Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: A Systematic Review and Meta-Analysis

Mingxin Liu, Tsuyoshi Okuhara, XinYi Chang, Ritsuko Shirabe, Yuriko Nishiie, Hiroko Okada, Takahiro Kiuchi

Submitted to: Journal of Medical Internet Research
on: May 22, 2024

Table of Contents

Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: A Systematic Review and Meta-Analysis

Mingxin Liu[1]; Tsuyoshi Okuhara[2] PhD; XinYi Chang[3]; Ritsuko Shirabe[2] PhD; Yuriko Nishiie[1]; Hiroko Okada[2] PhD; Takahiro Kiuchi[2] PhD

[1]Department of Health Communication Graduate School of Medicine The University of Tokyo Tokyo JP
[2]Department of Health Communication School of Public Health Graduate School of Medicine, The University of Tokyo Tokyo JP
[3]Department of Industrial Engineering and Economics School of Engineering Tokyo Institute of Technology Tokyo JP

Corresponding Author:
Mingxin Liu
Department of Health Communication
Graduate School of Medicine
The University of Tokyo
7-3-1 Hongo, Bunkyo
Tokyo
Tokyo
JP

Abstract

Background: Over the past two years, researchers have used various medical licensing examinations to test whether ChatGPT possesses accurate medical knowledge. The performance of each version of ChatGPT on the medical Licensing Exam in multiple environments showed significant differences. At this stage, there is still a lack of a comprehensive understanding of the variability in ChatGPT's performance on different medical licensing exams.

Objective: In this study, we reviewed all studies on ChatGPT performance in medical licensing examinations up to March 2024. This review aims to contribute to the evolving discourse on artificial intelligence (AI) in medical education by providing a comprehensive analysis of the performance of ChatGPT in various environments. The insights gained from this systematic review will guide educators, policymakers, and technical experts to effectively and judiciously utilize AI in medical education.

Methods: We searched the literature published between January 1, 2022, and March 29, 2024, by searching query strings in WOS, PubMed, and Scopus. Two authors screened the literature according to the inclusion and exclusion criteria, extracted data, and independently assessed the quality of the literature concerning Quality Assessment of Diagnostic Accuracy Studies-2. We conducted both qualitative and quantitative analyses.

Results: A total of 45 studies on the performance of different versions of ChatGPT in medical licensing examinations were included in this study. ChatGPT-4 achieved an overall accuracy rate of 81%, significantly surpassing ChatGPT-3.5, and, in most cases, passed the medical examinations, outperforming the average scores of medical students. Translating the exam questions into English improved ChatGPT-3.5's performance but did not affect ChatGPT-4. ChatGPT-3.5 showed no performance difference between exams from English-speaking and non-English-speaking countries, but ChatGPT-4 performed better on exams from English-speaking countries. ChatGPT-3.5 performed better on short-text questions than on long-text questions. The difficulty of the questions and the use of optimized prompts affected the performance of ChatGPT 3.5 and ChatGPT 4. In image-based multiple-choice questions (MCQ), ChatGPT's accuracy rate ranges from 13.1% to 100%. However, ChatGPT performed significantly worse on open-ended questions compared to MCQs.

Conclusions: Thus, ChatGPT-4 demonstrates considerable potential for future use in medical education. However, due to its incomplete accuracy, inconsistent performance, and the challenges posed by differing medical policies and knowledge across countries, ChatGPT-4 is not yet suitable for use in medical education. Clinical Trial: This systematic review was registered in the International Prospective Register of Systematic Reviews (PROSPERO) database on February 1, 2024 (CRD42024506687).

(JMIR Preprints 22/05/2024:60807)
DOI: https://doi.org/10.2196/preprints.60807

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: A Systematic Review and Meta-Analysis

Abstract

Background: Over the past two years, researchers have used various medical licensing examinations to test whether ChatGPT possesses accurate medical knowledge. The performance of each version of ChatGPT on the medical Licensing Exam in multiple environments showed significant differences. At this stage, there is still a lack of a comprehensive understanding of the variability in ChatGPT's performance on different medical licensing exams.

Objective: In this study, we reviewed all studies on ChatGPT performance in medical licensing examinations up to March 2024. This review aims to contribute to the evolving discourse on artificial intelligence (AI) in medical education by providing a comprehensive analysis of the performance of ChatGPT in various environments. The insights gained from this systematic review will guide educators, policymakers, and technical experts to effectively and judiciously utilize AI in medical education.

Methods: We searched the literature published between January 1, 2022, and March 29, 2024, by searching query strings in WOS, PubMed, and Scopus. Two authors screened the literature according to the inclusion and exclusion criteria, extracted data, and independently assessed the quality of the literature concerning Quality Assessment of Diagnostic Accuracy Studies-2. We conducted both qualitative and quantitative analyses.

Results: A total of 45 studies on the performance of different versions of ChatGPT in medical licensing examinations were included in this study. GPT-4 achieved an overall accuracy rate of 81% (95% CI: 78%-84%, *P*<.01), significantly surpassing the 58% (95% CI: 53%-63%, *P*<.01) accuracy rate of GPT-3.5. GPT-4 passed the medical examinations in 26 of 29 cases and outperforming the average scores of medical students in 13 of 17 cases. Translating the exam questions into English improved GPT-3.5's performance but did not affect GPT-4. GPT-3.5 showed no performance difference between exams from English-speaking and non-English-speaking countries (*P*=.72), but GPT-4 performed better on exams from English-speaking countries significantly (*P*=.02). Any type of prompts could significantly improve GPT-3.5’s (*P*=.03) and GPT-4’s (*P*<.01) performance. GPT-3.5 performed better on short-text questions than on long-text questions. The difficulty of the questions affected the performance of GPT-3.5 and GPT-4. In image-based multiple-choice questions (MCQ), ChatGPT's accuracy rate ranges from 13.1% to 100%. ChatGPT performed significantly worse on open-ended questions compared to MCQs.

Conclusions: GPT-4 demonstrates considerable potential for future use in medical education. However, due to its insufficient accuracy, inconsistent performance, and the challenges posed by differing medical policies and knowledge across countries, GPT-4 is not yet suitable for use in medical education.

Keywords: Large Language Model, ChatGPT, Medical Licensing Examination, Medical Education

Introduction

Background

In November 2022, the online artificial intelligence (AI) chatbot ChatGPT was released to the public and swiftly garnered global attention because of its ability to provide detailed answers to complex queries [1]. ChatGPT has been extensively applied across various domains, including programming, education, business, and law, with notable success in each [2-5]. Researchers have been actively exploring the potential roles and capabilities of ChatGPT in clinical diagnosis, health care and medical education [6,7]. The number of publications on this topic has increased dramatically since late 2022 [8,9]. Specifically, in medical education, ChatGPT can play several important roles, including, but not limited to, the following: First, compared to search engines like Google, which present a list of relevant pages, ChatGPT aims to provide concise and practical answers to users' questions, making it an effective knowledge resource [10, 11]. Second, in medical licensing exams comprising multiple-choice questions (MCQ), ChatGPT can act as a "virtual teaching assistant," providing insights for each question, analyzing common errors, and reinforcing concepts interactively [12]. Third, ChatGPT has the capability to analyze images. Although this feature is still in its early stages, it offers the potential for ChatGPT to serve as a "virtual mentor," capable of analyzing medical images such as skin rashes and X-rays [10]. Fourth, for most medical students who find it challenging to balance studying vast amounts of information, practicing evidence-based medicine, and fulfilling clinical duties, ChatGPT can provide concise summaries of clinical trials and generate key practical points from them [10].

However, a prerequisite for ChatGPT's ability to help medical students in their studies and play a role in medical education, both now and in the future, is that ChatGPT has solid and accurate knowledge of medicine. Medical licensing examinations are a crucial part of the medical education pathway as they assess the readiness of aspiring doctors to enter clinical practice. These exams vary in format and content across countries but typically test medical knowledge, clinical reasoning, and ethical decision-making [13]. Over the past two years, researchers have used medical licensing exams from various countries to test whether ChatGPT possesses accurate medical knowledge [14-57].

Although most of these studies used similar testing methods—inputting medical licensing exam questions into ChatGPT and recording the responses to calculate accuracy—the ChatGPT performance showed significant variation. A study conducted in the United States revealed that GPT-3.5 surpassed the 60% score threshold on the National Board of Medical Examiners (NBME)-Free-Step-1 question, reaching the level of a third-year medical student [21]. However, studies from South Korea, China, and Japan have indicated that GPT-3.5 failed to pass medical examinations in their respective countries [26,43,44,47,48,51,54]. Although GPT-4 performed better overall than GPT-3.5 [33,36,41,44,47], it did not pass the Japanese medical licensing exam [49]. Additionally, ChatGPT performance varies significantly across medical specialties within these examinations [23,25,26,27,33,34,32,34,35].

At this stage, there is still a lack of a comprehensive understanding of the variability in ChatGPT's performance on different medical licensing exams. We believe that prematurely utilizing ChatGPT for clinical diagnosis and medical education without thoroughly evaluating its performance across various medical licensing exams is irresponsible and could endanger human lives.

Literature Review

To the best of our knowledge, three systematic reviews have explored ChatGPT's performance in medical licensing exams [58-60].

A study from United States collected literature up to June 2, 2023, focusing on various types of medical licensing examinations in the United States [58]. Among the 19 included studies, only two were comprehensive medical licensing exams United States Medical Licensing Examination (USMLE), while the remaining 17 were medical specialty exams, such as plastic surgery, anesthesia, and ophthalmology [58]. In contrast to this study, our research extends the literature collection to a global scale and examines the performance of ChatGPT in medical licensing exams from different countries and languages. We believe that the worldwide perspective of the current review is crucial because medical education and licensure standards vary significantly across countries.

A study from Pakistan collected literature up to April 2023, focusing on the performance of GPT-3.5 in various medical licensing exams worldwide [59]. However, with the advent of the more advanced GPT-4, more studies have focused on GPT-4. Our research includes all ChatGPT versions and discusses their performance differences.

A study from China collected the literature up to July 15, 2023 [60]. This study reviewed the performance of ChatGPT for various medical questions. Of the 60 included studies, only three were medical licensing examinations. Additionally, this study created a framework to evaluate the quality of studies on the performance of large language models (LLMs) in medical questions [60]. We slightly modified this evaluation framework and applied it to our study.

Study Aims and Objectives

This study reviewed all studies on ChatGPT's performance in medical licensing exams from January 1, 2022, to March 29, 2024, to clarify the following issues:

1. Can ChatGPT pass the medical licensing exams?
2. How does ChatGPT's performance compare to that of medical students?
3. How did ChatGPT perform in different languages?
4. What is the relationship between question difficulty and ChatGPT's performance?
5. What is the relationship between question length and ChatGPT's performance?
6. How did ChatGPT perform on image-based MCQ?
7. How did ChatGPT perform on open-ended questions?
8. What is the difference in ChatGPT's performance with and without prompts?
9. Comparison of GPT-3.5's and GPT-4's performances.
10. How does ChatGPT perform in medical licensing exams of English-speaking countries and non-English-speaking countries?

By comprehensively evaluating the accuracy of the medical knowledge held by ChatGPT, we integrate these perspectives and offer comprehensive recommendations for applying ChatGPT in medical education.

Overall, this systematic review aimed to fill the knowledge gap regarding the application of ChatGPT in medical licensing exams. Further, it sought to contribute to the evolving discourse on AI in medical education and facilitate future developments and applications in this field. The insights gained from this systematic review will guide educators, policymakers, and technical experts to effectively and judiciously utilize AI in medical education.

To the best of our knowledge, this is the first study to comprehensively review the performance of all versions of ChatGPT on medical licensing exams across different countries.

Methods

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagrams and guidance [61]. This systematic review was registered in the International Prospective Register of Systematic Reviews (PROSPERO) database on February 1, 2024 (CRD42024506687).

Search Strategy

We searched for specific query strings (Multimedia Appendix 1) using the advanced search function in PubMed, Web of Science, and Scopus, with Google Scholar as a supplementary source. Literature published from January 1, 2022, to March 29, 2024, was included in the literature search. The RIS files exported from these three platforms were imported into Rayyan [62]. Two authors (Mingxin Liu and Xinyi Chang) independently screened the titles and abstracts of the retrieved studies using a search strategy to identify those that met the inclusion and exclusion criteria (Table 1). The full texts of these studies were then retrieved and independently assessed for eligibility by two authors. Any disagreements regarding the eligibility of specific studies were discussed and resolved by a third reviewer (Tsuyoshi Okuhara). In addition to the database searches, we searched Google Scholar for triangulations on March 29, 2024. When the preprint and peer-reviewed literature data were identical, we included the peer-reviewed literature in our analysis. As part of the screening process, we recorded the reasons for study exclusion and presented them in a prismatic flow diagram.

Table 1. Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
1. The study tested the performance of ChatGPT in medical licensing examinations 2. Any type of original research literature (peer-reviewed articles, conferences articles, preprints, letters, books, etc.) 3. Literature published from 2022 to 2024 4. Literature on the performance of ChatGPT in all languages 5. Literature on any version of ChatGPT 6. Literature on MCQ, open-ended questions, and all other types of questions for medical licensing examinations	1. Nonnational level medical licensing examination 2. Examinations other than comprehensive medical licensing examination (e.g., medical final exams at universities, medical questions created by the authors themselves, medical specialty exams) 3. Studies that are not related to ChatGPT 4. Duplicate studies 5. Studies that are not published in English 6. Systematic review

Data Extraction and Management

Two reviewers (ML and XC) independently extracted data from the included studies into an Excel spreadsheet by two reviewers (Mingxin Liu and Xinyi Chang). The data were compared, and inconsistencies were resolved via consensus or by a third reviewer (Tsuyoshi Okuhara). The general characteristics to be extracted include the following: 1) title, 2) authors, 3) publication year, 4) publication date, 5) type of publication, 6) country of the medical licensing examination, 7) name of the medical licensing exam, 8) ChatGPT version, 9) language in which ChatGPT was tested, 10) duration of the test, 11) type of questions, 12) counts of correct/total questions, 13) accuracy rate, 14) did ChatGPT pass the exam, 15) comparison between medical students, and 16) was a prompt used.

Assessing the Risk of Bias in the Included Studies

Previous study developed an LLM evaluation framework based on the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) [60,63]. We modified and applied this evaluation framework in our study (Multimedia Appendix 2).

Since this previous study collected articles on ChatGPT's performance across all types of medical questions [60], we modified the original framework, whereas our research focused on ChatGPT's performance in medical licensing exams. Specifically, we added two evaluation items, Items 4 and 5, to address aspects specific to medical licensing exams. We removed Item 8 (Are the questions individual standalone queries or a continuous conversation requiring multiple consecutive inquiries?) from the original evaluation framework, as it did not apply to our study.

In our modified evaluation framework, "Task Generation," "Conversation Structure," and

"Evaluation" correspond to "Patient Selection," "Index Test," and "Reference Standard" in QUADAS-2, respectively. Items 2 and 7 correspond to "Flow and Timing" in QUADAS-2.

Evidence Synthesis

Our analysis focuses on GPT-3.5 and GPT-4.

Qualitative Analyses

We performed a comprehensive summary using narrative analysis and descriptive statistics for the contents of the included studies that were narrative or lacked sufficient data.

Quantitative Analyses

We used the raw correct and total data in each included study to calculate the accuracy rate. The calculation rules are as follows: if a study used one set of questions for repeated testing, the displayed accuracy rate is the average score of all attempts and the total number of questions in the set. If the study tested both the original language and translated English questions, the displayed accuracy rate was based on the scores from the original language exam questions. For studies tested with and without optimized prompts, the displayed accuracy rate was based on the scores without optimized prompts. In studies that included multiple-choice and open-ended questions, the displayed accuracy rate excluded scores from the open-ended questions.

We conducted a meta-analysis of studies that tested ChatGPT using MCQ questions.

The I^2 statistic was used to assess the effect of heterogeneity on the pooled results. When significant heterogeneity was present ($I^2 > 50\%$), a random effects model was used; otherwise, a fixed effects model was used. Accuracy was reported with a 95% confidence interval (CI). The significance level was set at $p < 0.05$. Meta-regression and subgroup analysis was conducted to examine the potential sources of heterogeneity and compare performances across different subgroups. A sensitivity analysis was conducted to assess the robustness of the meta-analysis results. Accuracy was reported with 95 % confidence intervals (CIs). The "metafor" and "meta" package in R 4.4.0 were utilized for the meta-analysis, publication bias and sensitivity analyses.

Additionally, we conducted post hoc power analysis for random effects model results of each main group and subgroup. G*Power 3.1.9.7 was utilized for the power analysis.

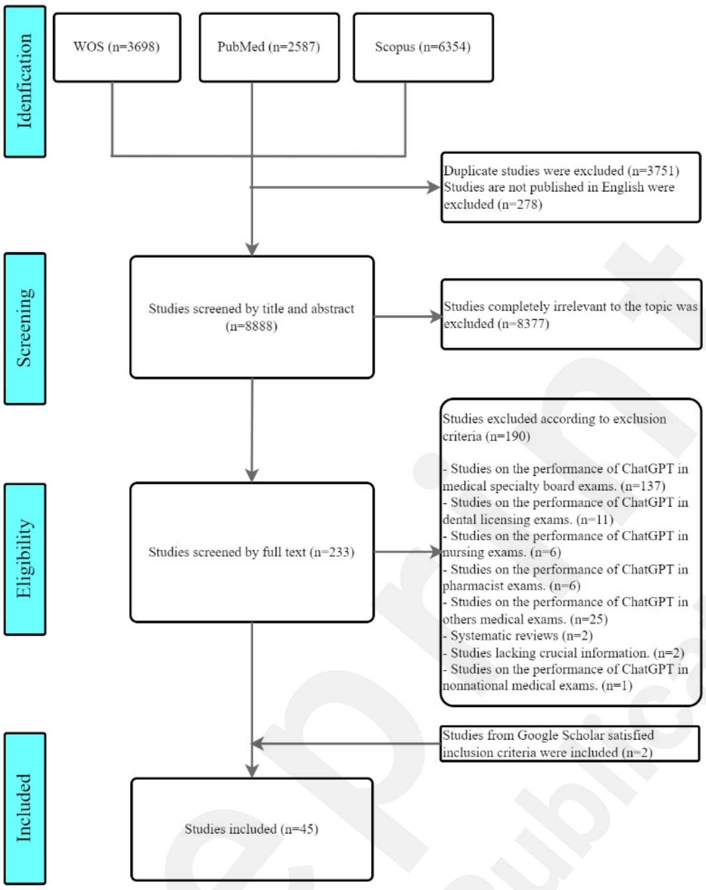
Results

Literature Screening and Selection

By searching the query strings in the WOS, Scopus, and PubMed, we retrieved 3,698 articles from the WOS, 6,354 articles from Scopus, and 2,587 articles from PubMed. After excluding 3,751 duplicate articles, 8,888 articles remained. We excluded 278 non-English articles, leaving 8,610 articles. After obliterating 8,377 articles unrelated to the research topic, 233 remained.

