	1	1	1	7	Moderate
Fang *et al.* [29]	1	1	1	1	1	1	1	1	8	Low
Flores-Cohaila *et al.* [30]	1	1	1	1	1	1	1	1	8	Low
Fuchs *et al.* [31]	1	1	1	1	1	1	1	1	8	Low
Huang [32]	1	1	0	1	1	1	1	1	7	Moderate
Kataoka *et al.* [33]	1	1	0	0	1	1	1	1	6	High

Kleinig *et al*. [34]	1	1	0	0	1	1	1	1	6	High
Kung *et al*. [11]	1	1	1	1	1	1	1	1	8	Low
Kunitsu [35]	1	1	1	1	1	1	1	1	8	Low
Lai *et al*. [36]	1	1	1	1	1	1	1	1	8	Low
Mihalache *et al*. [37]	1	1	1	1	1	1	1	1	8	Low
Morreel *et al*. [38]	1	1	1	1	1	1	1	1	8	Low
Taira *et al*. [39]	1	1	1	1	1	1	1	1	8	Low
Takagi *et al*. [40]	1	1	1	1	1	1	1	1	8	Low
Tanaka *et al*. [41]	1	1	1	1	1	1	1	1	8	Low
Tong *et al*. [42]	1	1	1	1	1	1	1	1	8	Low
Wang *et al*. [43]	1	1	1	1	1	1	1	1	8	Low
Wang *et al*. [44]	1	1	1	1	1	1	1	1	8	Low
Wang *et al*. [21]	1	1	1	1	1	1	1	1	8	Low
Yanagita *et al*. [45]	1	1	1	1	1	1	1	1	8	Low
Yang *et al*. [13]	1	1	1	1	1	1	1	1	8	Low

Q1: Inclusion criteria clearly defined; Q2: Study subjects and setting adequately described; Q3: Exposure measured reliably; Q4: Standardized criteria used for outcome measurement; Q5: Confounders identified; Q6: Strategies for confounders reported; Q7: Outcomes measured validly; Q8: Appropriate statistical methods applied.

Study characteristics

The 23 studies [11, 13, 14, 21, 27–45] were published between 2022 and 2024 **(Table 2)**. By region, five studies were from the United States [11, 13, 14, 28, 37], one from the UK [36], six from Japan [33, 35, 39–41, 45], four from China [29, 42–44], two from Taiwan [21, 32], and one each from Australia [34], Belgium [38], Peru [30], Switzerland [31], and Saudi Arabia [27]. In terms of discipline, 17 studies examined medicine [11, 13, 14, 27, 29, 30, 33, 34, 36–38, 40–45], three pharmacy [21, 28, 35], two nursing [32, 39], and one dentistry [31]. The number of questions per exam ranged from 21 (USMLE soft skills, ethics, and communication items [14]) to 1510 across four exams in the Registered Nurse License Exam [32].

Table 2. Characteristics of included studies

No.	Country	Study ID	Objectives Related to ChatGPT Models	Number of Questions	Field	Accuracy
1	Saudi Arabia	Aljindan *et al*. [27]	Evaluating GPT-4's proficiency in responding to questions from the Saudi Medical Licensing Exam	220	Medicine	GPT-4: Total accuracy 88.6%. Performance differed based on question difficulty and answer choices. Option A: 82.7% (43/52), Option B: 93.3% (42/45), Option C: 88.1% (37/42), Option D: 93.8% (76/81)
2	USA	Angel *et al*. [28]	Examining the strengths and weaknesses of GPT-3, GPT-4, and Bard on sample questions from the North American Pharmacist Licensure Examination	137	Pharmacy	GPT-4: 78.8%, representing a 27.7% improvement compared to GPT-3, and exceeding the minimum passing threshold of 75
3	USA	Brin *et al*. [14]	Assessing the performance of ChatGPT and GPT-4 on United States Medical Licensing Exam (USMLE) questions focused on	21	Medicine	GPT-4: 100% accuracy, ChatGPT: 66.6% accuracy

