

RESEARCH

Open Access



Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis

Hye Kyung Jin^{1,2}, Ha Eun Lee² and EunYoung Kim^{1,2,3*}

Abstract

Background ChatGPT, a recently developed artificial intelligence (AI) chatbot, has demonstrated improved performance in examinations in the medical field. However, thus far, an overall evaluation of the potential of ChatGPT models (ChatGPT-3.5 and GPT-4) in a variety of national health licensing examinations is lacking. This study aimed to provide a comprehensive assessment of the ChatGPT models' performance in national licensing examinations for medical, pharmacy, dentistry, and nursing research through a meta-analysis.

Methods Following the PRISMA protocol, full-text articles from MEDLINE/PubMed, EMBASE, ERIC, Cochrane Library, Web of Science, and key journals were reviewed from the time of ChatGPT's introduction to February 27, 2024. Studies were eligible if they evaluated the performance of a ChatGPT model (ChatGPT-3.5 or GPT-4); related to national licensing examinations in the fields of medicine, pharmacy, dentistry, or nursing; involved multiple-choice questions; and provided data that enabled the calculation of effect size. Two reviewers independently completed data extraction, coding, and quality assessment. The JBI Critical Appraisal Tools were used to assess the quality of the selected articles. Overall effect size and 95% confidence intervals [CIs] were calculated using a random-effects model.

Results A total of 23 studies were considered for this review, which evaluated the accuracy of four types of national licensing examinations. The selected articles were in the fields of medicine ($n = 17$), pharmacy ($n = 3$), nursing ($n = 2$), and dentistry ($n = 1$). They reported varying accuracy levels, ranging from 36 to 77% for ChatGPT-3.5 and 64.4–100% for GPT-4. The overall effect size for the percentage of accuracy was 70.1% (95% CI, 65–74.8%), which was statistically significant ($p < 0.001$). Subgroup analyses revealed that GPT-4 demonstrated significantly higher accuracy in providing correct responses than its earlier version, ChatGPT-3.5. Additionally, in the context of health licensing examinations, the ChatGPT models exhibited greater proficiency in the following order: pharmacy, medicine, dentistry, and nursing.

*Correspondence:
EunYoung Kim
eykimjcb777@cau.ac.kr

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

However, the lack of a broader set of questions, including open-ended and scenario-based questions, and significant heterogeneity were limitations of this meta-analysis.

Conclusions This study sheds light on the accuracy of ChatGPT models in four national health licensing examinations across various countries and provides a practical basis and theoretical support for future research. Further studies are needed to explore their utilization in medical and health education by including a broader and more diverse range of questions, along with more advanced versions of AI chatbots.

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

Background

Over the past decade, artificial intelligence (AI) technology has experienced rapid evolution, leading to significant advancements across various fields [1, 2]. One of the most recent and notable advancements is ChatGPT by OpenAI (San Francisco, CA), a natural language processing program designed to generate human-like language [3]. Since its launch, this innovative technology has demonstrated applicability in a variety of domains, including healthcare, education, research, business, and industry [4–6]. Moreover, ChatGPT is a robust, evolving AI chatbot with considerable potential as a support resource for healthcare professionals, educators, and learners. For instance, in the realm of education, cutting-edge ChatGPT versions can aid students through personalized learning experiences and tutoring [7]. In the healthcare domain, it can assist medical professionals in diagnoses, treatment plans, and patient education by integrating medical knowledge with interactive dialogue [1, 8–10].

With ongoing advancements in this novel technology, the accuracy of ChatGPT is expected to improve, thereby expanding their applicability, specifically in healthcare education. In a study by Gilson et al. emphasized that such technology could achieve equivalence with third-year US medical students, highlighting its potential to provide logical explanations and informative feedback [7]. Additionally, in the field of medical research, ChatGPT's efficacy has been proven in the United States Medical Licensing Examination (USMLE), in which it achieved accuracy rates of 40–60% [7, 11]. The latest ChatGPT version, GPT-4, outperforms its predecessor, ChatGPT-3.5, in reliability and accuracy, delivering expert-level quality responses [12]. Furthermore, Yang et al. found that ChatGPT-4's ability to respond to USMLE questions involving images resulted in a 90.7% accuracy rate across the entire USMLE examination, exceeding the passing threshold of approximately 60% accuracy [13]. Remarkably, GPT-4 also exhibited the capacity to address intricate interpersonal, ethical, and professional requirements within medical practice [14].

However, despite its capacity to significantly improve the efficiency of education, efforts to adopt it in healthcare professional education remain limited [15]. Specifically, this AI chatbot can produce inaccurate outputs

[16], and the reported response accuracy rates vary among different medical examinations and medical fields [17, 18]. Other studies have also shown that calculations are identified as a domain where large language models (LLMs) tend to exhibit comparatively lower precision [19] and fail to achieve the required accuracy level, as demonstrated in the National Medical Licensing Examination (NMLE), National Pharmacist Licensing Examination (NPLE), and National Nurse Licensing Examination (NNLE) in non-English language regions [20, 21]. Potential reasons for the performance differences have been proposed, including variations in curricula and examination content [22, 23]. Recent studies have shown that while ChatGPT's accuracy has improved over time, its scores are still lower than those of medical students [24, 25]. Consequently, existing studies are inconsistent and do not yield definitive conclusions.

Given these considerations, a meta-analysis is needed to offer a more comprehensive summary of the evidence regarding the performance of OpenAI's ChatGPT series in four types of national healthcare-related licensing examinations conducted in a range of countries. Therefore, this study aimed to conduct a meta-analysis of studies reporting the performance of both versions of ChatGPT (ChatGPT-3.5 and GPT-4) across different healthcare-related national licensing examinations without restriction to specific countries. The specific research questions were as follows: (1) What is the overall effect size of ChatGPT models in the national licensing examinations for medicine, pharmacy, dentistry, and nursing? (2) Which ChatGPT model has the greatest influence on the effect sizes measured in national licensing examinations? (3) Which medical field examinations (e.g., medicine, pharmacy, dentistry, and nursing) have the greatest influence on the effect sizes measured in ChatGPT model research?