A total of 137 focused on ChatGPT's performance in medical specialty exams, 11 on dental licensing exams, six on nursing exams, six on pharmacist exams, and 25 on other medical exams (e.g., university medical entrance exams, university medical final exams). Further, two were systematic reviews, one was about non-national medical exams, and two lacked the necessary information. These studies did not meet the inclusion criteria.

We then performed a supplementary search using Google Scholar and added two preprint articles on March 29, 2024. Ultimately, 45 articles were included in this systematic review [3, 14-57]. (Figure 1)



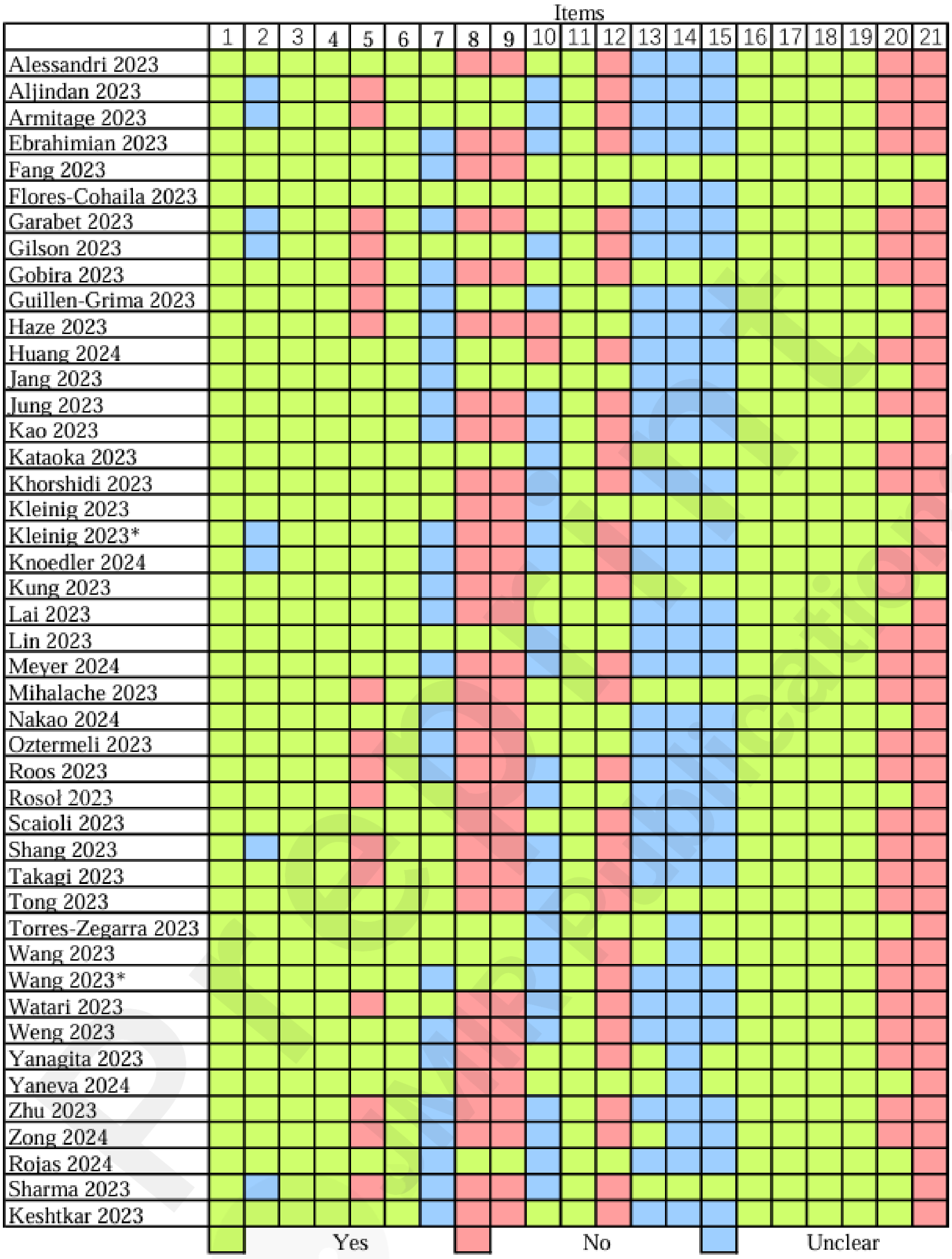
Quality Assessment of Included Studies

Two authors independently assessed the quality of the 45 studies using an evaluation framework, and any disagreements were resolved through discussion and consensus. (Figure 2). The literature we collected tested ChatGPT's performance using national medical licensing exams comprising MCQs with standard answers. Consequently, Items 13, 14, 15, and 21 pertain to evaluators were not mentioned in three-quarters of the included studies. Unlike open-ended questions, MCQs do not require multiple evaluators to adopt a double-anonymized approach to evaluate test results. Therefore, this does not increase the risk in the "Reference Standard" part.

For item seven, more than half of the studies did not specify the exact test dates. On November 6, 2023, OpenAI developers announced that the cutoff dates for ChatGPT versions 3.5 and 4 were updated from September 2021 to January 2022 and April 2023, respectively [64]. We believe that if the cutoff date of ChatGPT is updated during the testing period, this might affect the consistency of ChatGPT's performance before and after the update.

For Item 10, more than half of the studies did not specify whether a new chat session was used to test different questions. Conducting different questions in the same session might have affected the ChatGPT performance.

For reasons above, in the risk of bias assessment, only two studies and three studies were rated as high risk in the "Index Test" and "Flow and timing" categories, respectively (Figure 3).



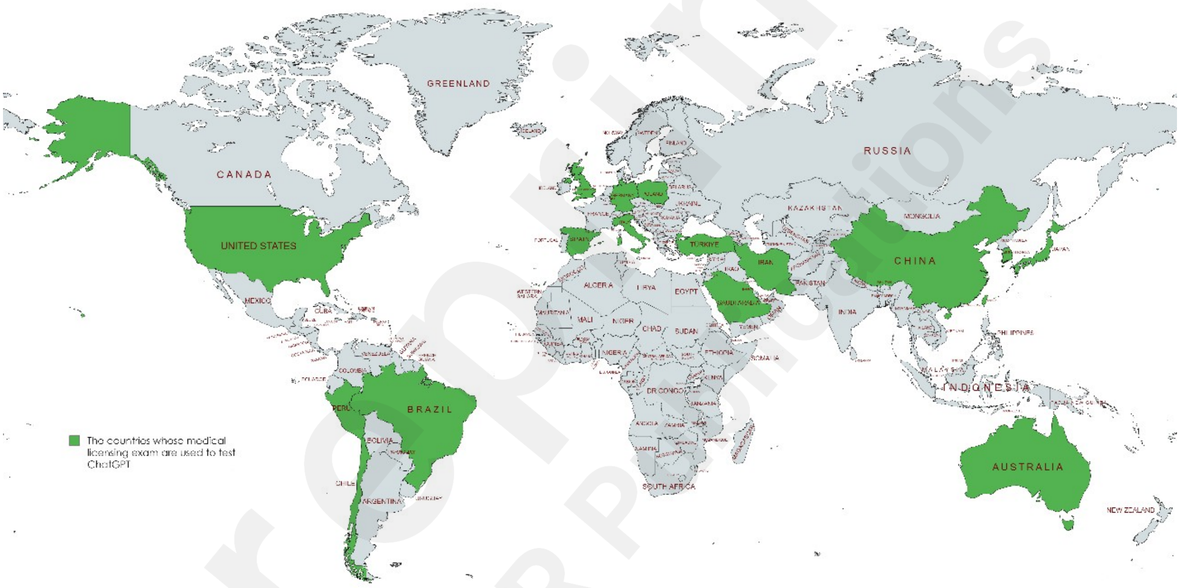
Risk of Bias				
Study	Patient Selection	Index Test	Reference Standard	Flow and Timing
Alessandri 2023	Low	Low	Low	Low
Aljindan 2023	Unclear	Unclear	Low	Low
Armitage 2023	Unclear	Unclear	Low	Low
Ebrahimian 2023	Low	Unclear	Low	Unclear
Fang 2023	Low	Low	Low	Unclear
Flores-Cohaila 2023	Low	Low	Low	Low
Garabet 2023	Unclear	Low	Low	High
Gilson 2023	Unclear	Unclear	Low	Low
Gobira 2023	Low	Low	Low	Unclear
Guillen-Grima 2023	Low	Unclear	Low	Unclear
Haze 2023	Low	High	Low	Unclear
Huang 2024	Low	High	Low	Unclear
Jang 2023	Low	Low	Low	Unclear
Jung 2023	Low	Unclear	Low	Unclear
Kao 2023	Low	Unclear	Low	Unclear
Kataoka 2023	Low	Unclear	Low	Low
Khorshidi 2023	Low	Unclear	Low	Low
Kleinig 2023	Low	Unclear	Low	Low
Kleinig 2023*	Low	Unclear	Low	High
Knoedler 2024	Low	Unclear	Low	Unclear
Kung 2023	Low	Low	Low	Unclear
Lai 2023	Low	Low	Low	Unclear
Lin 2023	Low	Unclear	Low	Low
Meyer 2024	Low	Unclear	Low	Unclear
Mihalache 2023	Low	Low	Low	Low
Nakao 2024	Low	Low	Low	Unclear
Oztermeli 2023	Low	Low	Low	Unclear
Roos 2023	Low	Unclear	Low	Unclear
Rosoł 2023	Low	Unclear	Low	Low
Scaioli 2023	Low	Low	Low	Low
Shang 2023	Unclear	Unclear	Low	Low
Takagi 2023	Low	Unclear	Low	Low
Tong 2023	Low	Unclear	Low	Low
Torres-Zegarra 2023	Low	Unclear	Low	Low
Wang 2023	Low	Unclear	Low	Low
Wang 2023*	Low	Unclear	Low	Unclear
Watari 2023	Low	Unclear	Low	Low
Weng 2023	Low	Unclear	Low	Unclear
Yanagita 2023	Low	Low	Low	Unclear
Yaneva 2024	Low	Low	Low	Low
Zhu 2023	Low	Unclear	Low	Unclear
Zong 2024	Low	Unclear	Low	Unclear
Rojas 2024	Low	Unclear	Low	Unclear
Sharma 2023	Unclear	Unclear	Low	High
Keshtkar 2023	Low	Low	Low	Low

General Characteristics of Included Studies

Among the 45 reviewed articles, the earliest was published on February 8, 2023 [21], and the latest on April 30, 2024 [55]. The general characteristics of the studies are shown in Multimedia Appendix 3.

The medical licensing exams applied to test ChatGPT's performance were from 17 countries and regions: Italy (2), Saudi Arabia (1), the United Kingdom (2), Iran (3), China (7), Peru (2), the United States (7), Brazil (1), Spain (1), Japan (6), Taiwan (4), South Korea (1), Germany (3), Australia (2), Turkey (1), Poland (1), and Chile (1)(Figure 4).

Of the 45 included studies, 29 tested the performance of GPT-4, and 26 tested the performance of GPT-3.5. A total of 14 studies tested both GPT-4 and GPT-3.5. Additionally, four studies tested the GPT-3, one tested the InstructGPT, and one tested the ChatGPT Plus.



Regarding the countries and languages of the medical licensing exam questions used to test ChatGPT, 11 studies used exams from an English-speaking country. Of the 34 medical licensing exams of non-English-speaking countries, 22 used only the native language for testing, three translated the original language into English, and nine used both the original and translated English questions.

All 45 studies included MCQs, with four studies including open-ended questions, 1 study including calculation questions, and 1 study including patient history inquiry questions.

Qualitative Analyses

Regarding performance of ChatGPT on passing medical licensing exam, among the 26 studies testing GPT-3.5, six reported that GPT-3.5 passed the medical licensing exam, and four reported satisfactory performance, making up 38.5% of the total. In the remaining studies, one was unclear, and 15 did not pass. Among the 29 studies testing GPT-4, 17 reported that GPT-4 passed the medical licensing exam, and nine reported satisfactory performance, making up 89.7% of the total. In the remaining studies, one was unclear, and two did not pass (Figure 5). For the other ChatGPT models, among the four studies testing the GPT-3 performance, two did not pass, one was unclear, and one showed a satisfactory performance. The studies that tested GPT-4 with Vision (GPT-4V, which is specifically designed for image tasks), InstructGPT, and ChatGPT Plus showed the following results: passed, did not pass, and did not pass.

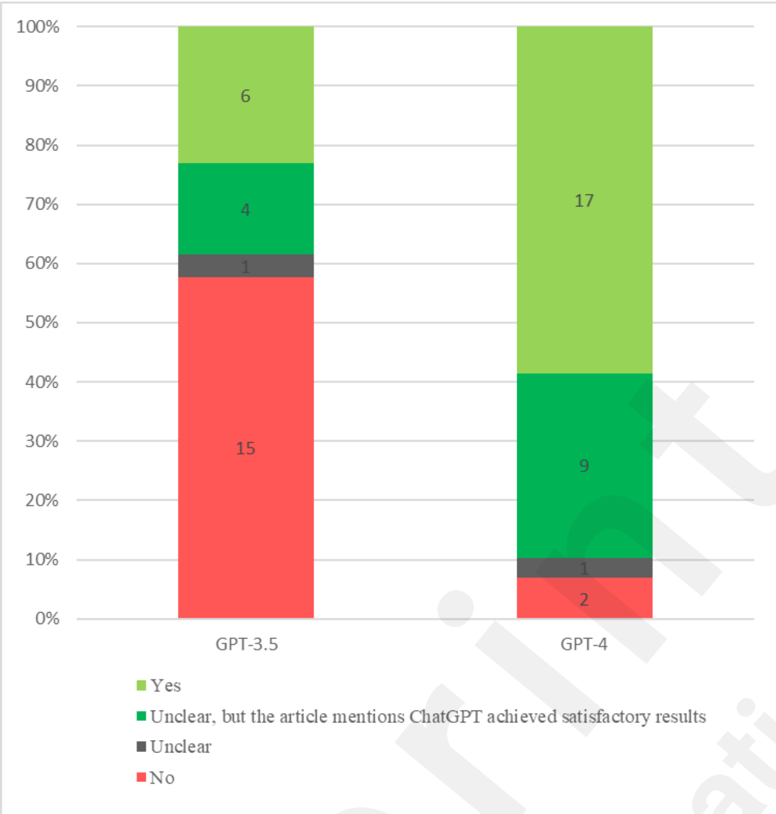
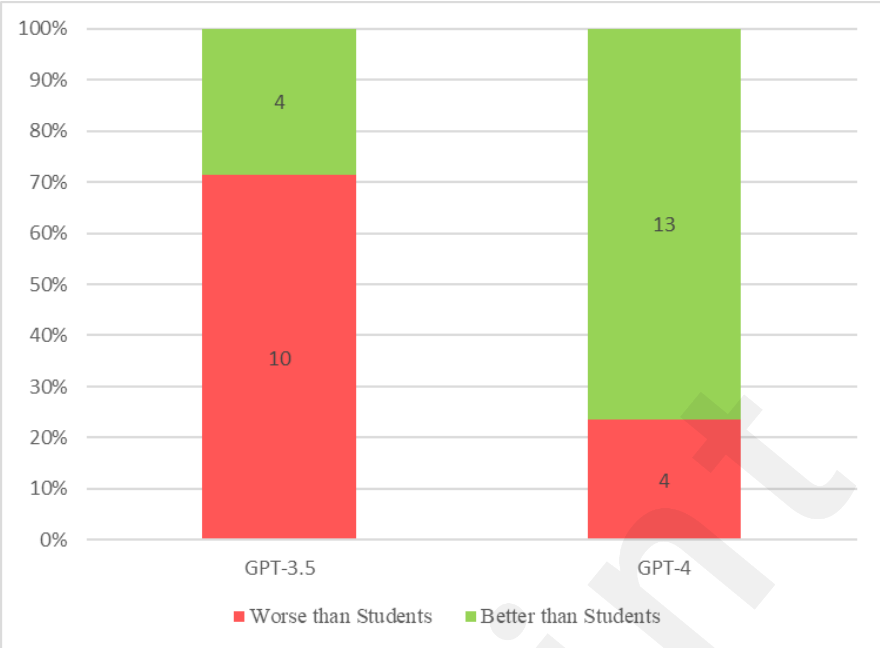


Figure 5. Performance of ChatGPT on passing medical licensing exam

Regarding performance of ChatGPT compared with medical students, 14 of 45 studies compared GPT-3.5's performance with medical students, and 17 of 45 compared GPT-4's performance with that of medical students. Four studies showed that GPT-3.5 surpassed medical students, accounting for 28.6% of the studies (4/14). Thirteen studies showed that GPT-4 surpassed medical students, accounting for 76.5% of the studies (13/17; Figure 6). For the other ChatGPT models, one study showed that GPT-3 surpassed medical students, while another showed that it performed worse. One study indicated that InstructGPT performed worse than the students.