			communication skills, ethics, empathy, and professionalism			
4	China	Fang *et al.* [29]	Evaluating ChatGPT's results on the Chinese National Medical Licensing Examination (NMLE)	600	Medicine	ChatGPT: Overall 73.7% (442/600), exceeding the passing score of 360/600. Results varied by unit, with the highest in Unit 4: 76.0%, and the lowest in Unit 2: 70.0%
5	Peru	Flores-Cohaila *et al.* [30]	Examining the precision of ChatGPT using GPT-3.5 and GPT-4 on the Peruvian National Licensing Medical Examination (ENAM)	180	Medicine	GPT-4: 86% (155/180), GPT-3.5: 77% (139/180) for multiple-choice questions requiring justification and 73% (133/180) for those without; passing score on the ENAM is 10.5 on a vigesimal scale: 95/180
6	Switzerland	Fuchs *et al.* [31]	Investigating ChatGPT and GPT-4 performance on self-assessment questions in dentistry from the Swiss Federal Licensing Examination in Dental Medicine (SFLEDM)	32	Dentistry	GPT-4: 64.4% (without priming), 66.7% (with priming), ChatGPT: 59.3% (without priming), 62.6% (with priming); typical passing threshold for the SFLEDM is 60%
7	Taiwan	Huang [32]	Evaluating ChatGPT's results on the Registered Nurse License Exam (RNLE), including its benefits and drawbacks	1510 (4 exams: 400 for 2022 1st, 370 each for 2022 2nd, 2023 1st, and 2023 2nd)	Nursing	ChatGPT: 63.75% (2022 1st exam), 58.37% (2022 2nd exam), 54.05% (2023 1st exam), 60% (2023 2nd exam); passing requires an average of 60 across five subjects, each scored out of 100
8	Japan	Kataoka *et al.* [33]	Examining the precision of ChatGPT and Bing responses to the National Medical Licensure Examination in Japan	281	Medicine	ChatGPT: Overall 38% (106/281), compulsory questions: 42% (34/81), other questions: 36% (72/200). Bing: 78% (219/281)
9	Australia	Kleinig *et al.* [34]	Evaluating ChatGPT-3.5, ChatGPT-4, and New Bing on publicly available Australian Medical Council Licensing Examination practice questions	50	Medicine	ChatGPT-3.5: 66% (99/150), ChatGPT-4: 79.3% (119/150)
10	USA	Kung *et al.* [11]	Examining ChatGPT's performance across the three United States Medical Licensing Exam (USMLE) steps: Step 1, Step 2CK, and Step 3	350	Medicine	Step 1: 55.8% (without justification), 64.5% (with justification), Step 2CK: 59.1% (without justification), 52.4% (with justification), Step 3: 61.3% (without justification), 65.2% (with justification)
11	Japan	Kunitsu [35]	Assessing GPT-4's capability on questions from the Japanese National Examination for Pharmacists (JNEP)	107th and 108th JNEP exams, with 345	Pharmacy	GPT-4: 107th exam: Overall 64.5% (222/344), 108th exam: Overall 62.9% (217/345). Rose to 78.2% (222/284) and 75.3% (217/288), respectively, for answerable

				questions across three blocks; GPT-4 answered 284 and 287 questions, respectively		questions. Reduced accuracy in physics, chemistry, and calculation questions
12	UK	Lai *et al.* [36]	Evaluating ChatGPT performance on the United Kingdom Medical Licensing Assessment	191	Medicine	GPT-4: Average score of 76.3% (437/573) across attempts. Attempt 1: 74.9% (143/191), Attempt 2: 78.0% (149/191), Attempt 3: 75.6% (145/191)
13	USA	Mihalache *et al.* [37]	Examining ChatGPT-4's results on practice questions for USMLE Step 1, Step 2CK, and Step 3	319	Medicine	GPT-4: 88% on Step 1 (82/93), 86% on Step 2CK (91/106), and 90% on Step 3 (108/120)
14	Belgium	Morrel *et al.* [38]	Evaluating the precision of ChatGPT with GPT-3.5 and GPT-4 on the University of Antwerp Medical Licensing Exam	95	Medicine	GPT-4: 76% (72/95), ChatGPT: 67% (64/95)
15	Japan	Taira *et al.* [39]	Examining ChatGPT's performance on Japanese National Nurse Examinations from 2019 to 2023	240 questions each year, divided into basic knowledge and general questions	Nursing	ChatGPT: Five-year average accuracy: 75.1% for basic knowledge questions, 64.5% for general questions; passing requires 80% for basic knowledge and about 60% for general
16	Japan	Takagi *et al.* [40]	Comparing GPT-3.5 and GPT-4 performance on the Japanese Medical Licensing Examination	254	Medicine	GPT-3.5: 50.8%, GPT-4: 79.9%
17	Japan	Tanaka *et al.* [41]	Assessing GPT-3.5 and GPT-4 on the NMLE and comparing to the exam's minimum passing rate	116th: 290, 117th: 262	Medicine	116th exam: GPT-3.5: 63.1% (183/290), GPT-4: 82.8% (240/290), 117th exam: GPT-4: 78.6% (206/262) with tuned prompt
18	China	Tong *et al.* [42]	Investigating the suitability and constraints of ChatGPT (4.0) for the Chinese Medical Licensing Examination in English and Chinese	160	Medicine	GPT-4: 81.25% for Chinese versions, 86.25% for English versions
19	China	Wang *et al.* [43]	Comparing GPT-3.5 and GPT-4 on the China National Medical Licensing Examination (CNMLE) and its English translation (ENMLE)	200 (100 in CNMLE, 100 in ENMLE)	Medicine	GPT-3.5: 56% (56/100) for CNMLE, 76% (76/100) for ENMLE. GPT-4: 84% (84/100) for CNMLE, 86% (86/100) for ENMLE