Methods

The reporting of this review and meta-analysis is based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [26]. See the completed PRISMA 2020 for abstracts (Supplementary Table 1) and the PRISMA 2020 checklists (Supplementary Table 2).

Data sources

The protocol for the review was not registered. Relevant articles were identified by searching the following databases: MEDLINE/PubMed, EMBASE, Cochrane Library, ERIC, and Web of Science. To identify further relevant studies, hand searches of key issues of journals related to healthcare education, as well as the reference lists of the retrieved studies [11, 13, 14, 21, 27–45] and previously conducted systematic reviews [17, 46, 47] were carried out. The retrieved records were imported into an End-Note™ library. The search criteria were restricted to articles published in English, with a publication date ranging from November 30, 2022, the release date of OpenAI, to February 27, 2024.

Search strategy

This study employed a structured search strategy that involved the use of specific search terms associated with ChatGPT (e.g., “Artificial intelligence” [MeSH Terms], “ChatGPT” [Title/Abstract], “GPT-4” [Title/Abstract], “Chatbot” [Title/Abstract], “Natural language processing” [Title/Abstract], “Large language models” [Title/Abstract]). These terms were combined with terms related to the licensing examination (e.g., “educational measurement” [MeSH Terms], “licensure, medical” [MeSH Terms], “licensure, pharmacy” [MeSH Terms], “licensure, dental” [MeSH Terms], “licensure, nursing” [MeSH Terms], “medical licensure exam*” [Title/Abstract]). Variations of this search strategy were applied across various sources. Controlled vocabulary such as MeSH (Medical Subject Headings) for MEDLINE (PubMed) and thesaurus terms from ERIC were utilized alongside advanced search techniques such as truncation for broader retrieval. The first author (HK) conducted the search, and the full search strategies are shown in Supplementary Table 3.

Inclusion and exclusion criteria

Two reviewers (HK and HE) independently reviewed the eligibility of studies for inclusion and exclusion. Disagreements between the reviewers were resolved by consensus or discussion with a third reviewer (EY). A study was considered for inclusion in the meta-analysis only if it met all of the following four criteria: (1) focused on evaluating the performance of a ChatGPT model (ChatGPT-3.5 or GPT-4); (2) related to national licensing examinations in the fields of medicine, pharmacy, dentistry, or nursing; (3) involved multiple-choice questions; and (4) provided data that enabled the calculation of effect size. Studies were excluded if they (1) were review articles or meta-analyses; (2) used any AI platform other than ChatGPT-3.5 or GPT-4; (3) were unrelated to national licensing exams; (4) lacked sufficient data for effect size calculation; (5) did not use multiple-choice

questions (e.g., used open-ended questions); (6) were unavailable in full text; and (7) were not in the English language.

Outcome

The outcome of this study was the accuracy of ChatGPT (GPT-3.5 and GPT-4) on the national licensing examinations for medicine, pharmacy, dentistry, and nursing.

Risk of bias in studies

The Joanna Briggs Institute (JBI) critical appraisal tools for analytical cross-sectional studies were used to assess the risk of bias [48]. This checklist consists of eight questions and each item is scored as 1 (yes) or 0 (no, unclear or not applicable), resulting in a total score of 8 if all questions are positively answered. Potential disagreements between the two reviewers were resolved by discussions. The risk of bias of the study was judged as follows: High risk if two or more domains were considered at high risk; moderate risk if one domain was considered at high risk or if two or more domains were considered unclear; low risk if no domains were considered at high risk.

Data extraction and coding

Data extraction was conducted independently by two reviewers (HK and HE) using a dual-review process, and any discrepancies in the data were resolved by consensus or adjudication by a third reviewer (EY). The data extraction form was designed to collect relevant information from studies, including descriptive details, such as the first author, publication year, journal, and country. Additionally, the specific characteristics of the studies included the ChatGPT models used, field of study, number of questions, accuracy rates, and key results observed.

Each of the 23 included studies was coded by the first author (HK), who subsequently trained other coders on the procedures, allowing them to execute the coding task independently.

Dependence

Independence violations occur when a single study yields multiple effect sizes [49]. This meta-analysis included 144 effect sizes from 23 studies and used multivariate data. Such data typically emerge when a study produces several effect sizes. To maintain the assumption of independence and to avoid information loss, this study adopted a “shifting unit of analysis” approach [50]. Specifically, in this meta-analysis, the total effect size was calculated using the study as the unit of analysis; however, for the subgroup analysis, individual effect sizes were used as the units to estimate the effect sizes.

Statistical analysis

A random-effects model was used to calculate the overall accuracy rate and all testing of moderators owing to the high level of heterogeneity among the included studies. All data analyses were conducted using the Comprehensive Meta-Analysis version 4 (Biostat, Englewood, NJ, USA). The number of correctly answered questions and the total number of questions were used to calculate the effect sizes in each study as proportions. Forest plot was employed to visually illustrate each study's contribution to this meta-analysis. The heterogeneity among the studies was measured using Q statistics, and the I^2 statistics illustrated the degree of heterogeneity. A sensitivity analysis was conducted to assess the robustness of the findings by excluding studies that were judged to be at high risk of bias. Additionally, subgroup analyses was performed to explore potential reasons for variance with the moderator variables (ChatGPT models and fields of health licensing examinations). Meta-regression was not performed in this study because there were no appropriate continuous variables available to assess the trends provided by the relevant variables through meta-regression analysis. Funnel plots, Egger's regression test, the Duval and Tweedie trim-and-fill methods, and classic fail-safe N test were used to evaluate publication bias. All analyses performed in this study were at 95% CIs.