We also compared ChatGPT’s performance in original language and English-translated questions of the same non-English medical licensing exam. In studies of medical licensing exams in non-English-speaking countries, nine used both the original language and English-translated questions to test ChatGPT's performance, with eight reporting comparative results (Table 2). Overall, for GPT-4, translating the original language into English had a limited effect on improving the performance. The accuracy improvement ranged from 0.17% to 8.65%, with six studies showing an accuracy increase of less than 5%. However, compared with GPT-4, GPT-3.5 showed significant improvement when tested in English in four studies. In two of these studies, GPT-3.5's accuracy was more than 20% higher in English than in the original language.

Table 2. ChatGPT's performance in original language and English-translated question [18,23,26,30,41,45,47,57]

Author, Year, Citation Number	**GPT-3.5 Accuracy Rate**		**GPT-4 Accuracy Rate**	
	Original language	English-translated	Original language	English-translated
Fang 2023 [18]	Untested		75.77% (197/260)	77.31% (201/260)
Guillen-Grima 2023 [23]	63.18% (115/182)	66.48% (121/182)	86.81% (158/182)	87.91% (160/182)
Jang 2023 [26]	Untested		51.82% (unclear)	60.47% (unclear)
Khorshidi 2023 [30]	Untested		81.30% (161/198)	84.30% (167/198)
Rosoł 2023 [41]	54.79% (320.5/585)	60.34% (353/585)	79.57% (465.5/585)	79.74% (466.5/585)
Tong 2023 [45]	Untested		81.25% (130/160)	86.25% (138/160)
Wang 2023 [47]	56% (56/100)	76% (76/100)	84% (84/100)	86% (86/100)

Keshtkar 2023 [57]	35.66% (394/1105)	61.36% (687/1105)	Untested

Two and three studies examined the correlation between GPT-3.5 or GPT-4 performance and the length of the question text. Both studies on GPT-3.5 showed a significant correlation between performance and the length of the question text; the longer the question text, the poorer the performance of GPT-3.5 [33,39]. By contrast, none of the three studies on GPT-4 found a significant difference in performance between long- and short-text questions [23,37,50].

Eight studies examined the correlation between the difficulty of the questions and ChatGPT's performance. Seven studies indicated that both GPT-4 and GPT-3.5 performed worse on difficult questions than easier ones [21.23,30,33,41,44,49]. Only one study showed that the difficulty of the questions did not affect GPT-4's performance. However, in this study, the difficulty was subjectively rated by three medical students rather than using official difficulty ratings [45].

Regarding ChatGPT's performance with and without optimized prompt, in our review of 45 articles, 13 stated that researchers provided ChatGPT with prompts before asking questions. Most of these prompts were designed to help ChatGPT better understand its task, such as "You are now an experienced clinician; please answer the following questions," or "You are a medical student, and we will be using medical licensing exam questions to test you; please provide your best answers." Researchers have not analyzed or elaborated on the impact of these task understanding prompts on ChatGPT's performance. However, three studies used optimized prompts [19,26,35]. A Korean study used four kinds of optimized prompts including: annotating Chinese terms in TKM, translating the instruction and question into English, providing exam-optimized instructions, and utilizing self-consistency in the prompt. The results showed ChatGPT's accuracy increased from 51.82% to 66.18% with optimized prompts [26]. In the other two studies, questions that ChatGPT initially answered incorrectly without prompts were re-asked with optimized prompts such as “Are you sure? Pretend to be a junior doctor with expertise in clinical practice and exam solving and retry” or “Could you double-check the answer?”. ChatGPT could correctly answer up to 88.9% and 84% of these questions, respectively [19,35]. For task understanding prompts, we conducted a subgroup analysis and meta-regression to examine whether they affected ChatGPT's performance.

Regarding the capability of ChatGPT on answering image-based MCQS, four studies have reported the performance of ChatGPT in image-based MCQs. Three tested GPT-4, and one compared GPT-4 and GPT-4V [16,23,38,55]. In a UK study, GPT-4 achieved an accuracy rate of 100% (3/3) for the image-based MCQs correctly [16]. In a Spanish study, the accuracy rate of GPT-4 for image-based MCQs in Spanish was 13%, and the accuracy rate was 26% after translating the image-based MCQs into English, twice as high as in Spanish [23]. Japanese researchers tested GPT-4's performance on image-based MCQs that provided both images and text and on image-based MCQs that provided only text. The rate of correctness was 68% (73/108) when both images and text were provided and 72% (78/108) when only text was provided [38]. Researchers in Chile compared the performance of chatgpt4 and chatgpt4v in image-based MCQs. Accuracy rates of GPT-4 and GPT-4V for image-based MCQs were 76.7% and 70%, respectively [55].

Regarding performance of ChatGPT on questions other than MCQs, 4 studies comparing ChatGPT's performance on open-ended questions versus MCQs. Among them, two showed that ChatGPT performed significantly worse on open-ended questions than on MCQs [3,19], one showed slightly better performance on open-ended questions, and another asked ChatGPT 10 short questions, all of which received an "A" grade [28,56]. In a study using calculation questions from the Japanese medical licensing exam, ChatGPT's performance on calculation questions was significantly worse than that of MCQs [24]. In a study using patient history inquiry questions from the Chinese medical licensing exam to assess medical students' clinical skills, ChatGPT passed the test and scored higher than the average medical student, achieving satisfactory performance [53].

Meta-analysis

We conducted a meta-analysis of the integrated accuracy of GPT-3.5 and GPT-4 in medical licensing examinations. The accuracy we involved to meta-analysis was displayed in Multimedia Appendix 3. 25 studies reporting the accuracy of GPT-3.5 and 29 studies reporting the accuracy of GPT-4 were included in this meta-analysis. Owing to significant heterogeneity (GPT-3.5: $I^2 = 95\%$, GPT-4: $I^2 = 93\%$), both groups were analyzed using a random-effects model.

The integrated accuracy for GPT-3.5 was 58% (95% CI: 53%-63%, $P<.01$), and the integrated accuracy for GPT-4 was 81% (95% CI: 78%-84%, $P<.01$) (Figures 7 and 8).

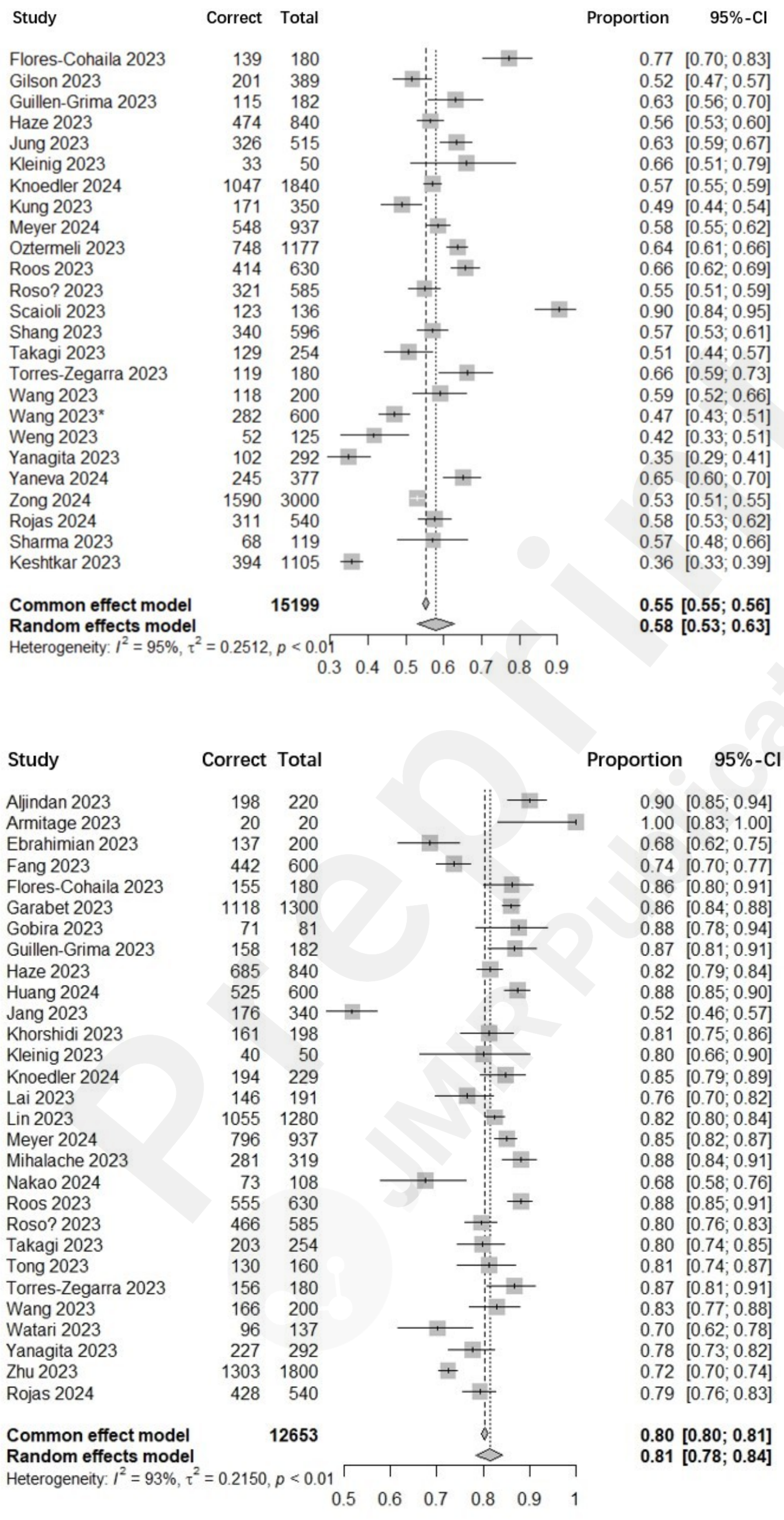


Figure 8. Performance of GPT-4 in medical licensing exam

Meta-regression and subgroup analysis

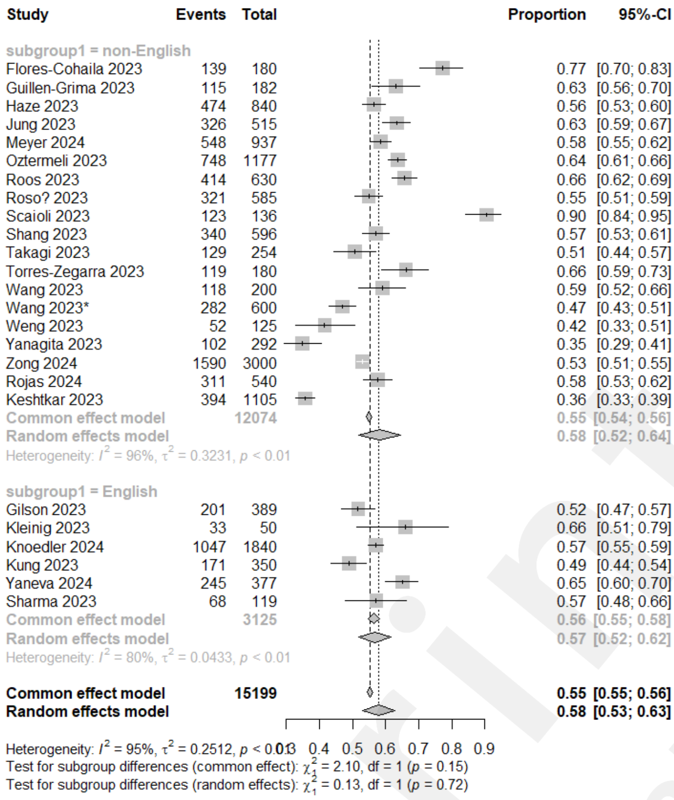
We divided studies with GPT-3.5 and GPT-4 in Figures 7 and 8 into three subgroups, respectively.

Subgroup 1 divided the studies into those using medical licensing exams from English-speaking countries to test ChatGPT and those using exams from non-English-speaking countries with native language. Subgroup 2 categorized studies based on whether they used prompts to test ChatGPT or not. In figure 11 and 12, “Yes” indicates the use of prompts, while "No" indicates the absence of prompts. Subgroup 3 categorized studies according to the "Flow and Timing" evaluation in Figure 3, with "Low risk" forming one category and "Unclear" and "High Risk" forming another. In figure 13 and 14, "Yes" means "Low risk," implying that ChatGPT's performance might not be affected by testing date and source date. "No" means "High risk" and "Unclear," implying that ChatGPT's performance might be influenced by testing date and source date. We conducted meta-regression and subgroup analyses for all subgroups to examine potential sources of heterogeneity and compare performances.

In subgroup analysis of subgroup 1, because of significant heterogeneity (GPT-3.5 tested in medical licensing exams of English-speaking countries: I^2 = 80%; GPT-3.5 tested in original language exams of non-English-speaking countries: I^2 = 96%; GPT-4 tested in medical licensing exams of English-speaking countries: I^2 = 69%; GPT-4 tested in original language exams of non-English-speaking countries: I^2 = 93%), all four groups were analyzed using a random-effects model.

The integrated accuracy for GPT-3.5 in exams from English-speaking countries was 57% (95% CI: 52%-62%, P<.01), and in exams from non-English-speaking countries with original languages, it was 58% (95% CI: 52%-64%, P<.01). No statistically significant differences were observed (P=.72). (Figure 9)

For GPT-4, the integrated accuracy in exams from English-speaking countries was 86% (95% CI: 82%–89%, P<.01), and in exams from non-English-speaking countries with original languages, it was 80% (95% CI: 76%–83%, P<.01). Statistically significant differences were observed between the results (P=.02). (Figure 10)



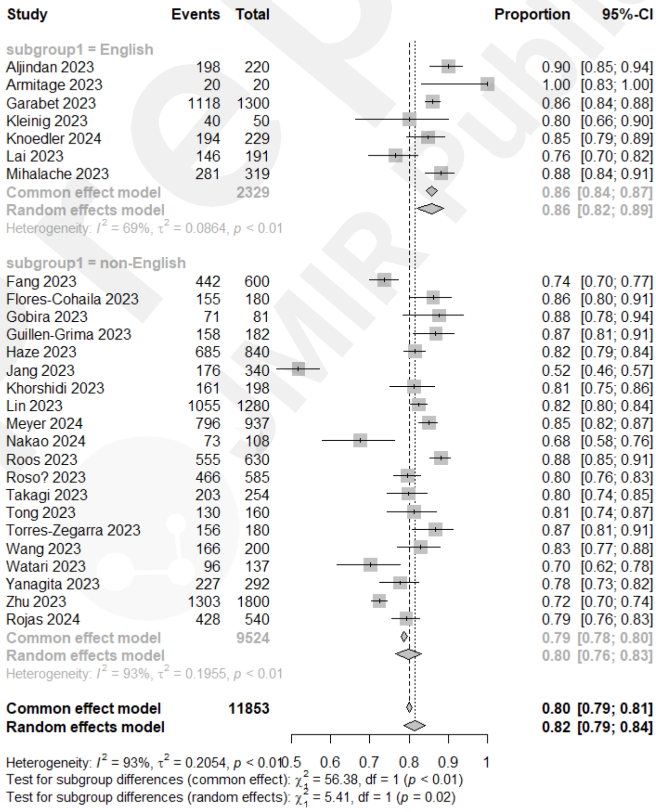


Figure 10. Subgroup 1: Performance of GPT-4 on medical licensing exam from English-speaking countries and non-English-speaking countries.

In subgroup analysis of subgroup 2, because of significant heterogeneity (GPT-3.5 in subgroup “Yes”: I² = 92%; GPT-3.5 in subgroup “No”: I² = 95%; GPT-4 in subgroup “Yes”: I² = 68%; GPT-4 in subgroup “No”: I^2 = 94%), all four groups were analyzed using a random-effects model.

The integrated accuracy for GPT-3.5 in exams with prompt was 68% (95% CI: 57%-77%, P<.01), and in exams without prompt, it was 54% (95% CI: 50%-59%, P<.01). Statistically significant differences were observed between the results (P=.03). (Figure 11)

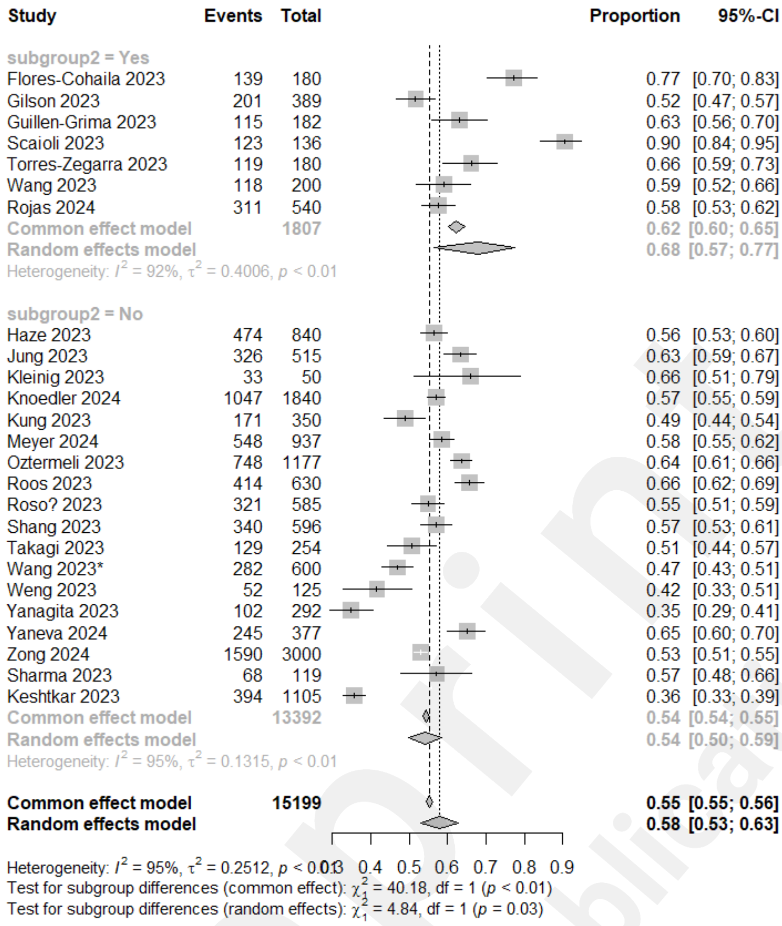


Figure 11. Subgroup 2: Performance of GPT-3.5 with or without prompts.