20	China	Wang *et al.* [44]	Comparing medical students' and ChatGPT's performance on the Chinese NMLE	600 (4 units with 150 per unit)	Medicine	ChatGPT: 47.0% (282/600) in 2020, 45.8% (275/600) in 2021, 36.5% (219/600) in 2022. ChatGPT scored lower than medical students on the Chinese NMLE
21	Taiwan	Wang *et al.* [21]	Assessing ChatGPT's precision on the Taiwanese Pharmacist Licensing Examination and its potential role in pharmacy education	826 (203 in Stage 1 and 210 in Stage 2 for Chinese and English, respectively)	Pharmacy	ChatGPT: Stage 1: 54.4% for Chinese, 56.9% for English. Stage 2: 53.8% for Chinese, 67.6% for English
22	Japan	Yanagita *et al.* [45]	Evaluating correct responses from GPT-3.5 and GPT-4 on Japan's NMLE	Out of 400 questions, 292 excluding those with unsupported charts	Medicine	GPT-4: 81.5% (237/292), exceeding the passing threshold of 72%. GPT-3.5: 42.8% (125/292)
23	USA	Yang *et al.* [13]	Comparing GPT-4 V, GPT-4, and ChatGPT on medical licensing exam questions	376 questions (119 in Step 1, 120 in Step 2CK, 137 in Step 3). Among these, 51 questions included images (19 in Step 1, 14 in Step 2CK, 18 in Step 3)	Medicine	GPT-4 V: 88.2–92.7%, GPT-4: 80.8–88.3%, ChatGPT: 55.1–60.9%. GPT-4 V met or exceeded USMLE standards except in anatomy, emergency medicine, and pathology (25–50%). Image-based questions: ChatGPT: 42.1–50.0%, GPT-4: 63.2–66.7%

ChatGPT performance

ChatGPT-3.5 showed considerable variation in performance, with scores ranging from 38% on the Japanese National Medical Licensure Examination [33] to 73% in the Peruvian National Licensing Medical Examination [30]. GPT-4 generally outperformed ChatGPT-3.5, with accuracy ranging from 64.4% in the Swiss Federal Licensing Examination for Dental Medicine [31] to 100% in USMLE assessments of soft skills [14]. These results demonstrate GPT-4's enhanced ability to perform accurately across different exams and regions.

Publication bias

Visual inspection of the funnel plot indicated slight asymmetry **(Figure 2)**. Egger's regression test returned a significant p-value ($p = 0.028$), suggesting potential publication bias. Nevertheless, the trim-and-fill method did not require the addition or removal of studies, leaving the overall effect size unchanged. The fail-safe N, calculated as 310 in CMA, exceeded the threshold defined by $5n + 10$ [51], confirming that publication bias is unlikely to have substantially influenced the results.

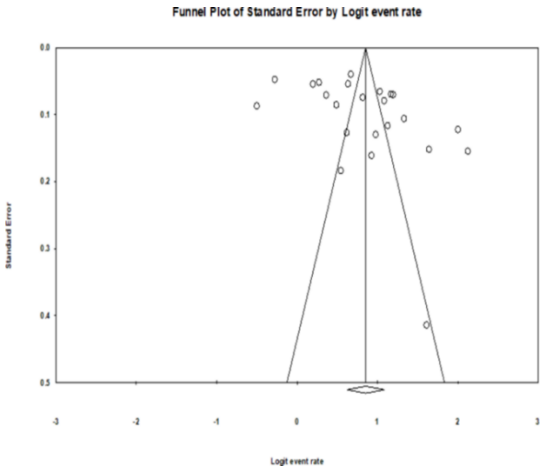

Figure 2. Funnel plot depicting the included studies

Overall analysis

A random-effects model was applied due to anticipated differences in effect sizes across studies stemming from variations in examination formats. Heterogeneity among studies was substantial ($Q = 1201.303$, $df = 22$, $p < 0.001$), with an I^2 value of 98.2%. The aggregated accuracy of ChatGPT models across all studies was 70.1% (95% CI, 65.0–74.8%). Individual study contributions and the pooled estimate are visualized in the forest plot **(Figure 3)**.