Ethical considerations

This meta-analysis was conducted using previously published studies. Therefore, evaluation by the institutional review board was not considered necessary.

Results

Study selection

The literature search yielded 433 articles after the initial search. After removing duplicates and performing title/abstract screening, a total of 31 articles were included for full-text review. Of these, one was excluded as it was reported in a language other than English, three for non-multiple-choice questions, and eight for insufficient data for calculating effect size. After assessing the articles for eligibility, 23 articles were included in this review (Fig. 1).

Risk of bias assessment

The findings of the risk of bias assessment using The Joanna Briggs Institute (JBI) Critical Appraisal Checklist for analytical cross-sectional studies are presented in Table 1. Among the studies, two were deemed to be at high risk [33, 34], and 18 at low risk [11, 13, 21, 27, 29–31, 35–45] with the most commonly observed weaknesses relating to the identification of confounding factors as well as the strategies to deal with them. The risk of bias for each of the other studies was considered to be moderate [14, 28, 32].

Study characteristics

The characteristics of the 23 included studies [11, 13, 14, 21, 27–45] are shown in Table 2. These studies were all published between 2022 and 2024. Five of the included studies were conducted in the United States [11, 13, 14, 28, 37], one in the UK [36], six in Japan [33, 35, 39–41, 45], four in China [29, 42–44], two in Taiwan [21, 32], and one each in Australia [34], Belgium [38], Peru [30], Switzerland [31], and Saudi Arabia [27]. In terms of field, 17 articles focused on medicine [11, 13, 14, 27, 29, 30, 33, 34, 36–38, 40–45], three on pharmacy [21, 28, 35], two on nursing [32, 39], and one on dentistry [31]. The range of questions per examination varied from 21 questions related to communication skills, ethics, empathy, and professionalism in the USMLE [14] to 1510 questions across four separate exams in 2022 and 2023 in the Registered Nurse License Exam [32].

The performance of ChatGPT-3.5 across these studies fluctuated drastically, with scores as low as 38% in the Japanese National Medical Licensure Examination [33] to as high as 73% in the Peruvian National Licensing Medical Examination [30]. Similarly, GPT-4's performance varied but generally demonstrated superior accuracy, achieving from 64.4% in the Swiss Federal Licensing Examination in Dental Medicine [31] to 100% in USMLE questions related to soft skills assessments [14]. This consistent improvement across different exams and regions indicates GPT-4's enhanced capability to handle diverse medical licensing examinations with higher accuracy compared to its predecessor.

Publication bias

A funnel plot of the included studies revealed somewhat asymmetry (Fig. 2). Egger's regression test resulted in a significant p -value ($p=0.028$), suggesting the presence of publication bias in the meta-analysis. Conversely, the trim-and-fill analysis revealed that no studies were omitted or required trimming, indicating that the overall effect size remained unchanged. Additionally, Rosenthal's fail-safe number, calculated in CMA as 310, suggests the absence of publication bias when compared to the Eq. $5n+10$ [51], where n represents the number of studies included in the meta-analysis. Taken together, we believe that publication bias does not appear likely in the study.

Overall analysis

A random-effects model was selected based on the assumption that effect sizes might vary among studies due to variations in examinations. Consequently, the effect sizes in the primary studies were also heterogeneous ($Q=1201.303$, $df=22$, $p<0.001$), with an I^2 of 98.2%. The overall effect size for the percentage of performance of ChatGPT models was 70.1% (95% CI,