The integrated accuracy for GPT-4 in exams with prompt was 85% (95% CI: 83%-88%, *P*<.01), and in exams without prompt, it was 79% (95% CI: 75%-82%, *P*<.01). Statistically significant differences were observed between the results (*P*<.01). (Figure 12)

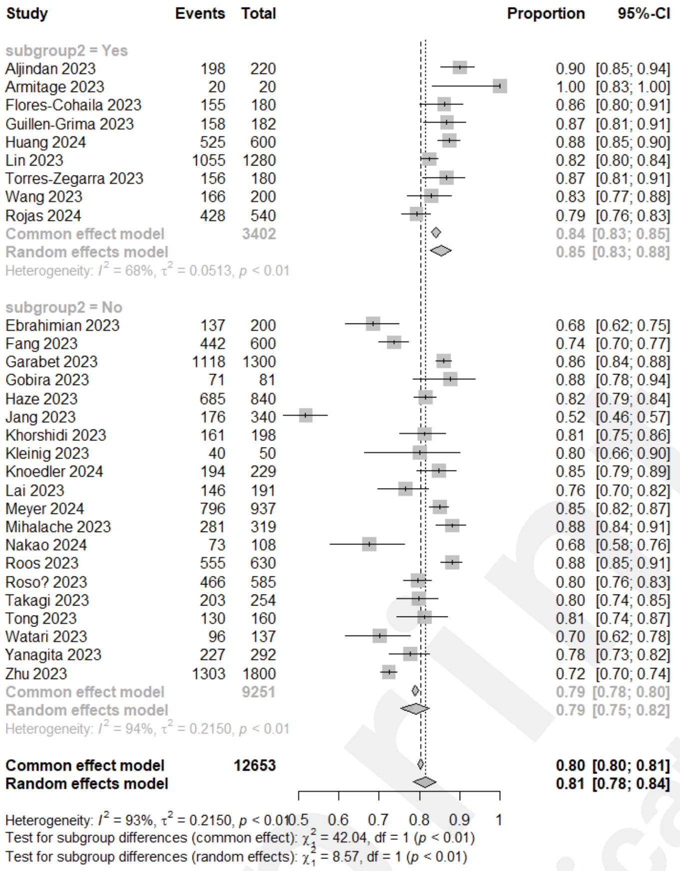


Figure 12. Subgroup 2: Performance of GPT-4 with or without prompts.

In subgroup analysis of subgroup 3, because of significant heterogeneity (GPT-3.5 in subgroup “Yes”: I^2 = 96%; GPT-3.5 in subgroup “No”: I^2 = 92%; GPT-4 in subgroup “Yes”: I^2 = 71%; GPT-4 in subgroup “No”: I^2 = 95%), all four groups were analyzed using a random-effects model.

The integrated accuracy for studies in which GPT-3.5’s performance may be influenced by testing date and source date was 55% (95% CI: 51%-60%, *P*<.01), and in studies in which GPT-3.5’s performance may not be influenced, it was 62% (95% CI: 53%-71%, *P*<.01). No statistically significant differences were observed (*P*=.19). (Figure 13)

The integrated accuracy for studies in which GPT-4’s performance may be influenced by testing date and source date was 80% (95% CI: 75%-83%, *P*<.01), and in studies in which GPT-4’s performance may not be influenced, it was 83% (95% CI: 80%-86%, *P*<.01). No statistically significant differences were observed (*P*=.12). (Figure 14)

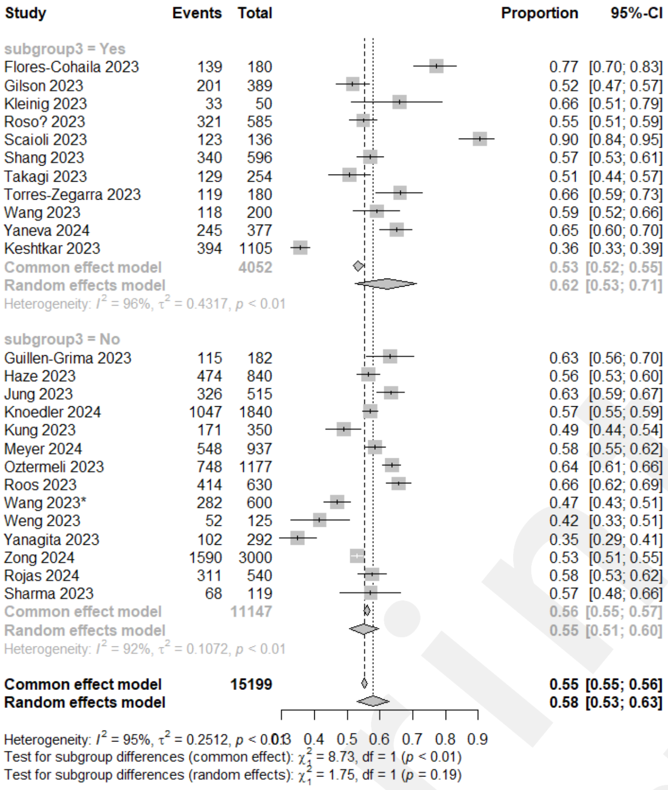


Figure 13. Subgroup 3: Performance of GPT-3.5 regarding “Flow and Timing”.

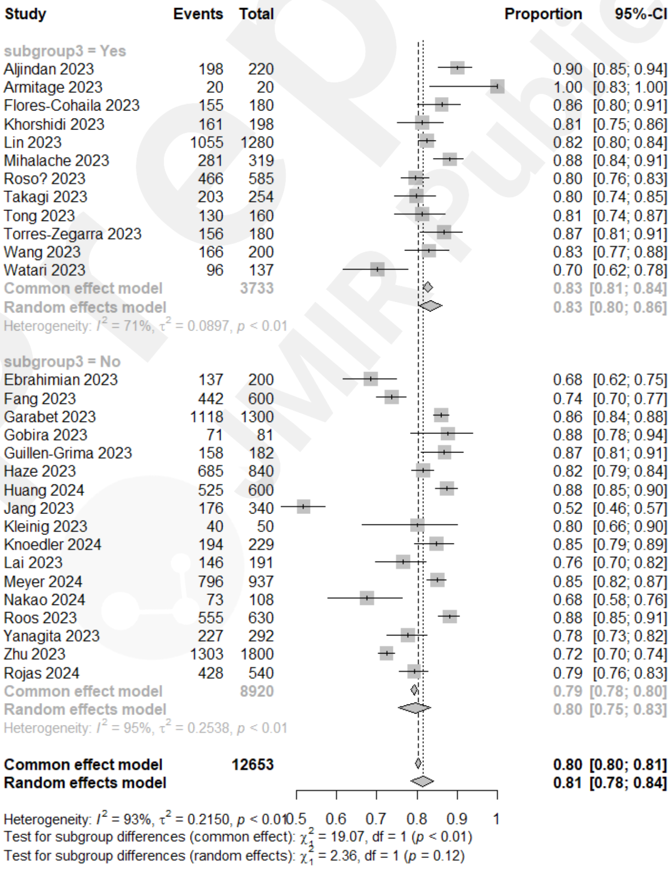


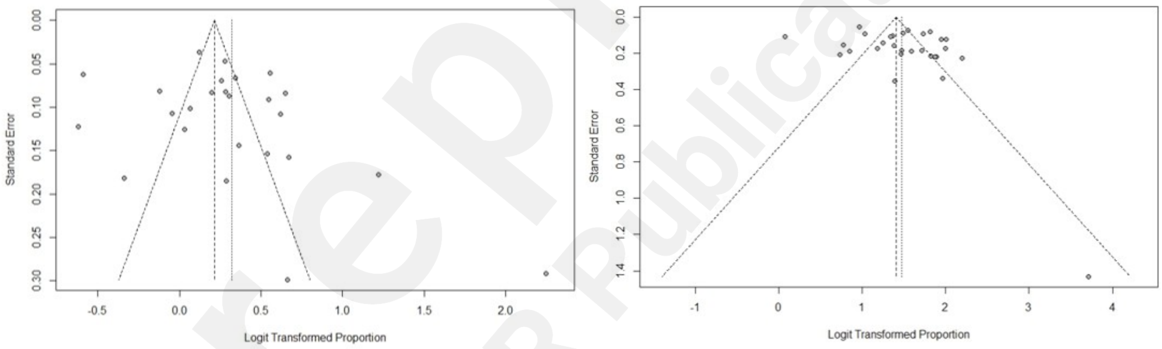
Figure 14. Subgroup 3: Performance of GPT-4 regarding “Flow and Timing”.

Regarding the meta-regression results for all subgroups, the use of prompts is likely to be a source of potential heterogeneity and showed a significant effect on the accuracy rates of GPT-3.5 and GPT-4 (subgroup 2), as indicated by an estimated regression coefficient of 0.54 (*P*=.01) and 0.46 (*P*=.02), respectively. Meta-regression of subgroups 1 and 3 did not show statistically significant effects on accuracy rates (all *P*>.05). (Table 3)

Table 3. Meta-regression results of three subgroups of GPT-3.5 and GPT-4.

Meta-regression			
Version		**Estimated regression coefficient**	***P*-value**
GPT-3.5	Subgroup 1	-0.03	.91
	Subgroup 2	0.54	.01
	Subgroup 3	0.28	.19
GPT-4	Subgroup 1	-0.39	.08
	Subgroup 2	0.46	.02
	Subgroup 3	0.25	.18

Publication Bias

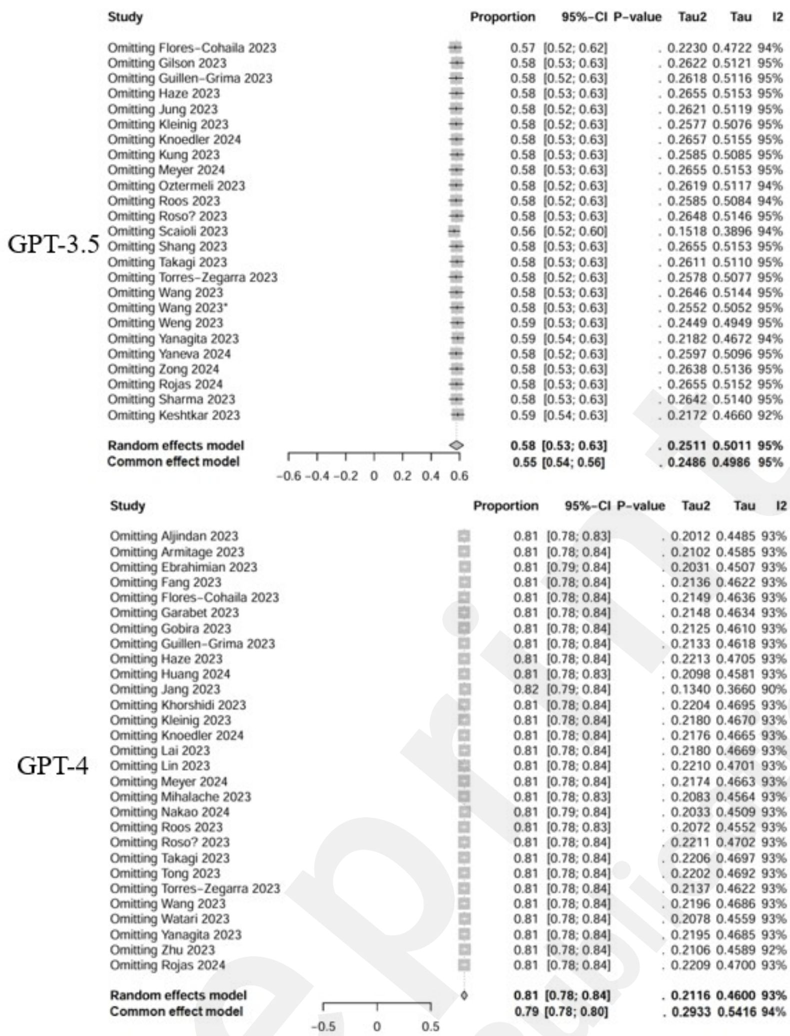


GPT-3.5 GPT-4

No publication bias was detected among the included studies, as indicated by the funnel plots. (Figure 15)

Sensitivity analyses

We used random effects model to assess the impact of excluding individual studies on overall effects. The sensitivity analysis plot showed that no single study significantly affected the overall meta-analysis results. This demonstrates the robustness of the meta-analysis results (Figure 16).



Power analysis

We conducted post hoc power analysis for the main groups and subgroups using results of random effects model. Subgroup 1 of GPT-3.5 had a power of 0.17. In this subgroup, we believe the sample size is adequate. The low power might be due to two main reasons. First, the inter-group difference is minimal, with effect sizes being very close (58% and 57%). Second, the data may have high heterogeneity (I^2 = 80% and 96%). In the main group and other subgroups, the power was 1 or close to 1, indicates sufficient power to detect the anticipated effect size with the given sample size for the random effects model.

Table 4. Power analysis results of main groups and subgroups.

Power analysis		
Version	**Group**	**Power**
GPT-3.5	Main group (Integrated accuracy rate in figure 7)	1
GPT-4	Main group (Integrated accuracy rate in figure 8)	1
GPT-3.5	Subgroup 1	0.17
	Subgroup 2	1
	Subgroup 3	1
GPT-4	Subgroup 1	1

	Subgroup 2	1
	Subgroup 3	0.98

Discussion

Principle Findings

Our systematic review and meta-analysis are the first to comprehensively evaluate the performance of all versions of ChatGPT across various medical licensing exam environments. Overall, GPT-4 significantly outperformed GPT-3.5; however, there are still some issues that make it difficult to use in medical education at this stage.

Regarding the accuracy of ChatGPT on MCQs, while two previous studies conducted meta-analyses that yielded accuracy rates of 61% and 56%, respectively, we noted that these accuracy rates reflected the performance of all versions of ChatGPT without differentiation by version [54,56]. Our review found that GPT-4 achieved an integrated accuracy rate of 81% for MCQs in medical licensing exams, passing nearly all tested exams and surpassing the average performance of medical students in three-quarters of the tests. In contrast, GPT-3.5 achieved an integrated accuracy rate of 58%, failing to pass more than half of the medical examinations and surpassing the average performance of medical students in only 4 of 14 tests. Therefore, regarding accuracy rate, passing rate, and comparison with medical students, GPT-4 significantly surpassed GPT-3.5.

In medical licensing exams from non-English-speaking countries, translating the original language questions into English significantly improved GPT-3.5's performance but did not affect GPT-4's performance. This indicates that GPT-4 has a much higher proficiency in languages other than English than GPT-3.5. However, based on the results of subgroup analysis for comparing GPT-3.5 and GPT-4 in medical licensing exams from English-speaking and non-English-speaking countries, we found that GPT-4 performed better in English-speaking countries. In contrast, GPT-3.5 showed no performance difference between exams from English-speaking and non-English-speaking countries.

Additionally, based on the results of qualitative analysis and subgroup analysis, we found that both “optimized prompts” and “task understanding prompts” could significantly improve ChatGPT's performance. When using prompts, the accuracy rates of GPT-3.5 and GPT-4 were 68% and 85% respectively, which were significantly higher than the accuracy rates of 54% and 79% without prompts.

The testing date and source date of each study were not sources of potential heterogeneity and did not significantly affect the performance of ChatGPT.

Challenge of Utilizing ChatGPT in Medical Education

First, although the AI hallucinations of GPT-4 have significantly been reduced compared to earlier versions, GPT-4 still generates incorrect information because the data used to train these models are not always correct [65]. We observed that in all tests of GPT-4, only two instances achieved an accuracy rate above 90%. The only example of a perfect accuracy rate was in UK study, in which GPT-4 correctly answered all 20 questions [16]. However, the number of questions used in this test was significantly lower than those used in other studies. We believe that this demonstrates ChatGPT's potential for future use in medical education but does not imply that medical students can rely on ChatGPT to acquire medical knowledge or prepare for exams. Traditional sources of medical knowledge, such as medical school courses and textbooks, are completely reliable. However, because most professional medical knowledge exists in book form [50], and medical expertise on the Internet is not always reliable [66], the medical knowledge that ChatGPT currently holds is not entirely accurate. In this context, if medical students rely on ChatGPT as a trusted source of expertise and acquire incorrect medical knowledge, the reliability of their knowledge and skills is significantly compromised. This is unacceptable in the medical field, as it directly impacts human lives. Therefore, GPT-4 passing medical licensing exams does not imply that it can be used as a source of knowledge in medical education.

Previous studies have noted that the responses generated by GPT-3.5 are nondeterministic and random [67-69]. Our study found that although the stability of GPT-4 has significantly improved compared to that of GPT-3.5, it still exhibits a degree of randomness in its outputs. Although GPT-4 achieved an overall accuracy of 81% across all tests, it only scored 52% on the Korean medical licensing exam, even lower than the overall accuracy of GPT-3.5 (58%) [25]. Additionally, in four studies using Japanese medical licensing exam questions, although GPT-4 passed three of the tests, it only achieved an accuracy of 67% in one and did not pass [23,38,44,51]. Furthermore, the use of optimized prompts and the difficulty of the questions can affect ChatGPT's performance stability. If millions of medical students use ChatGPT for learning, this randomness could be significantly magnified and affect their learning outcomes.