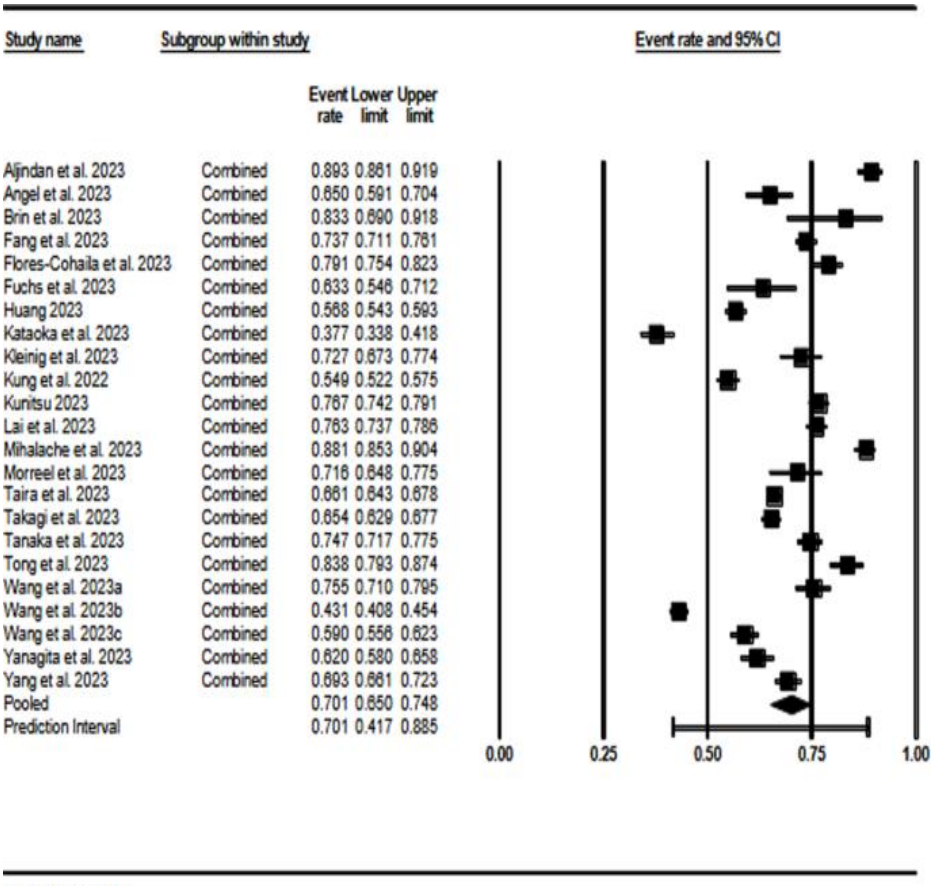

Figure 3. Forest plot illustrating effect sizes using a random-effects approach

Subgroup analyses

Subgroup analyses under the random-effects framework were conducted for two categorical moderators: (1) the ChatGPT version (ChatGPT-3.5 vs. GPT-4) and (2) the discipline of the health licensing examination (medicine, pharmacy, dentistry, nursing), as summarized in **Table 3**. Examination of additional potential moderators was limited due to infrequent reporting or incomplete information across studies.

Table 3. Subgroup analysis results

Variables	Classification	Point Estimate	95% CI Upper	95% CI Lower	df	Q	p	Notes
ChatGPT models	ChatGPT-3.5	0.589	0.616	0.562	1	177.027	< 0.001	
	GPT-4	0.804	0.820	0.786				
Fields of health licensing examinations	Pharmacy[a]	0.715	0.762	0.663	3	15.334	0.002	a, b > d
	Medicine[b]	0.697	0.732	0.659				
	Dentistry[c]	0.632	0.711	0.546				
	Nursing[d]	0.618	0.649	0.587				

By ChatGPT version

The pooled effect size for ChatGPT-3.5 was 58.9% (95% CI, 56.2–61.6%), whereas GPT-4 achieved 80.4% (95% CI, 78.6–82.0%). This difference was statistically significant ($Q = 177.027$, $p < 0.001$), highlighting the superior performance of the more recent model.