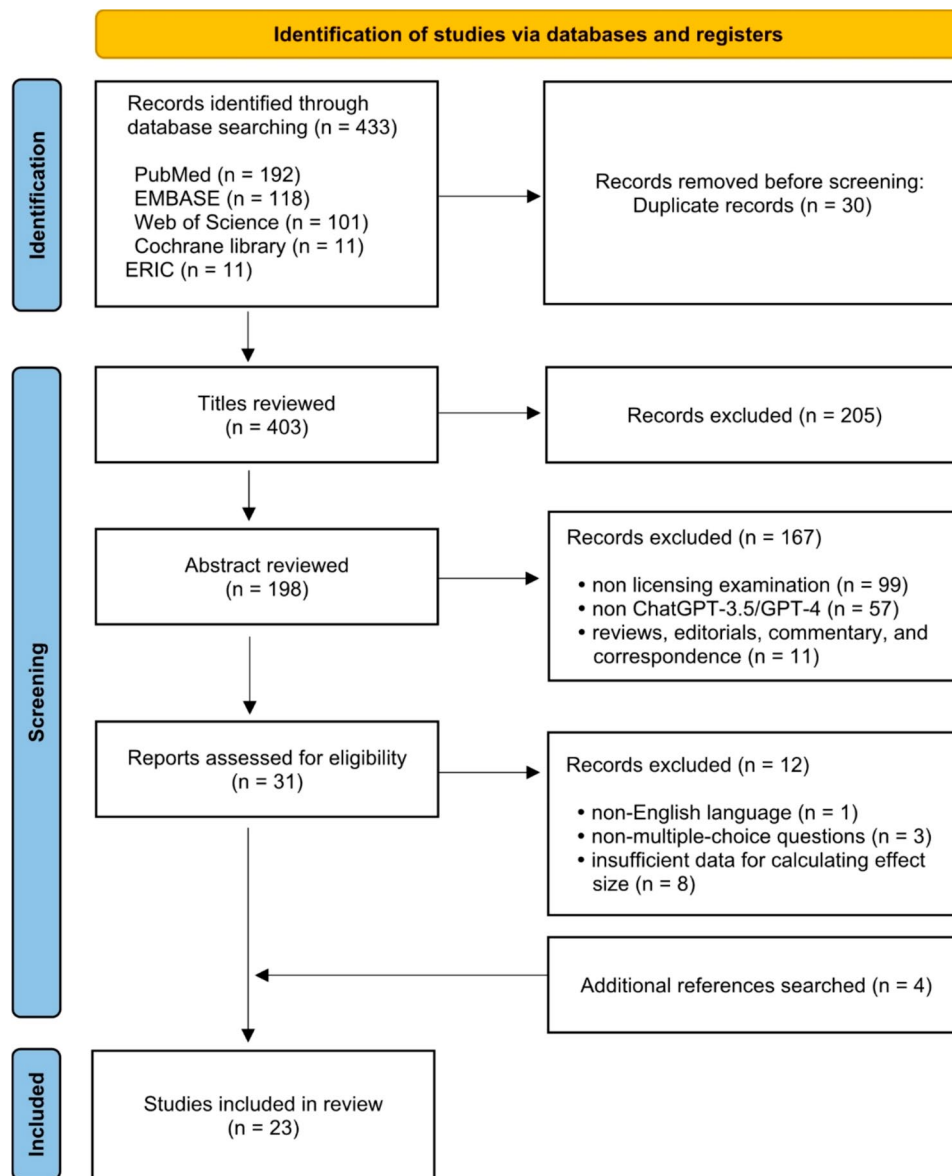


Fig. 1 PRISMA flowchart of included studies

65.0-74.8%). The forest plot of individual and overall effect sizes is shown in Fig. 3.

Subgroup analyses

Under a random-effects model, subgroup analyses were conducted with the two categorical variables: (1) the type of ChatGPT model (ChatGPT-3.5 vs. GPT-4); and (2) the field of health licensing examinations (medicine, pharmacy, dentistry, and nursing), as detailed in Table 3. Other potential moderating factors were limited, as they were either too infrequently reported or inadequately described to facilitate a comprehensive subgroup analysis.

ChatGPT models

The effect size for the percentage of performance with regard to the ChatGPT models was 58.9% (95% CI, 56.2-61.6%) for ChatGPT-3.5 and 80.4% (95% CI, 78.6-82.0%) for GPT-4. The test results demonstrated a significant difference between the effect sizes of the studies ($Q=177.027, p<0.001$).

Fields of health licensing examinations

The effect size for percentage of performance was at an average rate of 71.5% (95% CI, 66.3-76.2%) in pharmacy, 69.7% (95% CI, 65.9-73.2%) in medicine, 63.2% (95% CI, 54.6-71.1%) in dentistry, and 61.8% (95% CI, 58.7-64.9%) in nursing. This indicated a significant difference between the fields ($Q=15.334, p=0.002$).

Table 1 Quality assessment of the included studies

Study ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	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	1	1	1	0	1	1	7	Moderate
Brin et al. [14]	1	1	1	1	1	0	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	1	1	1	0	1	1	7	Moderate
Kataoka et al. [33]	1	1	1	1	0	0	1	1	6	High
Kleinig et al. [34]	1	1	1	1	0	0	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

Joanna Briggs Institute (JBI) critical appraisal tools for analytical cross-sectional studies

^aThe range of overall quality scores is 0–8

Q1. Were the criteria for inclusion in the sample clearly defined?; Q2. Were the study subjects and the setting described in detail?; Q3. Was the exposure measured in a valid and reliable way?; Q4. Were objective, standard criteria used for measurement of the condition?; Q5. Were confounding factors identified?; Q6. Were strategies to deal with confounding factors stated?; Q7. Were the outcomes measured in a valid and reliable way?; Q8. Was appropriate statistical analysis used?

Sensitivity analysis

The sensitivity analysis demonstrated that removing high-risk studies did not significantly alter the overall accuracy, which remained at 71.3% (95% CI, 66.3–75.8%), indicating that the results were robust ($p < 0.001$), with an I^2 of 98.1% (Fig. 4).

Discussion

Overall, though exhibiting varying levels of risk of bias, the majority of the included studies were considered to be of high quality. The largest threats were related to the lack of identification of confounding factors and the strategies to address them. To the best of our knowledge, this is the first meta-analysis focusing on the performance of ChatGPT technology in four types of health licensing examinations, specifically comparing GPT-3.5 and GPT-4. Most previous studies have either focused solely on academic testing or knowledge assessment rather than national licensing exams [25, 52–54] or have combined these contexts without distinction [17, 46, 47]. Additionally, they have primarily reported the accuracy rate of a single version of ChatGPT instead of comparing different versions [17, 29, 32, 37, 39, 44].

The findings clearly indicate that GPT-4's accuracy rates were significantly higher than those of ChatGPT-3.5 across various national medical licensing exams conducted in different countries. Furthermore, these results demonstrate that GPT-4 not only surpasses GPT-3.5 in overall accuracy but also shows more reliable performance, making it a more advanced tool for such assessments. A previous study suggested that achieving an accuracy rate exceeding 95% could make ChatGPT a reliable education tool [55]. Although it remains uncertain whether future models will attain this level of proficiency, the rapid development of these LLM technologies, owing to user feedback and deep learning, will lead to their potential utilization in learning and education in the medical field.

Regarding the types of ChatGPT models, this study provides insights into the potential of GPT-4, the latest version, particularly within the scope of its training dataset. Our finding is similar to what a previous study showed regarding USMLE questions, where accuracy rates of 84.7% were observed for GPT-4 and 56.9% for its earlier version [56]. Another study also reported that GPT-4 had an accuracy of 86.1% and ChatGPT-3.5 had