Moreover, different countries' medical policies, cultural, ethics, and unique local traditional medical knowledge pose significant challenges for ChatGPT [70]. Regarding varying medical policies and ethics, a Chinese study mentioned that abortion is prohibited in the United States but allowed in certain circumstances in China [48]. Although euthanasia is legal in many countries, it is illegal in Japan; ChatGPT chose the option of euthanasia in the Japanese medical licensing exam [25]. ChatGPT may struggle to adapt to localized medical policies and ethics. Additionally, East Asian countries still use local traditional medicine (e.g., Chinese medicine), and most local traditional medicine learning materials are written in the native languages. These materials might not be accessible on the Internet and included in ChatGPT's training dataset, making it difficult for ChatGPT to provide accurate answers to such topics [18,26,50,54].

In the evaluation of image-based questions, we observed significant variations in the performance of GPT-4, with accuracy rates ranging from 13% to 100% [16, 23, 38, 55]. However, there were only three questions in which GPT-4 achieved 100% accuracy, which is too small a sample size to demonstrate its proficiency in handling image-based questions [16]. In addition, a study from Japan tested the performance of ChatGPT when provided with images and text versus text only. Surprisingly, ChatGPT performed better when given only text than when provided with both images and text [38]. Similarly, Chile found that GPT-4V, designed explicitly for image tasks, performed worse on image-based questions than GPT-4 [55]. We believe that studies testing the ChatGPT performance on image-based questions are limited at this stage. Therefore, comprehensive and reliable conclusions cannot be drawn. Consequently, using ChatGPT for image-based medical education is extremely risky.

Finally, human teachers usually recognize their knowledge limitations when faced with uncertain questions and correct their mistakes by consulting resources. However, the fatal issue with ChatGPT is that owing to the nature of AI language models, it can provide detailed and logically sound explanations for incorrect answers [24, 40, 44]. Given ChatGPT's authoritative writing style, students are likely to believe and memorize the incorrect information provided by ChatGPT [71].

Limitation

This systematic review did not include studies on the performance of ChatGPT in various medical specialty examinations, dental licensing examinations, pharmacy examinations, and other medical-related assessments. Future studies should review the performance of ChatGPT in these specific medical fields.

Studies published in languages other than English were excluded from the systematic review. This may omit the literature that tests the performance of ChatGPT on non-English-speaking medical licensing exams.

Conclusion

A total of 45 studies on the performance of different versions of ChatGPT in medical licensing examinations were included in this systematic review. GPT-4 achieved an overall accuracy rate of 81%, significantly surpassing GPT-3.5, and, in most cases, passed the medical examinations, outperforming the average scores of medical students. Thus, GPT-4 demonstrates considerable potential for future use in medical education. However, because the knowledge of ChatGPT is not entirely accurate and its performance can be inconsistent, and because of the challenges posed by differing medical policies and knowledge across countries, we believe that GPT-4 is not yet suitable for use in medical education.

Conflicts of Interest

None declared.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24KJ0830.

Multimedia Appendix 1: [Query strings of WOS, Scopus, and PubMed.]
Multimedia Appendix 2: [Evaluation framework used in this systematic review.]
Multimedia Appendix 3: [General characteristics of included studies.]
Multimedia Appendix 4: [PRISMA 2020 checklist.]

Reference

1. OpenAI. ChatGPT. https://chat.openai.com/chat [accessed Feb 12, 2024]
2. Khlaif ZN, Mousa A, Hattab MK, et al. The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation. JMIR Med Educ. 2023;9:e47049. Published 2023 Sep 14. doi:10.2196/47049. PMID: 37707884
3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. Published 2023 Feb 9. doi:10.1371/journal.pdig.0000198. PMID: 36812645
4. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the Situational Judgement Test-A Professional Dilemmas-Based Examination for Doctors in the United Kingdom. JMIR Med Educ. 2023;9:e48978. Published 2023 Aug 7. doi:10.2196/48978. PMID: 37548997
5. Rahman MM, Watanobe Y. ChatGPT for Education and Research: Opportunities, Threats, and Strategies. Applied Sciences. 2023; 13(9):5783. https://doi.org/10.3390/app13095783
6. Wang S, Mo C, Chen Y, Dai X, Wang H, Shen X. Exploring the Performance of ChatGPT-4 in the Taiwan Audiologist Qualification Examination: Preliminary Observational Study Highlighting the Potential of AI Chatbots in Hearing Care. JMIR Med Educ. 2024;10:e55595. Published 2024 Apr 26. doi:10.2196/55595. PMID: 38693697
7. Kuroiwa T, Sarcon A, Ibara T, et al. The Potential of ChatGPT as a Self-Diagnostic Tool in Common Orthopedic Diseases: Exploratory Study. J Med Internet Res. 2023;25:e47621. Published 2023 Sep 15. doi:10.2196/47621. PMID: 37713254
8. Tian S, Jin Q, Yeganova L, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. 2023;25(1):bbad493. doi:10.1093/bib/bbad493. PMID: 38168838
9. Gödde D, Nöhl S, Wolf C, et al. A SWOT (Strengths, Weaknesses, Opportunities, and Threats) Analysis of ChatGPT in the Medical Literature: Concise Review. J Med Internet Res. 2023;25:e49368. Published 2023 Nov 16. doi:10.2196/49368. PMID: 37865883
10. Tsang R. Practical Applications of ChatGPT in Undergraduate Medical Education. J Med Educ Curric Dev. 2023;10:23821205231178449. Published 2023 May 24. doi:10.1177/23821205231178449. PMID: 37255525
11. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for Queries Related to Dementia and Other Cognitive Decline: Comparison of Results. J Med Internet Res. 2023;25:e48966. Published 2023 Jul 25. doi:10.2196/48966. PMID: 37490317
12. Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. Published online March 14, 2023. doi:10.1002/ase.2270. PMID: 36916887
13. Price T, Lynn N, Coombes L, et al. The International Landscape of Medical Licensing Examinations: A Typology Derived From a Systematic Review. Int J Health Policy Manag. 2018;7(9):782-790. Published 2018 Sep 1. doi:10.15171/ijhpm.2018.32. PMID: 30316226
14. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How Does ChatGPT Perform on the Italian Residency Admission National Exam Compared to 15,869 Medical Graduates?. Ann Biomed Eng. 2024;52(4):745-749. doi:10.1007/s10439-023-03318-7. PMID: 37490183
15. Aljindan FK, Al Qurashi AA, Albalawi IAS, 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. Published 2023 Sep 11. doi:10.7759/cureus.45043. PMID: 37829968
16. Armitage RC. Performance of Generative Pre-trained Transformer-4 (GPT-4) in Membership of the Royal College of General Practitioners (MRCGP)-style examination questions. Postgrad Med J. 2024;100(1182):274-275. doi:10.1093/postmj/qgad128. PMID: 38142282
17. Ebrahimian M, Behnam B, Ghayebi N, Sobhrakhshankhah E. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health Care Inform. 2023;30(1):e100815. Published 2023 Dec 11. doi:10.1136/bmjhci-2023-100815. PMID: 38081765
18. Fang C, Wu Y, Fu W, 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. Published 2023 Dec 1. doi:10.1371/journal.pdig.0000397. PMID: 38039286
19. Flores-Cohaila JA, García-Vicente A, Vizcarra-Jiménez SF, et al. Performance of ChatGPT on the Peruvian

National Licensing Medical Examination: Cross-Sectional Study. JMIR Med Educ. 2023;9:e48039. Published 2023 Sep 28. doi:10.2196/48039. PMID: 37768724

20. Garabet R, Mackey BP, Cross J, Weingarten M. ChatGPT-4 Performance on USMLE Step 1 Style Questions and Its Implications for Medical Education: A Comparative Study Across Systems and Disciplines. Med Sci Educ. 2023;34(1):145-152. Published 2023 Dec 27. doi:10.1007/s40670-023-01956-z. PMID: 38510401
21. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment [published correction appears in JMIR Med Educ. 2024 Feb 27;10:e57594]. JMIR Med Educ. 2023;9:e45312. Published 2023 Feb 8. doi:10.2196/45312. PMID: 36753318
22. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R Jr. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. Rev Assoc Med Bras (1992). 2023;69(10):e20230848. Published 2023 Sep 25. doi:10.1590/1806-9282.20230848. PMID: 37792871
23. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al. Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. Clin Pract. 2023;13(6):1460-1487. Published 2023 Nov 20. doi:10.3390/clinpract13060130. PMID: 37987431
24. Haze T, Kawano R, Takase H, Suzuki S, Hirawa N, Tamura K. Influence on the accuracy in ChatGPT: Differences in the amount of information per medical field. Int J Med Inform. 2023;180:105283. doi:10.1016/j.ijmedinf.2023.105283. PMID: 37931432
25. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. Digit Health. 2024;10:20552076241233144. Published 2024 Feb 16. doi:10.1177/20552076241233144. PMID: 38371244
26. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. PLOS Digit Health. 2023;2(12):e0000416. Published 2023 Dec 15. doi:10.1371/journal.pdig.0000416. PMID: 38100393
27. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted. Dtsch Arztebl Int. 2023;120(21):373-374. doi:10.3238/arztebl.m2023.0113. PMID: 37530052
28. Kao YS, Chuang WK, Yang J. Use of ChatGPT on Taiwan's Examination for Medical Doctors. Ann Biomed Eng. 2024;52(3):455-457. doi:10.1007/s10439-023-03308-9. PMID: 37432530
29. 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):536-538. doi:10.31662/jmaj.2023-0043. PMID: 37941716
30. Khorshidi, H., Mohammadi, A., Yousem, D. M., Abolghasemi, J., Ansari, G., Mirza-Aghazadeh-Attari, M., ... & Ardakani, A. A. (2023). Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination. *Informatics in Medicine Unlocked*, *41*, 101314. doi:10.1016/j.imu.2023.101314
31. 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. doi:10.5694/mja2.52061. PMID: 37528548
32. Kleinig O, Kovoor JG, Gupta AK, Bacchi S. Universal precautions required: Artificial intelligence takes on the Australian Medical Council's trial examination. Aust J Gen Pract. 2023;52(12):863-865. doi:10.31128/AJGP-02-23-6708. PMID: 38049136
33. Knoedler L, Alfertshofer M, Knoedler 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. Published 2024 Jan 5. doi:10.2196/51148. PMID: 38180782
34. 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. Published 2023 Sep 19. doi:10.3389/fmed.2023.1240915. PMID: 37795422
35. Lin SY, Chan PK, Hsu WH, Kao CH. Exploring the proficiency of ChatGPT-4: An evaluation of its performance in the Taiwan advanced medical licensing examination. Digit Health. 2024;10:20552076241237678. Published 2024 Mar 5. doi:10.1177/20552076241237678. PMID: 38449683
36. Meyer A, Riese J, Streichert T. Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. JMIR Med Educ. 2024;10:e50965. Published 2024 Feb 8. doi:10.2196/50965. PMID: 38329802

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):366-372. doi:10.1080/0142159X.2023.2249588. PMID: 37839017
38. Nakao T, Miki S, Nakamura Y, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study. JMIR Med Educ. 2024;10:e54393. Published 2024 Mar 12. doi:10.2196/54393. PMID: 38470459
39. Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: An observational study. Medicine (Baltimore). 2023;102(32):e34673. doi:10.1097/MD.0000000000034673. PMID: 37565917
40. Roos J, Kasapovic A, Jansen T, Kaczmarczyk R. Artificial Intelligence in Medical Education: Comparative Analysis of ChatGPT, Bing, and Medical Students in Germany. JMIR Med Educ. 2023;9:e46482. Published 2023 Sep 4. doi:10.2196/46482. PMID: 37665620
41. 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. Published 2023 Nov 22. doi:10.1038/s41598-023-46995-z. PMID: 37993519
42. Scaioli G, Lo Moro G, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam. Ann Ist Super Sanita. 2023;59(4):267-270. doi:10.4415/ANN_23_04_05. PMID: 38088393
43. Shang L, Xue M, Hou Y, Tang B. Can ChatGPT pass China's national medical licensing examination?. Asian J Surg. 2023;46(12):6112-6113. doi:10.1016/j.asjsur.2023.09.089. PMID: 37775381
44. 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. Published 2023 Jun 29. doi:10.2196/48002. PMID: 37384388
45. Tong W, Guan Y, Chen J, 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. Published 2023 Oct 19. doi:10.3389/fmed.2023.1237432. PMID: 38020160
46. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, et al. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. 2023;20:30. doi:10.3352/jeehp.2023.20.30. PMID: 37981579
47. 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 Inform. 2023;177:105173. doi:10.1016/j.ijmedinf.2023.105173. PMID: 37549499
48. Wang X, Gong Z, Wang G, et al. ChatGPT Performs on the Chinese National Medical Licensing Examination. J Med Syst. 2023;47(1):86. Published 2023 Aug 15. doi:10.1007/s10916-023-01961-0. PMID: 37581690
49. Watari T, Takagi S, Sakaguchi K, et al. Performance Comparison of ChatGPT-4 and Japanese Medical Residents in the General Medicine In-Training Examination: Comparison Study. JMIR Med Educ. 2023;9:e52202. Published 2023 Dec 6. doi:10.2196/52202. PMID: 38055323
50. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. 2023;86(8):762-766. doi:10.1097/JCMA.0000000000000946. PMID: 37294147
51. 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. Published 2023 Oct 13. doi:10.2196/48023. PMID: 37831496
52. Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT Performance on USMLE Sample Items and Implications for Assessment. Acad Med. 2024;99(2):192-197. doi:10.1097/ACM.0000000000005549. PMID: 37934828
53. Zhu Z, Ying Y, Zhu J, Wu H. ChatGPT's potential role in non-English-speaking outpatient clinic settings. Digit Health. 2023;9:20552076231184091. Published 2023 Jun 26. doi:10.1177/20552076231184091. PMID: 37434733
54. 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. Published 2024 Feb 14. doi:10.1186/s12909-024-05125-7. PMID: 38355517
55. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the Performance of ChatGPT Versions 3.5, 4, and 4 With Vision in the Chilean Medical Licensing Examination: Observational Study. JMIR Med Educ. 2024;10:e55048. Published 2024 Apr 29. doi:10.2196/55048. PMID: 38686550
56. Sharma, P., Thapa, K., Dhakal, P., Upadhaya, M. D., Adhikari, S., & Khanal, S. R. (2023). Performance of chatgpt on usmle: Unlocking the potential of large language models for ai-assisted medical education. *arXiv preprint arXiv:2307.00112*. doi:10.48550/arXiv.2307.00112. PMID: 36812645

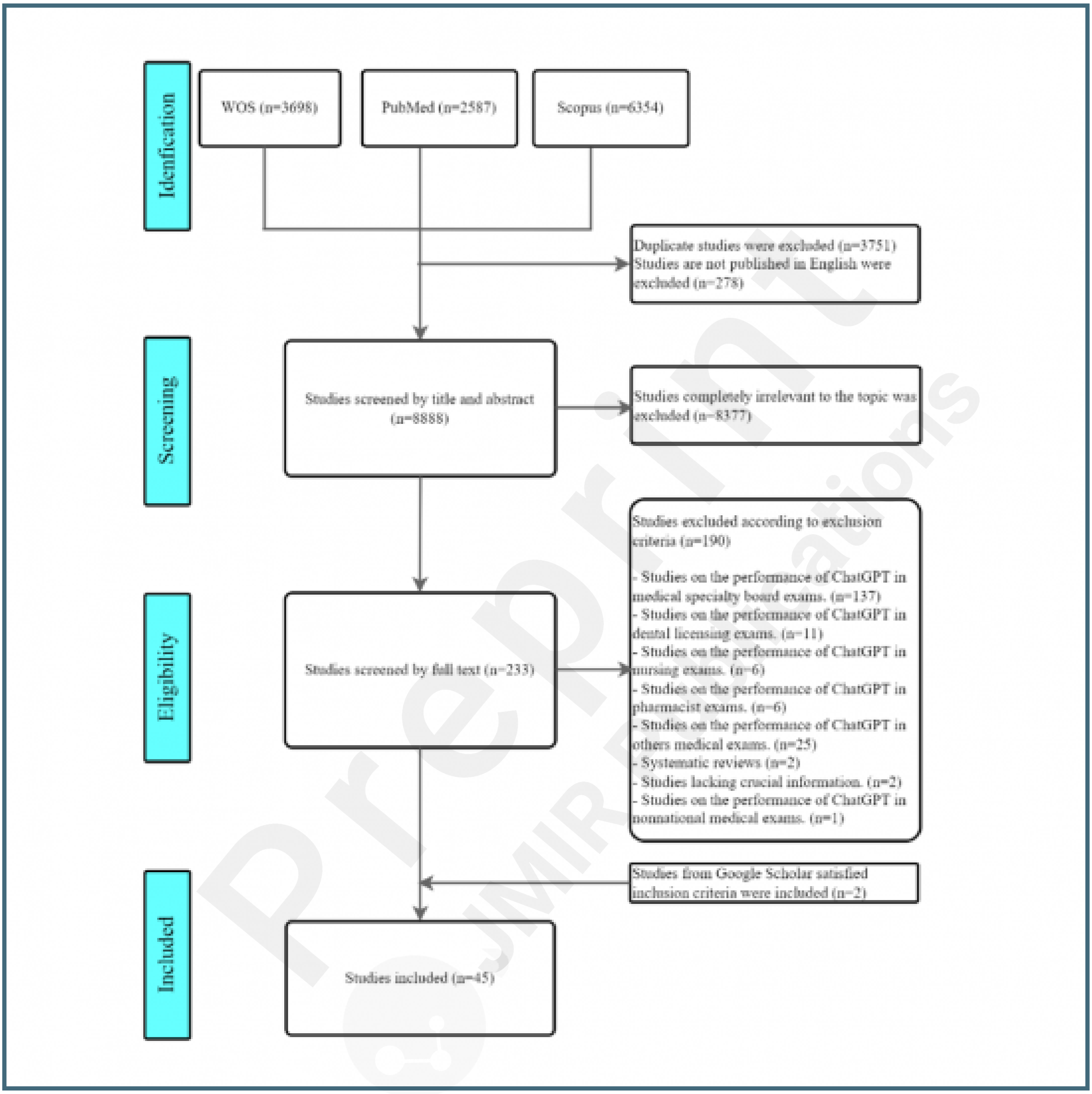
57. Keshtkar, A., Hayat, A. A., Atighi, F., Ayare, N., Keshtkar, M., Yazdanpanahi, P., ... & Hashempur, M. H. (2023). ChatGPT's Performance on Iran's Medical Licensing Exams. doi: https://doi.org/10.21203/rs.3.rs-3253417/v1
58. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: A systematic review and a meta-analysis. BJOG. 2024;131(3):378-380. doi:10.1111/1471-0528.17641. PMID: 37604703
59. 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;11:23821205241238641. Published 2024 Mar 13. doi:10.1177/23821205241238641. PMID: 38487300
60. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. J Biomed Inform. 2024;151:104620. doi:10.1016/j.jbi.2024.104620. PMID: 38462064
61. McInnes MDF, Moher D, Thombs BD, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement [published correction appears in JAMA. 2019 Nov 26;322(20):2026]. JAMA. 2018;319(4):388-396. doi:10.1001/jama.2017.19163. PMID: 29362800
62. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5(1):210. Published 2016 Dec 5. doi:10.1186/s13643-016-0384-4. PMID: 27919275
63. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009. PMID: 22007046
64. Heath A. All the news from OpenAI's first developer conference. The Verge. November 19, 2023. https://www.theverge.com/2023/11/6/23948619/openai-chatgpt-devday-developer-conference-news [accessed Apr 20, 2024]
65. Wong RS, Ming LC, Raja Ali RA. The Intersection of ChatGPT, Clinical Medicine, and Medical Education. JMIR Med Educ. 2023;9:e47274. Published 2023 Nov 21. doi:10.2196/47274. PMID: 37988149
66. Battineni G, Baldoni S, Chintalapudi N, et al. Factors affecting the quality and reliability of online health information. Digit Health. 2020;6:2055207620948996. Published 2020 Aug 30. doi:10.1177/2055207620948996. PMID: 32944269
67. He Z, Bhasuran B, Jin Q, et al. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study. J Med Internet Res. 2024;26:e56655. Published 2024 Apr 17. doi:10.2196/56655. PMID: 38630520
68. Choi W. Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: a study using MCQs. BMC Med Educ. 2023;23(1):864. Published 2023 Nov 13. doi:10.1186/s12909-023-04832-x. PMID: 37957666
69. Wu Y, Zheng Y, Feng B, Yang Y, Kang K, Zhao A. Embracing ChatGPT for Medical Education: Exploring Its Impact on Doctors and Medical Students. JMIR Med Educ. 2024;10:e52483. Published 2024 Apr 10. doi:10.2196/52483. PMID: 38598263
70. Zhang G, Jin Q, Jered McInerney D, et al. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. J Biomed Inform. 2024;153:104640. doi:10.1016/j.jbi.2024.104640. PMID: 38608915
71. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ. 2023;9:e48291. Published 2023 Jun 1. doi:10.2196/48291. PMID: 37261894