By examination discipline

Average performance varied by field: pharmacy, 71.5% (95% CI, 66.3–76.2%); medicine, 69.7% (95% CI, 65.9–73.2%); dentistry, 63.2% (95% CI, 54.6–71.1%); nursing, 61.8% (95% CI, 58.7–64.9%). The differences between disciplines were significant ($Q = 15.334$, $p = 0.002$), indicating field-specific variability in model performance.

Sensitivity analysis

Excluding studies with high risk of bias did not materially affect the overall estimate, which remained 71.3% (95% CI, 66.3–75.8%), with an I^2 of 98.1%, confirming the robustness of the findings ($p < 0.001$) **(Figure 4)**.

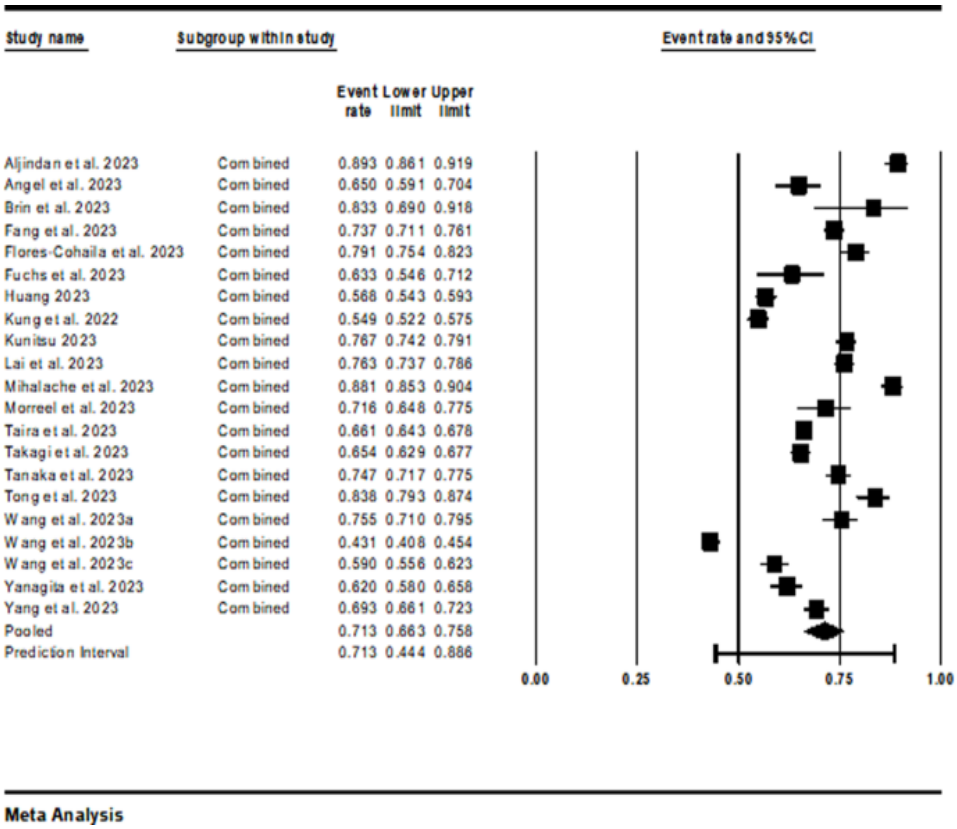

Figure 4. Sensitivity analysis excluding high-risk studies

Overall, most included studies were of high methodological quality, although the main limitations pertained to incomplete identification of confounding factors and inadequate strategies to address them. To our knowledge, this is the first meta-analysis systematically evaluating the performance of ChatGPT (3.5 and GPT-4) across four types of national health licensing exams. Prior research predominantly focused on academic knowledge assessments rather than formal licensing examinations [25, 52–54], or combined educational and licensing contexts without distinction [17, 46, 47]. Furthermore, earlier studies often reported results for a single ChatGPT version rather than comparing multiple versions [17, 29, 32, 37, 39, 44].

Our findings clearly demonstrate that GPT-4 consistently outperforms ChatGPT-3.5 in accuracy across diverse national medical licensing exams. GPT-4 not only exhibits higher overall accuracy but also demonstrates greater reliability, suggesting its superiority as a tool for professional assessments. Previous research indicated that a model accuracy above 95% could render ChatGPT a dependable educational resource [55]. While it is uncertain if future iterations will reach this threshold, ongoing improvements through user feedback and deep learning algorithms are likely to enhance its utility in medical education.