Table 2 Characteristics of the included studies

No.	Study ID	Country	Objectives related to ChatGPT models	Field	No. of questions	Accuracy
1	Aljindan et al. [27]	Saudi Arabia	Assessment of proficiency of GPT-4 in answering Saudi Medical Licensing Exam questions	Medicine	220	GPT-4: Overall accuracy 88.6%. Accuracy varied by question difficulty and response option. 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	Angel et al. [28]	USA	Assessment of the capabilities and limitations of GPT-3, GPT-4, and Bard using a sample North American Pharmacist Licensure Examination	Pharmacy	137	GPT-4: 78.8%, 27.7% improvement over GPT-3, surpassing the minimum passing score of 75
3	Brin et al. [14]	USA	Evaluation of ChatGPT and GPT-4's performance in the United States Medical Licensing Exam (USMLE) questions related to communication skills, ethics, empathy, and professionalism	Medicine	21	GPT-4: 100% accuracy, ChatGPT: 66.6% accuracy
4	Fang et al. [29]	China	Assessment of ChatGPT's performance on the Chinese National Medical Licensing Examination (NMLE)	Medicine	600	ChatGPT: Overall 73.7% (442/600), surpassing the passing threshold of 360/600. Performance varied across four units, highest in Unit 4: 76.0%, lowest in Unit 2: 70.0%
5	Flores-Cohaila et al. [30]	Peru	Assessment the accuracy of ChatGPT using GPT-3.5 and GPT-4 on the Peruvian National Licensing Medical Examination (ENAM)	Medicine	180	GPT-4: 86% (155/180), GPT-3.5: 77% (139/180) for multiple-choice questions with justification and 73% (133/180) for those without justification; passing score on the ENAM 10.5 on a vigesimal scale: 95/180
6	Fuchs et al. [31]	Switzerland	Evaluating the performance of ChatGPT and GPT-4 on self-assessment dentistry questions via the Swiss Federal Licensing Examination in Dental Medicine (SFLEDM)	Dentistry	32	GPT-4: 64.4% (without priming), 66.7% (with priming), ChatGPT: 59.3% (without priming), 62.6% (with priming); commonly used passing threshold in the SFLEDM assessment: 60%
7	Huang [32]	Taiwan	Assessment of ChatGPT's performance on the Registered Nurse License Exam (RNLE), and evaluation of the advantages and disadvantages of using ChatGPT	Nursing	1510 (4 exams: 400 for 2022 1st, 370 each for 2022 2nd, 2023 1st, and 2023 2nd)	ChatGPT: 63.75% (2022 1st exam), 58.37% (2022 2nd exam), 54.05% (2023 1st exam), 60% (2023 2nd exam); passing average score of 60 across five subjects, each out of 100
8	Kataoka et al. [33]	Japan	Assessment of the accuracy of ChatGPT and Bing based on their responses to the National Medical Licensure Examination in Japan	Medicine	281	ChatGPT: Overall 38% (106/281), compulsory questions: 42% (34/81), other questions: 36% (72/200). Bing: 78% (219/281)
9	Kleinig et al. [34]	Australia	Assessment of ChatGPT-3.5, ChatGPT-4, and New Bing's performance using the publicly accessible Australian Medical Council Licensing Examination practice questions	Medicine	50	ChatGPT-3.5: 66% (99/150), ChatGPT-4: 79.3% (119/150)
10	Kung et al. [11]	USA	Assessment of ChatGPT's performance on the United States Medical Licensing Exam (USMLE), which consists of three exams: Step 1, Step 2CK, and Step 3	Medicine	350	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)

Table 2 (continued)

No.	Study ID	Country	Objectives related to ChatGPT models	Field	No. of questions	Accuracy
11	Kunitsu [35]	Japan	Assessment of GPT-4's ability to answer questions from the Japanese National Examination for Pharmacists (JNEP)	Pharmacy	107th and 108th JNEP exams, with 345 questions across three blocks; GPT-4 responded to 284 and 287 questions, respectively	GPT-4: 107th exam: Overall 64.5% (222/344), 108th exam: Overall 62.9% (217/345). Improved to 78.2% (222/284) and 75.3% (217/288), respectively, when only considering questions it could answer. Lower accuracy was observed in physics, chemistry, and calculation questions
12	Lai et al. [36]	UK	Assessment the performance of ChatGPT in the United Kingdom Medical Licensing Assessment	Medicine	191	GPT-4: Average score of 76.3% (437/573) across these attempts. Attempt 1: 74.9% (143/191), Attempt 2: 78.0% (149/191), Attempt 3: 75.6% (145/191)
13	Mihalache et al. [37]	USA	Assessment of ChatGPT-4's performance in responding to USMLE Step 1, Step 2CK, and Step 3 practice questions	Medicine	319	GPT-4: 88% on USMLE Step 1 (82/93), 86% on Step 2CK (91/106), and 90% on Step 3 (108/120)
14	Morreel et al. [38]	Belgium	Assessment of the accuracy of ChatGPT using GPT-3.5 and GPT-4 on the University of Antwerp Medical Licensing Exam	Medicine	95	GPT-4: 76% (72/95), ChatGPT: 67% (64/95)
15	Taira et al. [39]	Japan	Assessment of the performance of ChatGPT on the Japanese National Nurse Examinations from 2019 to 2023	Nursing	240 questions each year, categorized into two types: basic knowledge questions and general questions	ChatGPT: Average accuracy over five years: 75.1% for basic knowledge questions, 64.5% for general questions; passing criteria: 80% for basic knowledge questions, approximately 60% for general
16	Takagi et al. [40]	Japan	Comparison of the performances of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination	Medicine	254	GPT-3.5: 50.8%, GPT-4: 79.9%
17	Tanaka et al. [41]	Japan	Evaluating the performance of ChatGPT models (GPT-3.5 and GPT-4) on the NMLE and comparing it to the exam's minimum passing rate	Medicine	116th: 290 117th: 262	116th exam: GPT-3.5: 63.1% (183/290), GPT-4: 82.8% (240/290), 117th exam: GPT-4: 78.6% (206/262) with a tuned prompt
18	Tong et al. [42]	China	Evaluating the applicability and limitations of ChatGPT(4.0) in the context of taking the Chinese Medical Licensing Examination in both English and Chinese versions	Medicine	160	GPT-4: 81.25% for the Chinese versions, 86.25% for the English versions
19	Wang et al. [43]	China	Comparison of ChatGPT's (GPT-3.5 and GPT-4) performance in the China National Medical Licensing Examination (CNMLE) and its English translated version (ENMLE)	Medicine	200 (100 in CNMLE, 100 in ENMLE)	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	Wang et al. [44]	China	Comparison the performance of medical students and ChatGPT in the Chinese NMLE	Medicine	600 (4 units with 150 per unit)	ChatGPT: 47.0% (282/600) in 2020, 45.8% (275/600) in 2021, 36.5% (219/600) in 2022. ChatGPT's accuracy was lower than that of medical students in the Chinese NMLE
21	Wang et al. [21]	Taiwan	Evaluating ChatGPT's accuracy on the Taiwanese Pharmacist Licensing Examination and exploring its potential in pharmacy education	Pharmacy	826 (203 in Stage1 and 210 in Stage2 for Chinese and English, respectively)	ChatGPT: Stage 1: 54.4% for Chinese, 56.9% for English. Stage 2: 53.8% for Chinese, 67.6% for English