Abbreviations
AI: artificial intelligence
LLMs: large language models
USMLE: United States Medical Licensing Examination
MCQs: multiple-choice questions
NBME: National Board of Medical Examiners
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
CI: confidence interval

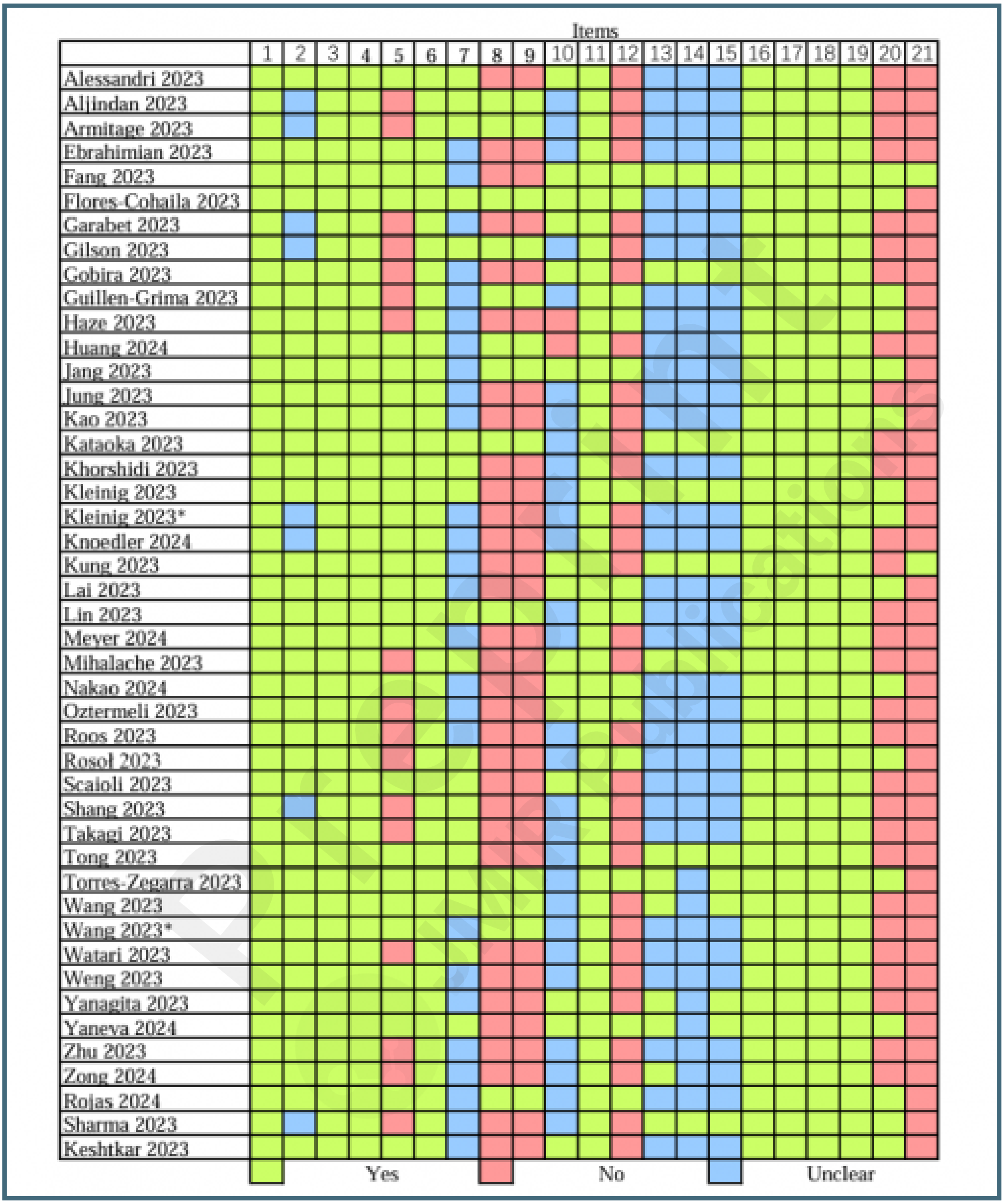
Supplementary Files

Figures

Prisma flow diagram.



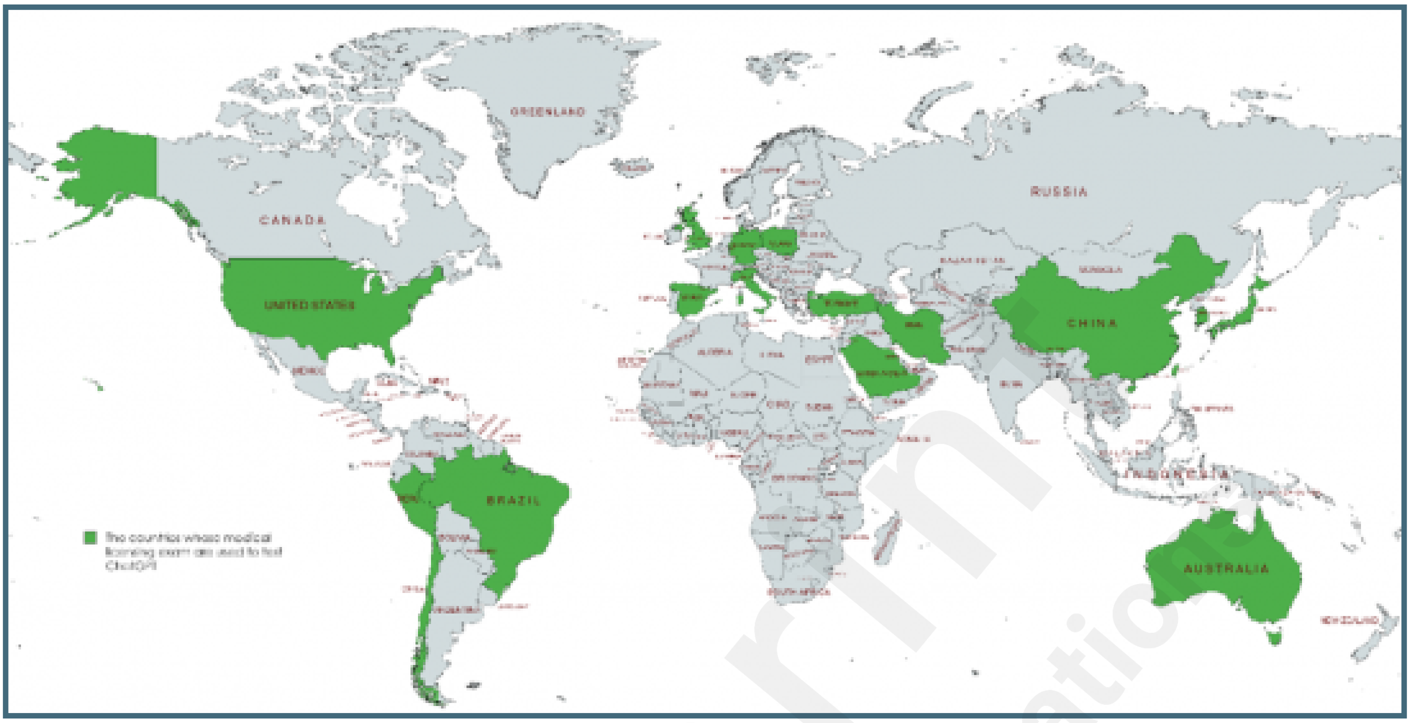
Quality assessment of included studies using evaluation framework.



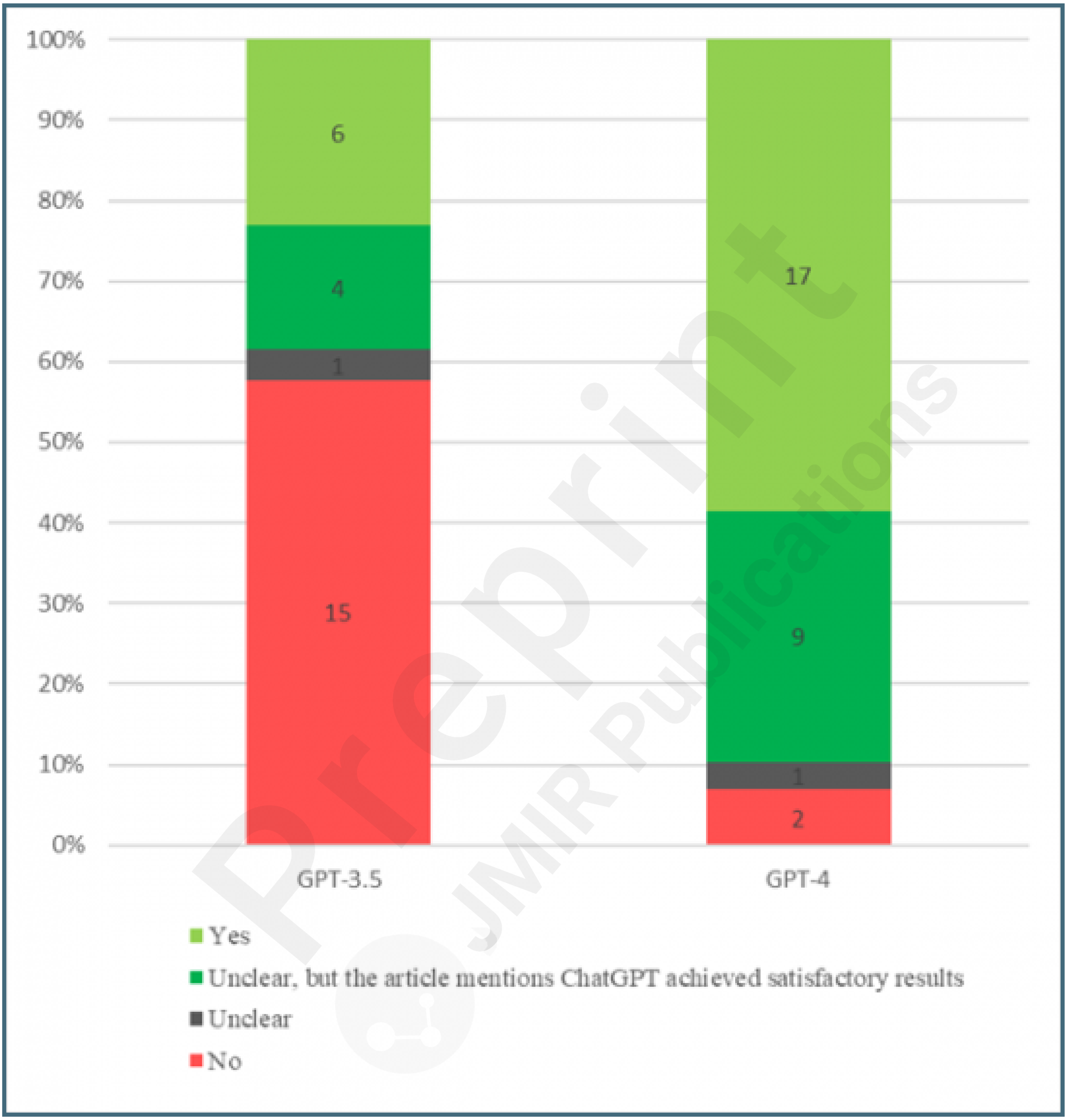
Risk of bias.

Risk of Bias				
Study	Patient Selection	Index Test	Reference Standard	Flow and Timing
Alessandri 2023	Low	Low	Low	Low
Aljindan 2023	Unclear	Unclear	Low	Low
Armitage 2023	Unclear	Unclear	Low	Low
Ebrahimian 2023	Low	Unclear	Low	Unclear
Fang 2023	Low	Low	Low	Unclear
Flores-Cohaila 2023	Low	Low	Low	Low
Garabet 2023	Unclear	Low	Low	High
Gilson 2023	Unclear	Unclear	Low	Low
Gobira 2023	Low	Low	Low	Unclear
Guillen-Grima 2023	Low	Unclear	Low	Unclear
Haze 2023	Low	High	Low	Unclear
Huang 2024	Low	High	Low	Unclear
Jang 2023	Low	Low	Low	Unclear
Jung 2023	Low	Unclear	Low	Unclear
Kao 2023	Low	Unclear	Low	Unclear
Kataoka 2023	Low	Unclear	Low	Low
Khorshidi 2023	Low	Unclear	Low	Low
Kleinig 2023	Low	Unclear	Low	Low
Kleinig 2023*	Low	Unclear	Low	High
Knoedler 2024	Low	Unclear	Low	Unclear
Kung 2023	Low	Low	Low	Unclear
Lai 2023	Low	Low	Low	Unclear
Lin 2023	Low	Unclear	Low	Low
Meyer 2024	Low	Unclear	Low	Unclear
Mihalache 2023	Low	Low	Low	Low
Nakao 2024	Low	Low	Low	Unclear
Oztermeli 2023	Low	Low	Low	Unclear
Roos 2023	Low	Unclear	Low	Unclear
Rosoł 2023	Low	Unclear	Low	Low
Scaioli 2023	Low	Low	Low	Low
Shang 2023	Unclear	Unclear	Low	Low
Takagi 2023	Low	Unclear	Low	Low
Tong 2023	Low	Unclear	Low	Low
Torres-Zegarra 2023	Low	Unclear	Low	Low
Wang 2023	Low	Unclear	Low	Low
Wang 2023*	Low	Unclear	Low	Unclear
Watari 2023	Low	Unclear	Low	Low
Weng 2023	Low	Unclear	Low	Unclear
Yanagita 2023	Low	Low	Low	Unclear
Yaneva 2024	Low	Low	Low	Low
Zhu 2023	Low	Unclear	Low	Unclear
Zong 2024	Low	Unclear	Low	Unclear
Rojas 2024	Low	Unclear	Low	Unclear
Sharma 2023	Unclear	Unclear	Low	High
Keshtkar 2023	Low	Low	Low	Low