Regarding model comparisons, our analysis underscores GPT-4's advantages, reflecting the impact of its expanded training dataset. This aligns with earlier findings on USMLE questions, where GPT-4 achieved 84.7% accuracy compared to 56.9% for ChatGPT-3.5 [56]. Similarly, in the Peruvian National Licensing Medical Examination, GPT-4 scored 86.1%, whereas ChatGPT-3.5 scored 77.2% [30]. These observations suggest that GPT-4's superior performance may be attributed to enhanced reasoning and critical thinking capabilities [57].

The present analysis demonstrated that, within the context of health licensing examinations, ChatGPT models performed best in pharmacy, followed sequentially by medicine, dentistry, and nursing. These findings contrast with prior studies from China, which reported superior performance in the NNLE, with the NMLE and NPLE ranking lower; this discrepancy may stem from differences in the GPT model versions as well as variations in the complexity and difficulty of exam questions [20]. Language and cultural factors may further contribute to differences in performance. In line with this, Seghier highlighted that these AI models face challenges with non-English inputs, resulting in notably reduced accuracy for responses in other languages [58]. To address these limitations, expanding training datasets to include multiple languages is recommended, enhancing ChatGPT's utility as an educational assistant for both students and professionals across diverse learning environments. Additionally, while this meta-analysis included studies spanning several disciplines, research on dentistry remains limited, with only a single study available; further investigation is required to strengthen the generalizability of these conclusions across educational fields.

Another consideration is the rapid evolution of medical knowledge, which underscores the need for high-quality, up-to-date data to improve model performance. Given the annual introduction of new medications and updated clinical guidelines, ChatGPT's knowledge—currently limited to data up to 2021—may not reflect the most recent information. Prior studies have also noted that ChatGPT can produce inaccurate or seemingly plausible but incorrect outputs, commonly referred to as "hallucinations" [16, 32, 38, 43]. The versions analyzed here, ChatGPT-3.5 and GPT-4, represent earlier and more recent iterations, respectively, but rapid advancements in AI mean that even these versions may soon appear limited. Research indicates that several limitations observed in ChatGPT-3.5, including hallucinations, persist in GPT-4 [57]. Future model development must account for these limitations while leveraging potential improvements, emphasizing continuous refinement and retraining to enhance reliability and practical applicability.

Certain limitations of this study warrant attention when interpreting the results. First, although this systematic review followed rigorous PRISMA guidelines, the review protocol was not registered. Second, heterogeneity in study designs and inclusion criteria across countries represents a major limitation. Differences in exam structure, difficulty, and content may have influenced ChatGPT's performance, complicating cross-country comparisons. Third, all studies included only multiple-choice questions, which capture a limited aspect of medical knowledge and differ from real-world educational scenarios; future research should incorporate open-ended or scenario-based questions to better simulate clinical reasoning. Fourth, despite demonstrating potential in answering licensing exam questions, ChatGPT models should not be solely relied upon for exam preparation. These LLMs can

produce unreliable or false outputs [59], so results must be corroborated against authoritative educational resources, particularly for high-stakes exams. Fifth, this review did not examine grey literature, which could provide additional insights and reduce publication bias. Finally, while the focus was on ChatGPT, other LLMs, such as Google Bard, Microsoft Bing Chat, and Meta LLaMA, are advancing rapidly. Future research should explore these alternative models to provide a more comprehensive view of AI efficacy in medical and health education.