Table 2 (continued)

No.	Study ID	Country	Objectives related to ChatGPT models	Field	No. of questions	Accuracy
22	Yanagi-ta et al. [45]	Japan	Assessment of correct answers output by GPT-3.5 and GPT-4 from the NMLE in Japan	Medicine	Out of 400 questions, 292 excluding those with charts unsupported by ChatGPT	GPT-4: 81.5% (237/292), surpassing the passing threshold of 72%. GPT-3.5: 42.8% (125/292)
23	Yang et al. [13]	USA	Comparison of the performance of GPT-4 V, GPT-4, and ChatGPT on medical licensing examination questions	Medicine	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)	GPT-4 V: 88.2–92.7%, GPT-4: 80.8–88.3%, ChatGPT: 55.1–60.9%. GPT-4 V 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%

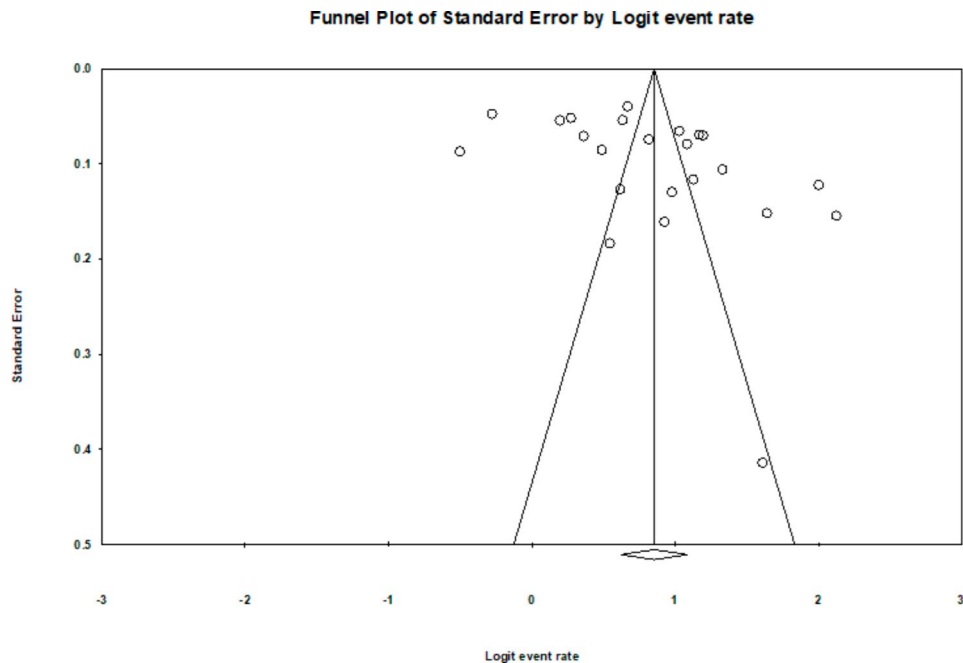
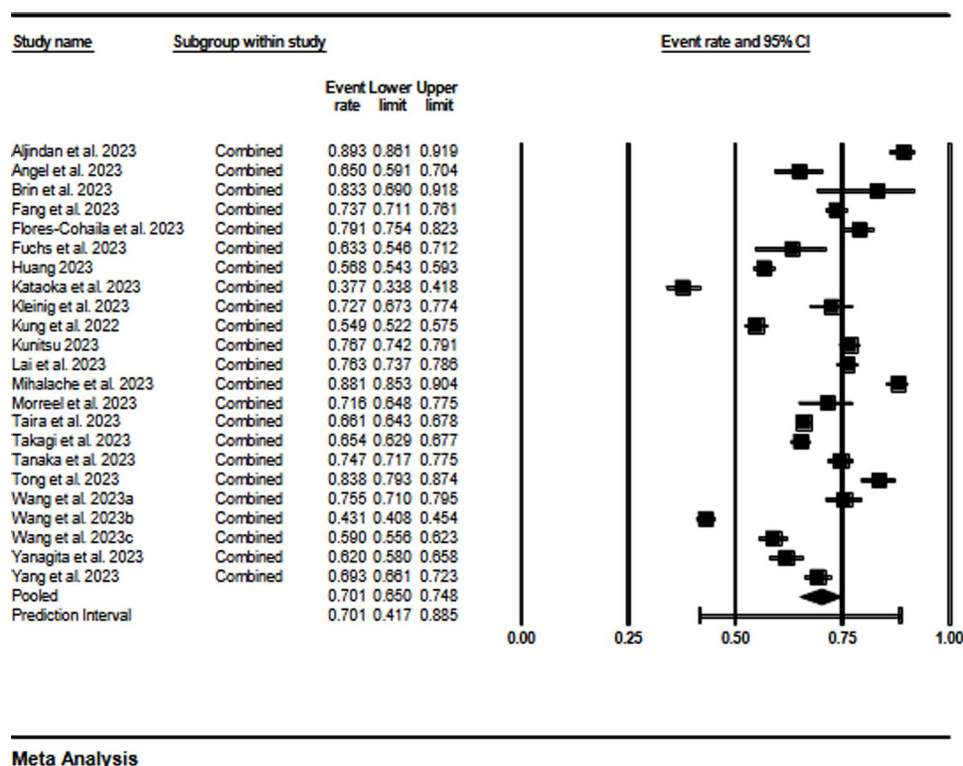


Fig. 2 Funnel plot for studies included in the meta-analysis

an accuracy of 77.2% in the Peruvian National Licensing Medical Examination [30]. All of this suggests that GPT-4’s outperformance is perhaps attributable to its improved reasoning capabilities or critical thinking [57].