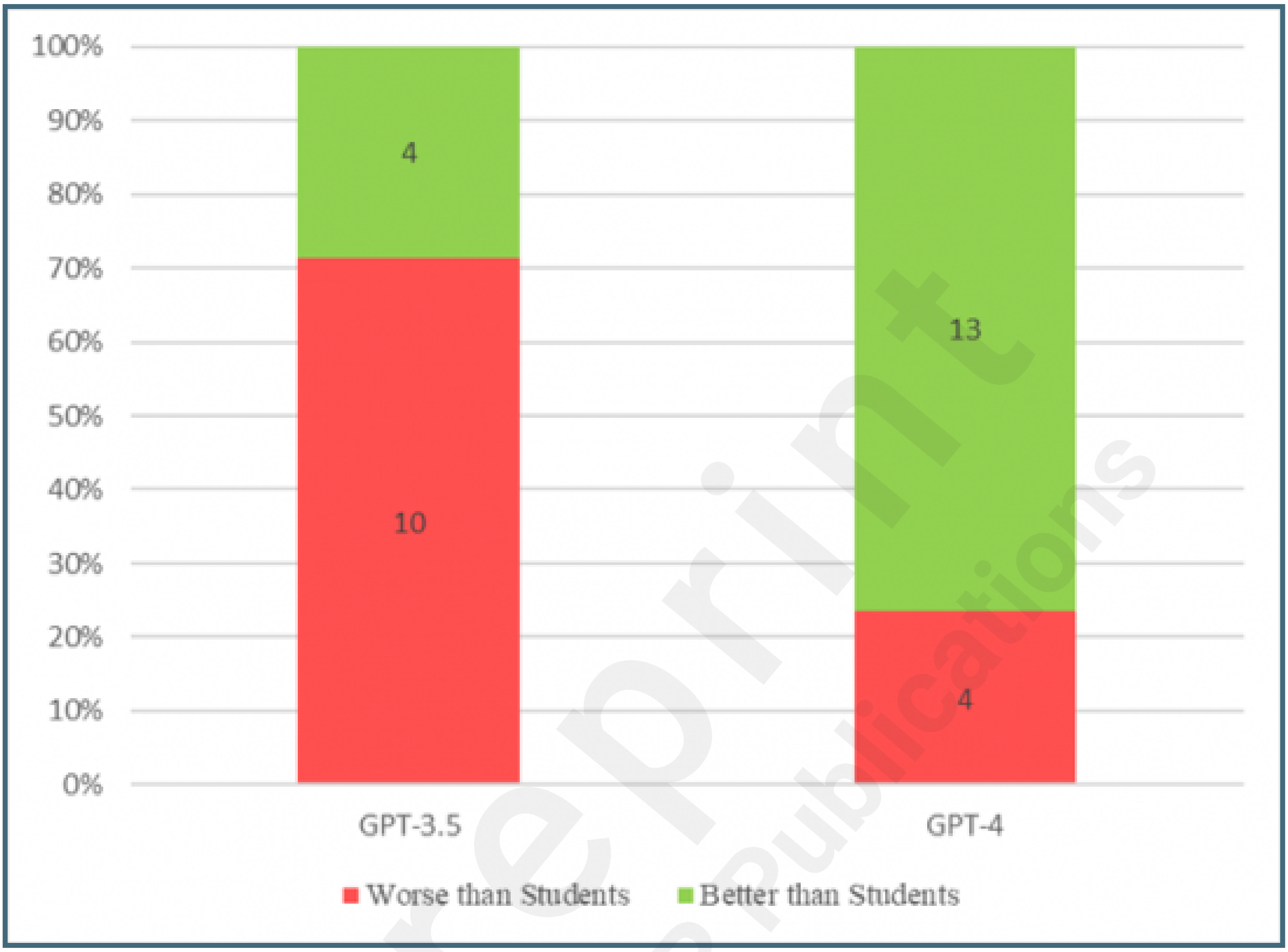
Countries which medical licensing exam have been used to test ChatGPT.



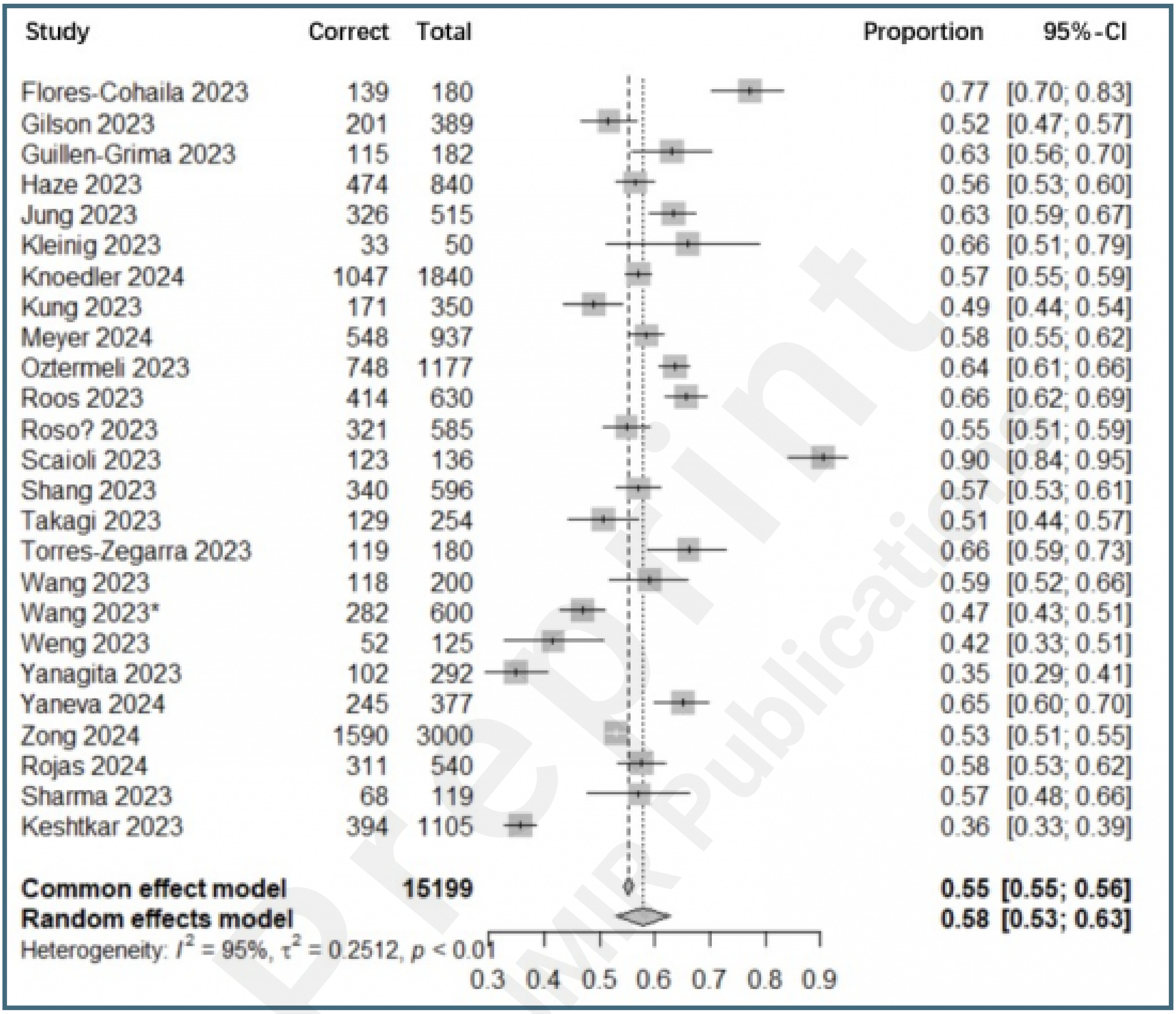
Performance of ChatGPT on passing medical licensing exam.



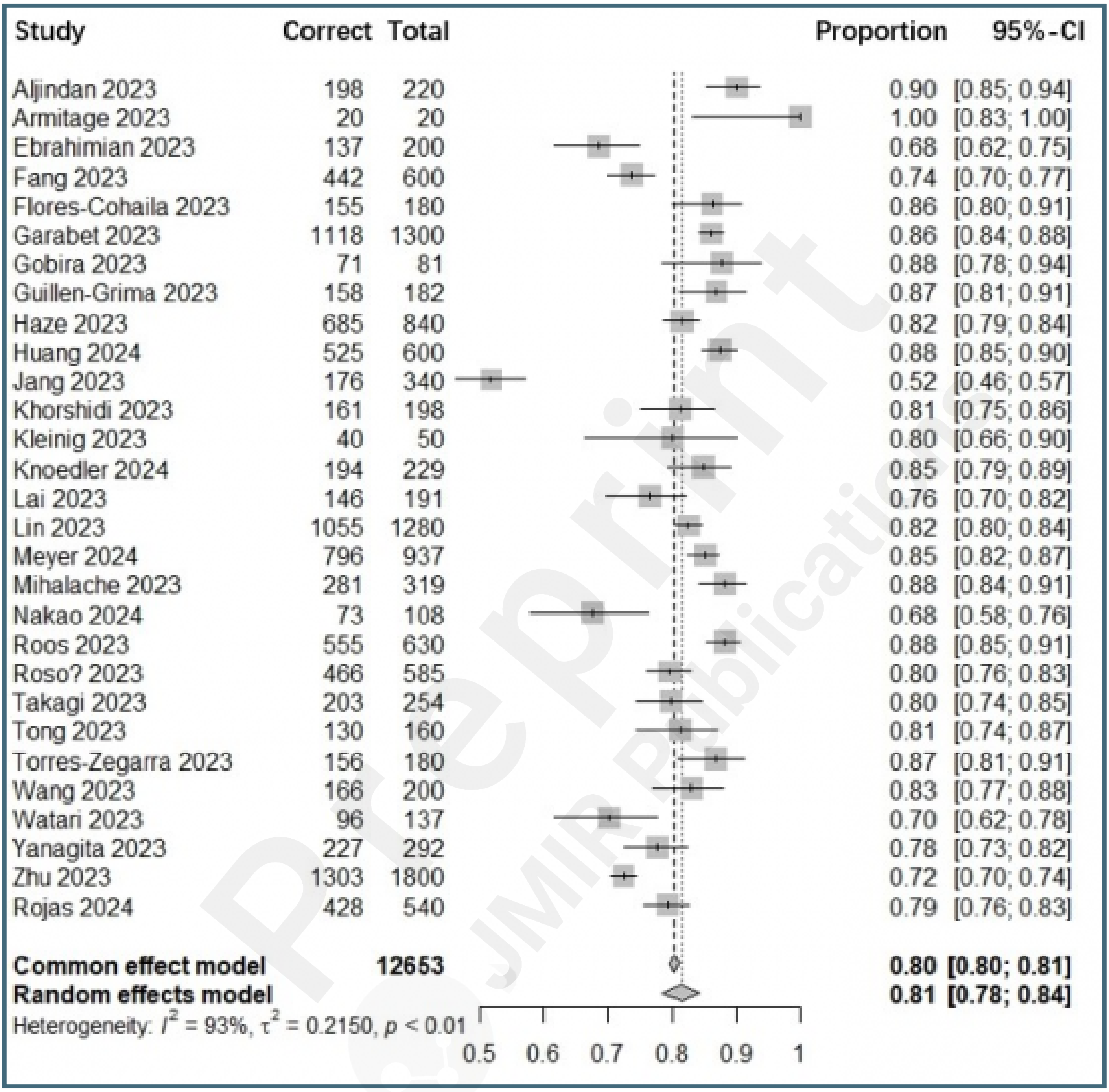
Performance of ChatGPT compared with medical students.



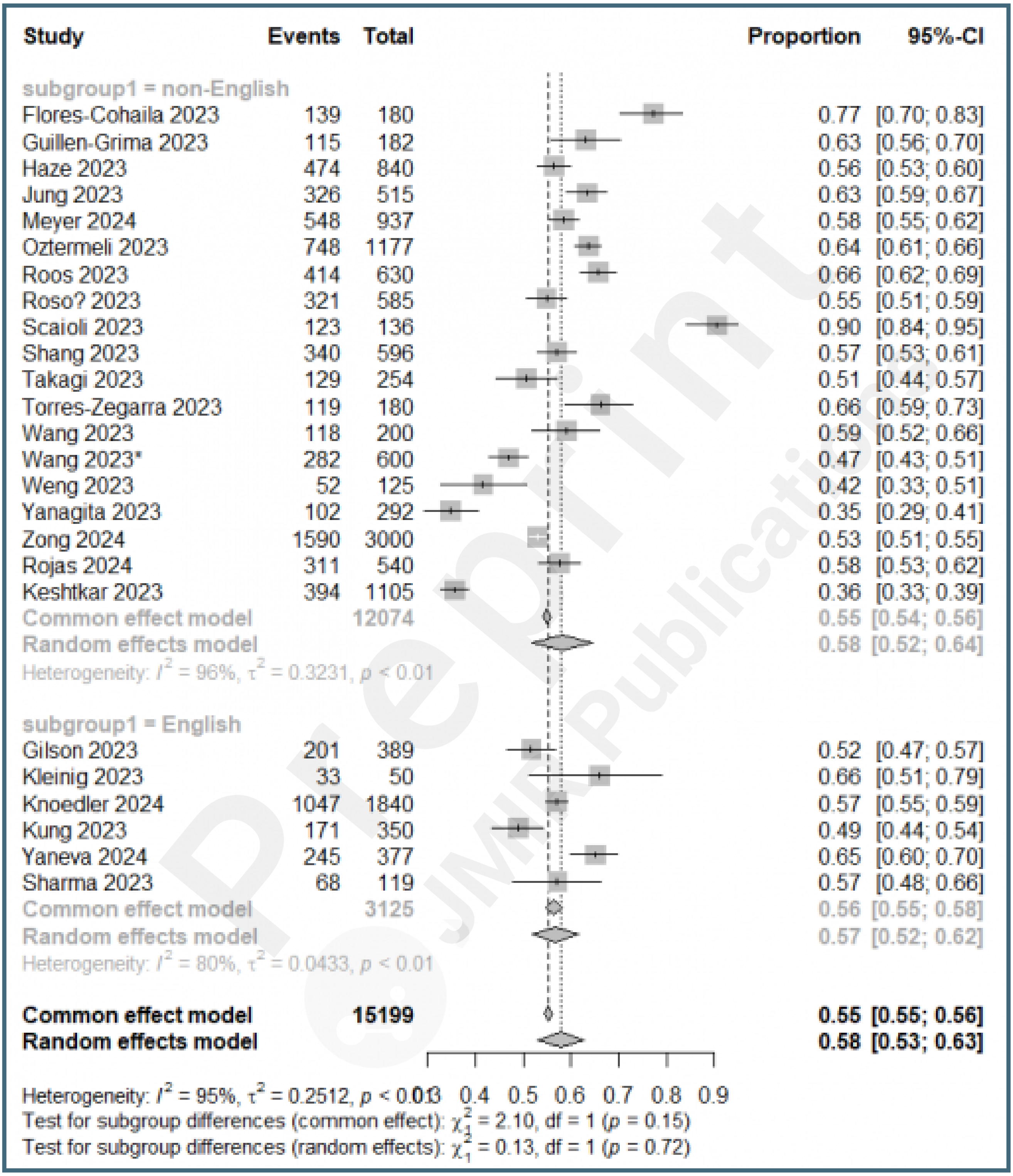
Performance of ChatGPT-3.5 in medical licensing exam.



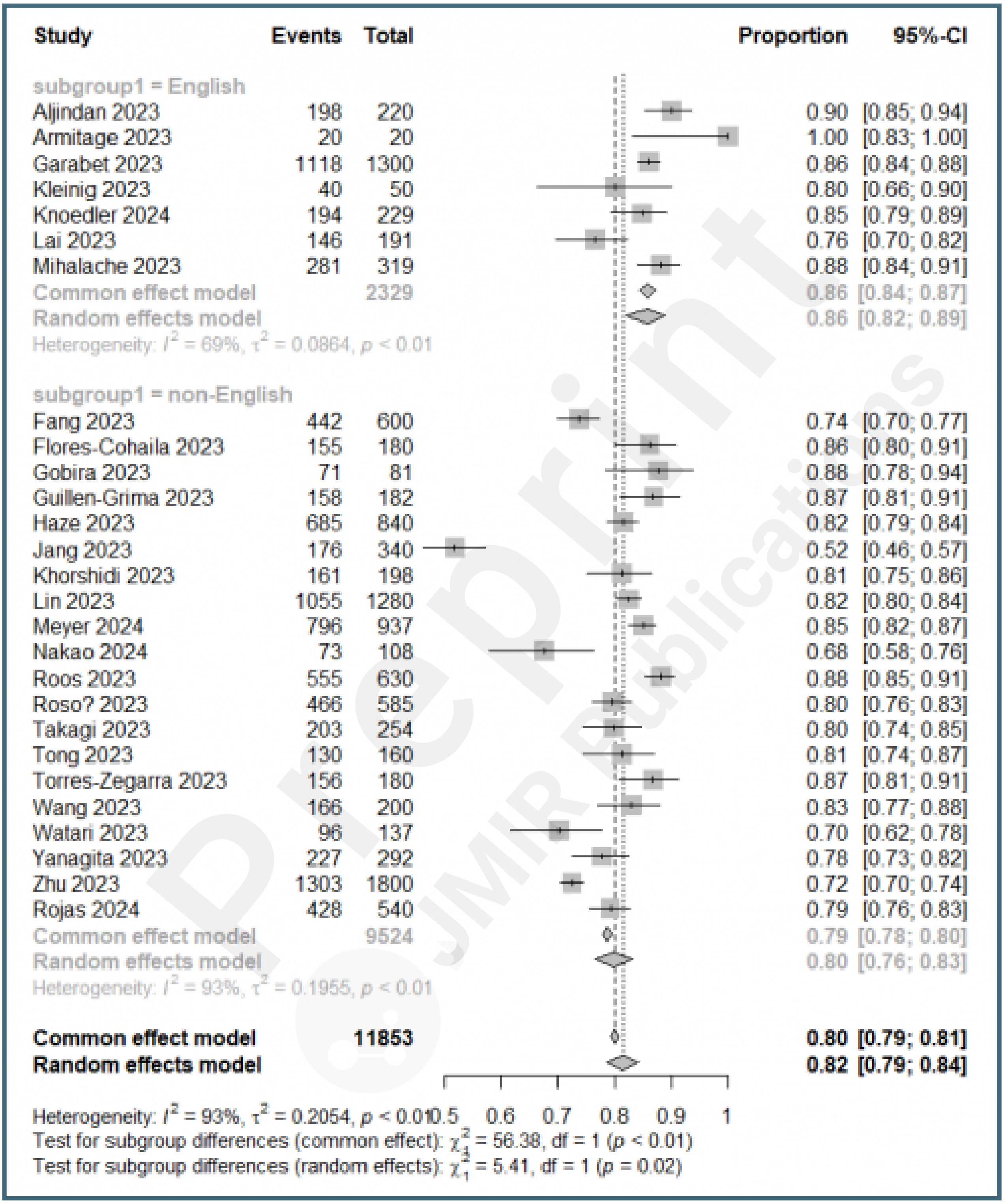
Performance of ChatGPT-4 in medical licensing exam.