Conclusion

This meta-analysis evaluated the performance of ChatGPT-3.5 and GPT-4 across four types of health licensing examinations. The results indicate that GPT-4 consistently outperformed ChatGPT-3.5 in terms of accuracy. Additionally, the models demonstrated the highest proficiency in pharmacy, followed by medicine, dentistry, and nursing. Future studies should incorporate larger and more diverse question sets, as well as examine newer generations of AI chatbots, to gain a deeper understanding of their applications in health education and clinical practice.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Holzinger A, Keiblinger K, Holub P, Zatloukal K, Müller H. AI for life: trends in artificial intelligence for biotechnology. N Biotechnol. 2023;74:16–24. 10.1016/j.nbt.2023.02.001.
2. Montejo-Ráez A, Jiménez-Zafra SM. Current approaches and applications in natural language processing. Appl Sci. 2022;12(10):4859. 10.3390/app12104859.
3. Open AI. Introducing ChatGPT. San Francisco. https://openai.com/blog/chatgpt. Accessed 10 2024.
4. Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L, Generative. AI and ChatGPT: applications, challenges, and AI-human collaboration. J Inf Technol Case Appl Res. 2023;25(3):277–304. 10.1080/15228053.2023.2233814.
5. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber Phys Syst. 2023;3:121–54. 10.1016/j.iotcps.2023.04.003.
6. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388:1233–9. 10.1056/NEJMsr2214184.
7. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. 10.2196/45312.
8. Nakhleh A, Spitzer S, Shehadeh N. ChatGPT's response to the diabetes knowledge questionnaire: implications for diabetes education. Diabetes Technol Ther. 2023;25(8):571–3. 10.1089/dia.2023.0134.
9. Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. Cureus. 2023;15(5):e38755. 10.7759/cureus.38755.
10. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for Ai-assisted medical education and decision making in radiation oncology. Front Oncol. 2023;13:1265024. 10.3389/fonc.2023.1265024.
11. Kung TH, Cheatham M, Medenilla A, Sillos C, de Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198. 10.1371/journal.pdig.0000198.
12. OpenAI. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. https://openai.com/product/gpt-4. Accessed 10 Jan 2024.
13. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Ouyang F et al. Performance of multimodal GPT-4V on USMLE with image: potential for imaging diagnostic support with explanations. medRxiv 202310.26.23297629. 10.1101/2023.10.26.23297629

14. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. Sci Rep. 2023;13:16492. 10.1038/s41598-023-43436-9.
15. O'Connor S, Yan Y, Thilo FJS, Felzmann H, Dowding D, Lee JJ. Artificial intelligence in nursing and midwifery: a systematic review. J Clin Nurs. 2023;32(13–14):2951–68. 10.1111/jocn.16478.
16. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. 2023;27(1):120. 10.1186/s13054-023-04393-x.
17. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG. 2024;131:378–80. 10.1111/1471-0528.17641.
18. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng. 2024;52(6):1542–5. 10.1007/s10439-023-03338-3.
19. Shakarian P, Koyyalamudi A, Ngu N, Mareedu L. An independent evaluation of ChatGPT on mathematical word problems (MWP). 10.48550/arXiv.2302.13814
20. Zong H, Li J, Wu E, Wu R, Lu J, Shen B. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med Educ. 2024;24(1):143. 10.1186/s12909-024-05125-7.
21. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc. 2023;86(7):653–8. 10.1097/JCMA.0000000000000942.
22. Price T, Lynn N, Coombes L, Roberts M, Gale T, de Bere SR, et al. The international landscape of medical licensing examinations: a typology derived from a systematic review. Int J Health Policy Manag. 2018;7(9):782–90. 10.15171/ijhpm.2018.32.
23. Zawiślak D, Kupis R, Perera I, Cebula G. A comparison of curricula at various medical schools across the world. Folia Med Cracov. 2023;63(1):121–34. 10.24425/fmc.2023.145435.
24. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical final examination. Sci Rep. 2023;13(1):20512. 10.1038/s41598-023-46995-z.
25. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination? A descriptive study. J Educ Eval Health Prof. 2023;20:1. 10.3352/jeehp.2023.20.1.
26. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. 10.1136/bmj.n71.
27. Aljindan FK, Al Qurashi AA, Albalawi IAS, Alanazi AMM, Aljuhani HAM, Falah Almutairi F, et al. ChatGPT conquers the Saudi medical licensing exam: exploring the accuracy of artificial intelligence in medical knowledge assessment and implications for modern medical education. Cureus. 2023;15(9):e45043. 10.7759/cureus.45043.
28. Angel M, Patel A, Alachkar A, Baldi B. Clinical knowledge and reasoning abilities of AI large language models in pharmacy: a comparative study on the NAPLEX exam. bioRxiv 2023.06.07.544055. 10.1101/2023.06.07.544055
29. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-english national medical licensing examination? An evaluation in Chinese language. PLOS Digit Health. 2023;2(12):e0000397. 10.1371/journal.pdig.0000397.
30. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, De la Cruz-Galán JP, Gutiérrez-Arratia JD, Quiroga Torres BG, et al. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. JMIR Med Educ. 2023;9:e48039. 10.2196/48039.
31. Fuchs A, Trachsel T, Weiger R, Eggmann F. ChatGPT's performance in dentistry and allergy-immunology assessments: a comparative study. Swiss Dent J. 2023;134(5). Epub ahead of print.
32. Huang H. Performance of ChatGPT on registered nurse license exam in Taiwan: a descriptive study. Healthc (Basel). 2023;11(21):2855. 10.3390/healthcare11212855.
33. Kataoka Y, Yamamoto-Kataoka S, So R, Furukawa TA. Beyond the pass mark: accuracy of ChatGPT

and Bing in the national medical licensure examination in Japan. JMA J. 2023;6(4):536–8. 10.31662/jmaj.2023-0043.