The current study revealed that in the context of health licensing examinations, ChatGPT models exhibited the highest proficiency for pharmacy, followed by medicine, dentistry, and nursing. Our results were inconsistent with previous research indicating that ChatGPT

exhibited greater proficiency in NNLE, with NMLE and NPLE following behind in China, which is likely associated with the difference in GPT model versions as well as the difficulty and complexity of the test questions [20]. Moreover, variations in language or culture can lead to performance disparities. Similarly, a study by Seghier also highlighted that these innovative models encountered linguistic challenges and emphasized that their performance in non-English responses was significantly lower



Meta Analysis

Fig. 3 Forest plot of effect sizes using random effect model

Table 3 Subgroup analyses

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

[58]. Consequently, the present study recommends incorporating additional training data in languages other than English to improve performance, enabling ChatGPT to be an effective educational tool as an AI assistant in broader educational contexts, for both students and professionals. Along with this, while this meta-analysis includes studies from various educational fields, it is important to address the scarcity of research specifically in the field of dentistry, with only one study available. Further research in this area is necessary to ensure that our final findings are robust and generalizable across different educational settings.

In addition to the suggestions discussed above, it must be noted that the rapid evolution of medical fields necessitates high quality, up-to-date data to enhance performance. For instance, considering the annual introduction of new and updated medications and medical guidelines, ChatGPT might not always have access to the most

recent information due to its available knowledge being limited to 2021. Furthermore, previous studies have reported that ChatGPT might provide incorrect information or plausible but wrong answers, commonly referred to as “hallucinations” [16, 32, 38, 43]. The models used in this study, ChatGPT-3.5 and GPT-4, represent earlier and more recent versions, respectively, and as AI technology rapidly evolves, these versions may soon become obsolete and appear limited. Previous research has suggested that many limitations identified in ChatGPT-3.5, including hallucinations, still apply to GPT-4 [57]. Therefore, it is imperative for future versions to recognize both the limitations and possibilities of these generative AI chatbots, and to ensure continuous refinement and training, in order to achieve a more reliable resource and practical application.

Based on the results of this study, certain limitations that could affect the interpretation of our findings should

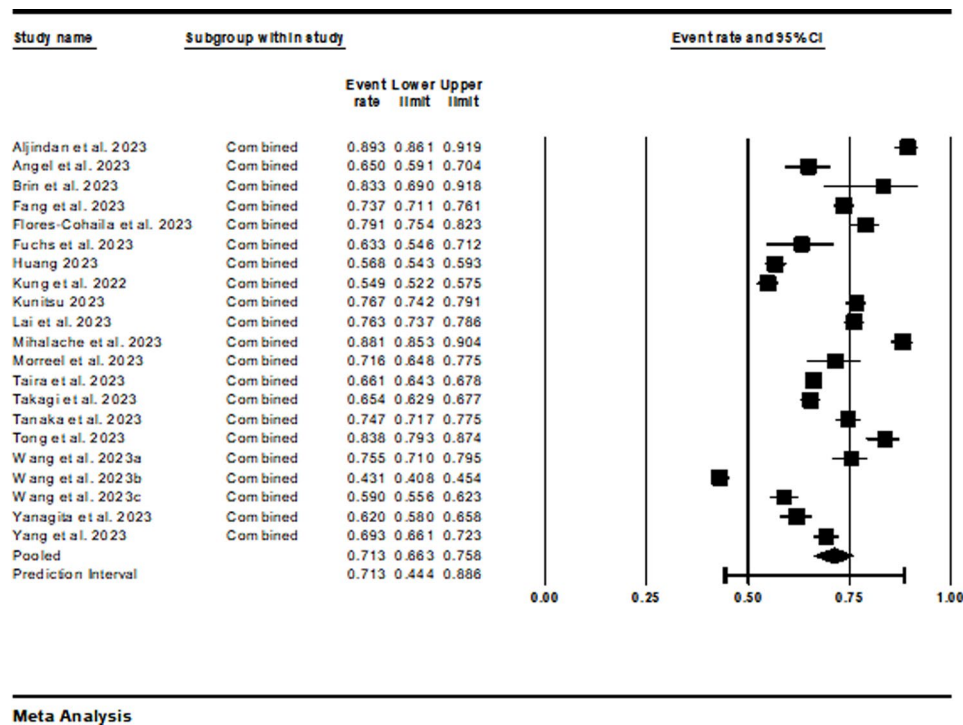


Fig. 4 Sensitivity analysis omitting high risk studies

be acknowledged. First, while this systematic review employed rigorous methods and has been reported in accordance with PRISMA, it is acknowledged that the protocol was not registered. Second, the variability in the inclusion criteria and the design of the studies across different countries poses a significant limitation. Health licensing examinations’ structures, difficulty levels, and content vary by country, which might influence the performance outcomes of ChatGPT models. Therefore, this variability is possible to make it challenging to accurately assess the effectiveness of the AI models. Third, the questions included in the studies were multiple choice, covering only a limited scope of knowledge and requiring the selection of the best answer. This approach deviates significantly from real-world medical education settings; hence, specialized training data is needed for the different types of questions, including open-ended or scenario-based formats. Fourth, it must be emphasized that although the ChatGPT versions demonstrated potential in answering questions from the licensing examination, they should not be solely depended upon for studying for tests. As these LLMs often deliver false or unreliable data [59], it is essential to handle their results with caution and verify them against reliable educational resources, especially in high-stakes exams such as health licensing exams. Fifth, this study did not include any grey literature. Incorporating these sources would help provide a more balanced and comprehensive view of the effectiveness of ChatGPT. Finally, this study concentrated solely

on ChatGPT models, but it is important to consider that other LLMs, such as Google’s Bard, Microsoft’s Bing Chat, and Meta’s LLaMA, have also made significant advancements and ongoing improvements. Therefore, future research should explore the application of LLMs beyond ChatGPT to offer an up-to-date perspective on their efficacy in medical and health education.

Conclusions

This study evaluated the performance of both ChatGPT versions across four types of health licensing examinations using a meta-analysis. The findings indicated that GPT-4 significantly outperformed its predecessor, ChatGPT-3.5, in terms of accuracy in providing correct responses. Additionally, the ChatGPT models showed higher proficiency for pharmacy, followed by medicine, dentistry, and nursing. However, future research needs to incorporate larger and more varied sets of questions, as well as advanced generations of AI chatbots, to achieve a more in-depth understanding in health educational and clinical settings.

Abbreviations

CNMLE	China National Medical Licensing Examination
ENMLE	English version of China National Medical Licensing Examination
JNPE	Japanese National Examination for Pharmacists
LLMs	Large Language Models
NMLE	National Medical Licensing Examination
NNLE	National Nurse Licensing Examination
NPLE	National Pharmacist Licensing Examination

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analysis
USMLE United States Medical Licensing Examination

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12909-024-05944-8>.

Supplementary Material 1

Author contributions

Conceptualization HKJ and EYK; literature searches HKJ; data extraction and coding HKJ and HUL; data analysis HKJ; writing-original draft HKJ; writing-review and editing HKJ, and EYK; overall study supervision EYK. All the authors have read and approved the final manuscript.

Funding

This research was supported by the Chung-Ang University Graduate Research Scholarship in 2023 and by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A03044296).

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Declarations

Ethics approval and consent to participate

This study did not involve human participants; therefore, this is not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Research Institute of Pharmaceutical Sciences, College of Pharmacy, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, South Korea

²Data Science, Evidence-Based and Clinical Research Laboratory, Department of Health, Social, and Clinical Pharmacy, College of Pharmacy, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, South Korea

³Division of Licensing of Medicines and Regulatory Science, The Graduate School of Pharmaceutical Management, and Regulatory Science Policy, The Graduate School of Pharmaceutical Regulatory Sciences, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, South Korea

Received: 29 March 2024 / Accepted: 22 August 2024

Published online: 16 September 2024

References

- 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. <https://doi.org/10.1016/j.nbt.2023.02.001>.
- Montejo-Ráez A, Jiménez-Zafra SM. Current approaches and applications in natural language processing. *Appl Sci*. 2022;12(10):4859. <https://doi.org/10.3390/app12104859>.
- Open AI. Introducing ChatGPT. San Francisco. <https://openai.com/blog/chat-gpt>. Accessed 10 2024.
- Fui-Hoon Nah F, Zheng R, Cai J, Sia K, Chen L. Generative AI and ChatGPT: applications, challenges, and AI-human collaboration. *J Inf Technol Case Appl Res*. 2023;25(3):277–304. <https://doi.org/10.1080/15228053.2023.2233814>.
- 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. <https://doi.org/10.1016/j.iotcps.2023.04.003>.
- 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. <https://doi.org/10.1056/NEJMs2214184>.
- 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. <https://doi.org/10.2196/45312>.
- 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. <https://doi.org/10.1089/dia.2023.0134>.
- Webb JJ. Proof of concept: using ChatGPT to teach emergency physicians how to break bad news. *Cureus*. 2023;15(5):e38755. <https://doi.org/10.7759/cureus.38755>.
- 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. <https://doi.org/10.3389/fonc.2023.1265024>.
- 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. <https://doi.org/10.1371/journal.pdig.0000198>.
- 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.
- 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. <https://doi.org/10.1101/2023.10.26.23297629>.
- 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. <https://doi.org/10.1038/s41598-023-43436-9>.
- 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. <https://doi.org/10.1111/jocn.16478>.
- Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. 2023;27(1):120. <https://doi.org/10.1186/s13054-023-04393-x>.
- Levin G, Horeish N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. *BJOG*. 2024;131:378–80. <https://doi.org/10.1111/1471-0528.17641>.
- Alfertschofer 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. <https://doi.org/10.1007/s10439-023-03338-3>.
- Shakaran P, Koyyalamudi A, Ngu N, Mareedu L. An independent evaluation of ChatGPT on mathematical word problems (MWP). <https://doi.org/10.48550/arXiv.2302.13814>.
- 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. <https://doi.org/10.1186/s12909-024-05125-7>.
- 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. <https://doi.org/10.1097/JCMA.0000000000000942>.
- 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. <https://doi.org/10.15171/ijhpm.2018.32>.
- 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. <https://doi.org/10.24425/fmc.2023.145435>.
- Rosol M, Gąsior JS, Łaba J, Korzeniowski 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. <https://doi.org/10.1038/s41598-023-46995-z>.
- 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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/10.5694/mja2.52061>.
35. Kunitzu 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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/10.1016/j.jmedinf.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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. *JBIM 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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/10.3171/2023.2.JNS.23419>.
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. <https://doi.org/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. <https://doi.org/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. <https://doi.org/10.1038/d41586-023-00680-3>.
59. Mello MM, Guha N. ChatGPT and physicians' malpractice risk. *JAMA Health Forum*. 2023;4(5):e231938. <https://doi.org/10.1001/jamahealthforum.2023.1938>.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.