34. Kleinig O, Gao C, Bacchi S. This too shall pass: the performance of ChatGPT-3.5, ChatGPT-4 and New Bing in an Australian medical licensing examination. Med J Aust. 2023;219(5):237. 10.5694/mja2.52061.
35. Kunitsu Y. The potential of GPT-4 as a support tool for pharmacists: analytical study using the Japanese national examination for pharmacists. JMIR Med Educ. 2023;9:e48452. 10.2196/48452.
36. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom medical licensing assessment. Front Med (Lausanne). 2023;10:1240915. 10.3389/fmed.2023.1240915.
37. Mihalache A, Huang RS, Popovic MM, Muni RH. ChatGPT-4: an assessment of an upgraded artificial intelligence chatbot in the United States medical licensing examination. Med Teach. 2024;46(3):366–72. 10.1080/0142159X.2023.2249588.
38. Morreel S, Verhoeven V, Mathysen D. Microsoft Bing outperforms five other generative artificial intelligence chatbots in the Antwerp University multiple choice medical license exam. PLOS Digit Health. 2024;3(2):e0000349. 10.1371/journal.pdig.0000349.
39. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse examinations in Japan: evaluation study. JMIR Nurs. 2023;6:e47305. 10.2196/47305.
40. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. 2023;9:e48002. 10.2196/48002.
41. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of generative pretrained transformer on the national medical licensing examination in Japan. PLOS Digit Health. 2024;3(1):e0000433. 10.1371/journal.pdig.0000433.
42. Tong W, Guan Y, Chen J, Huang X, Zhong Y, Zhang C, et al. Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT, in the Chinese national medical licensing examination. Front Med (Lausanne). 2023;10:1237432. 10.3389/fmed.2023.1237432.
43. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inf. 2023;177:105173. 10.1016/j.ijmedinf.2023.105173.
44. Wang X, Gong Z, Wang G, Jia J, Xu Y, Zhao J, et al. ChatGPT performs on the Chinese national medical licensing examination. J Med Syst. 2023;47(1):86. 10.1007/s10916-023-01961-0.
45. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. JMIR Form Res. 2023;7:e48023. 10.2196/48023.
46. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. J Med Educ Curric Dev. 2024;11:1–12. 10.1177/23821205241238641.
47. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024;1–10. 10.1111/medu.15402.
48. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, et al. Chapter 7: systematic reviews of etiology and risk. In: Aromataris E, Munn Z, editors. Editors). JBI Manual for evidence synthesis. JBI; 2020. https://jbi.global/critical-appraisal-tools.
49. Becker BJ. Multivariate meta-analysis. In: Tinsley HEA, Brown SD, editors. Handbook of applied multivariate statistics and mathematical modeling. San Diego: Academic; 2000. pp. 499–525.
50. Cooper. Synthesizing research: a guide for literature reviews. 3rd ed. Thousand Oaks, CA: Sage; 1998.
51. Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull. 1979;86(3):638–41. 10.1037/0033-2909.86.3.638.
52. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589–97. 10.1001/jamaophthalmol.2023.1144.
53. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first-year plastic surgery residents: evaluation of ChatGPT on the plastic

surgery in-service examination. Aesthet Surg J. 2023;43(12):NP1085–9. 10.1093/asj/sjad130.

54. Hopkins BS, Nguyen VN, Dallas J, Texakalidis P, Yang M, Renn A, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. J Neurosurg. 2023;139(3):904–11. 10.3171/2023.2.JNS23419.
55. Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American College of Gastroenterology self-assessment test. Am J Gastroenterol. 2023;118:2280–2. 10.14309/ajg.0000000000002320.
56. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure wisdom or potemkin villages? A comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE step 3 style questions: quantitative analysis. JMIR Med Educ. 2024;10:e51148. 10.2196/51148.
57. OpenAI. GPT-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf. Accessed 10 2024.
58. Seghier ML. ChatGPT: not all languages are equal. Nature. 2023;615(7951):216. 10.1038/d41586-023-00680-3.
59. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. JAMA Health Forum. 2023;4(5):e231938. 10.1001/jamahealthforum.2023.1938.