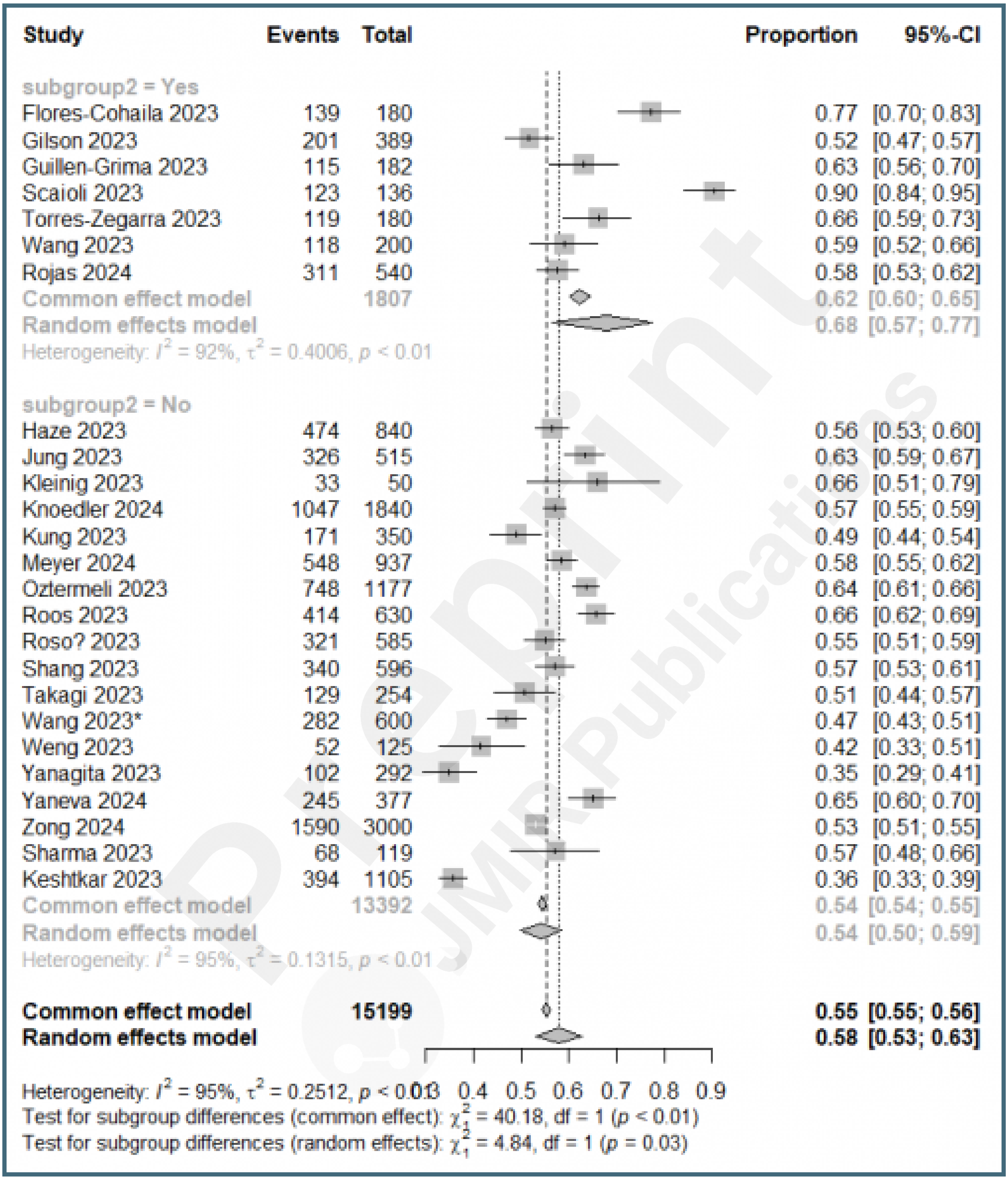
Subgroup 1: Performance of GPT-3.5 on medical licensing exam from English-speaking countries and non-English-speaking countries.



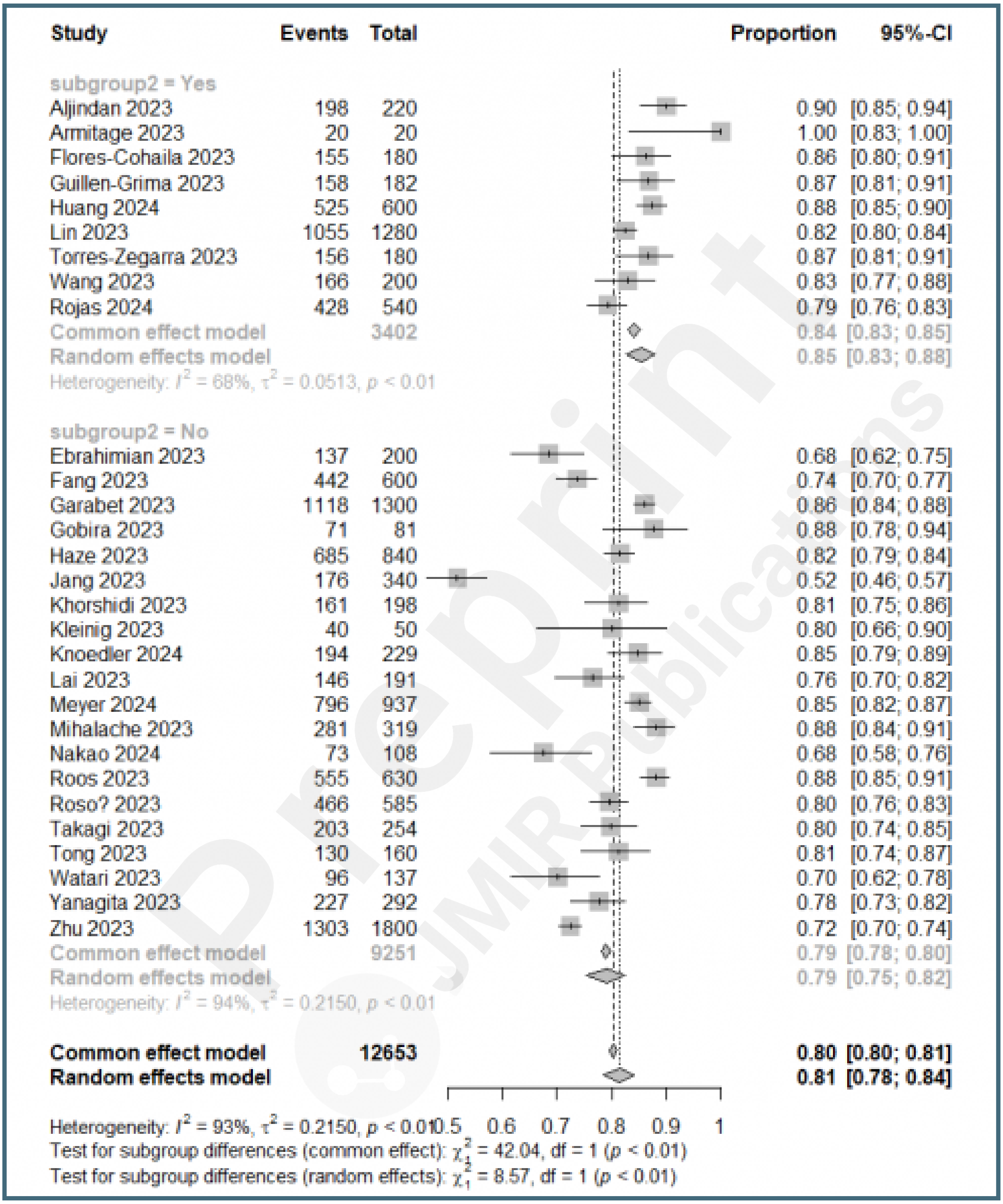
Subgroup 1: Performance of GPT-4 on medical licensing exam from English-speaking countries and non-English-speaking countries.



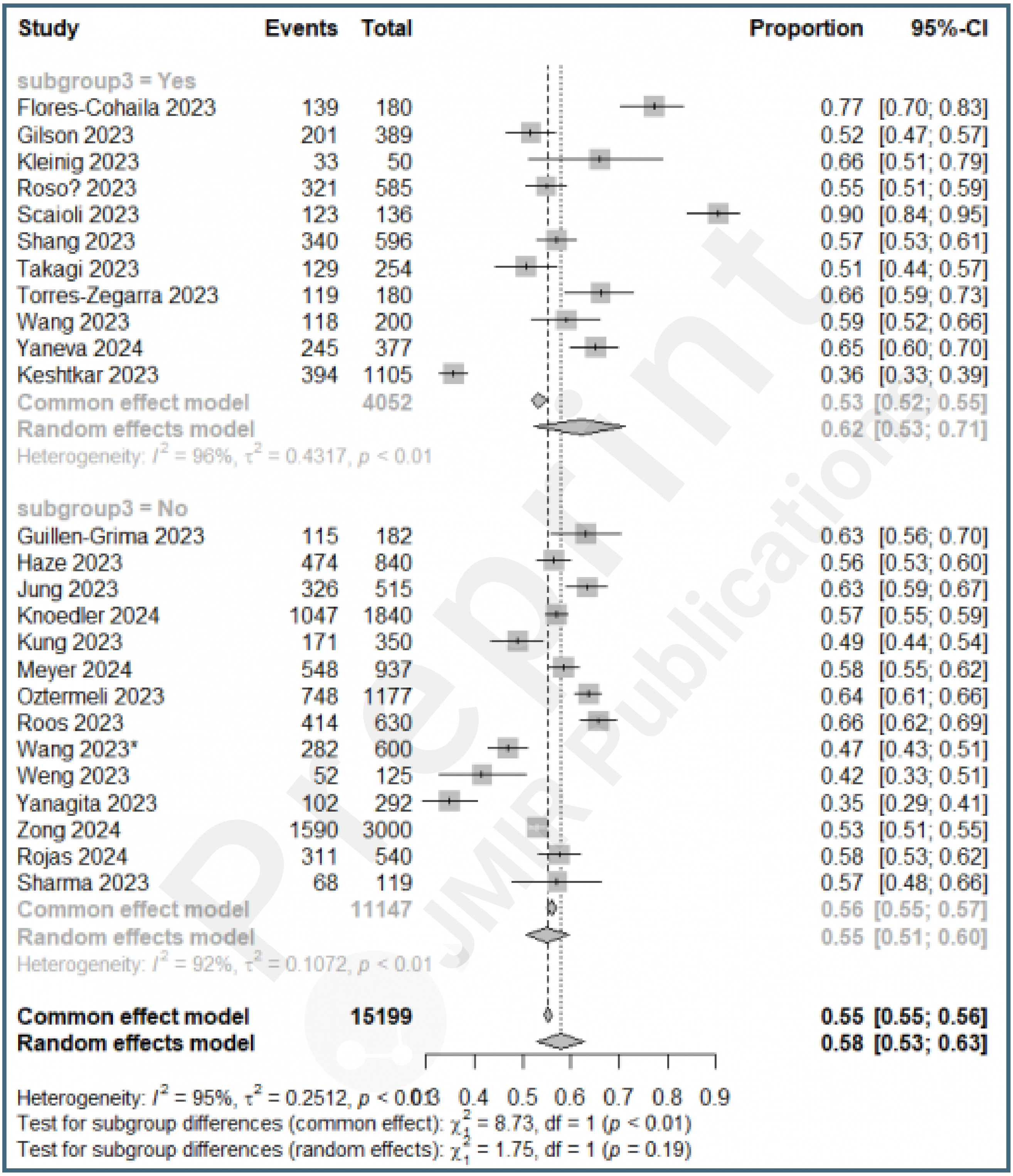
Subgroup 2: Performance of GPT-3.5 with or without prompts.



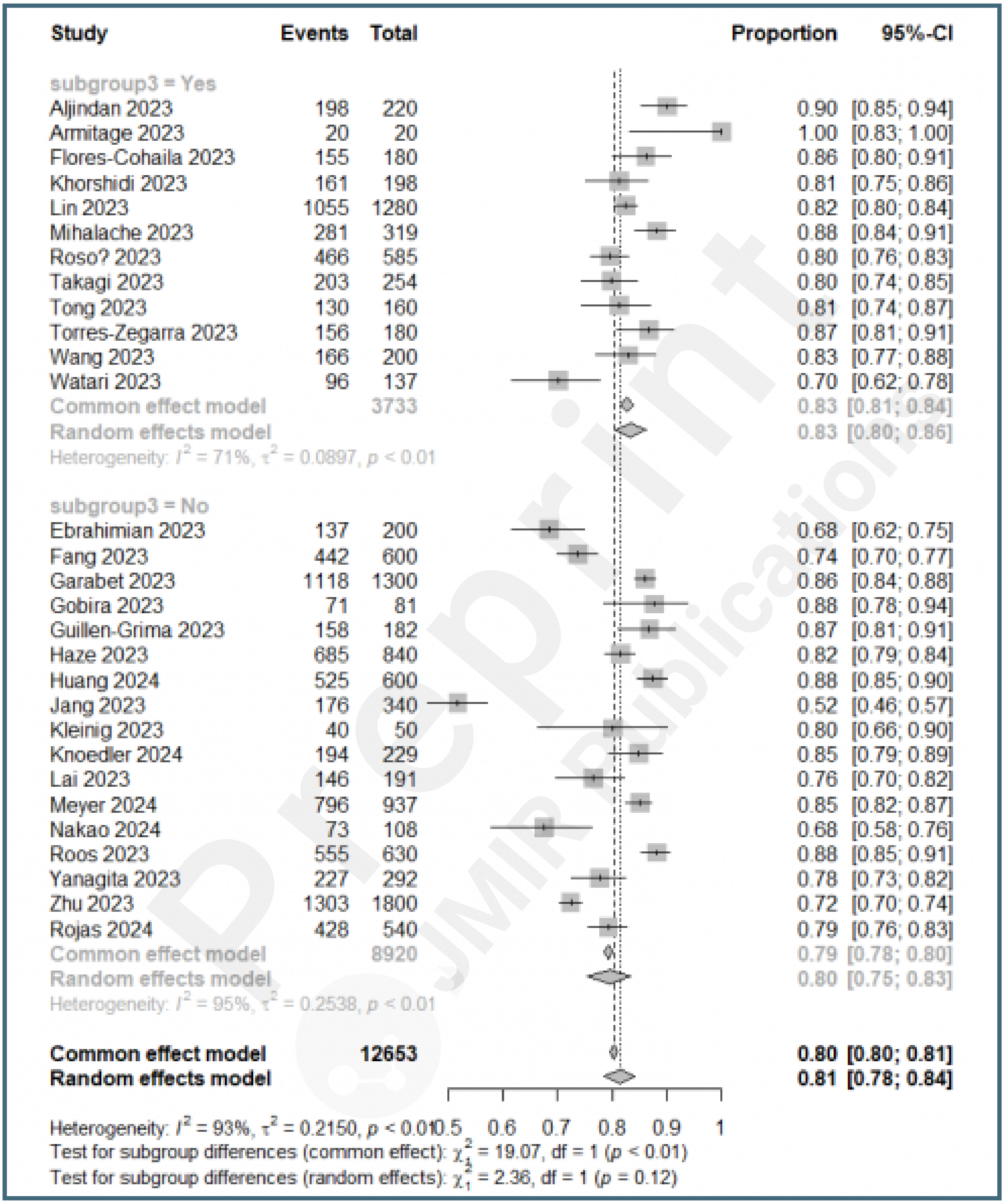
Subgroup 2: Performance of GPT-4 with or without prompts.



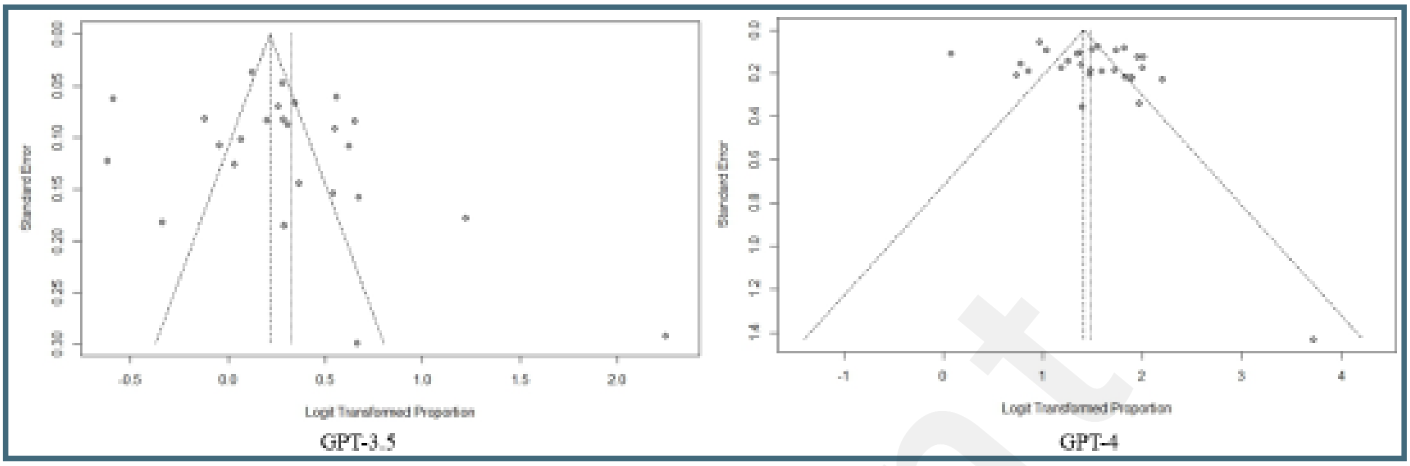
Subgroup 3: Performance of GPT-3.5 regarding “Flow and Timing”.



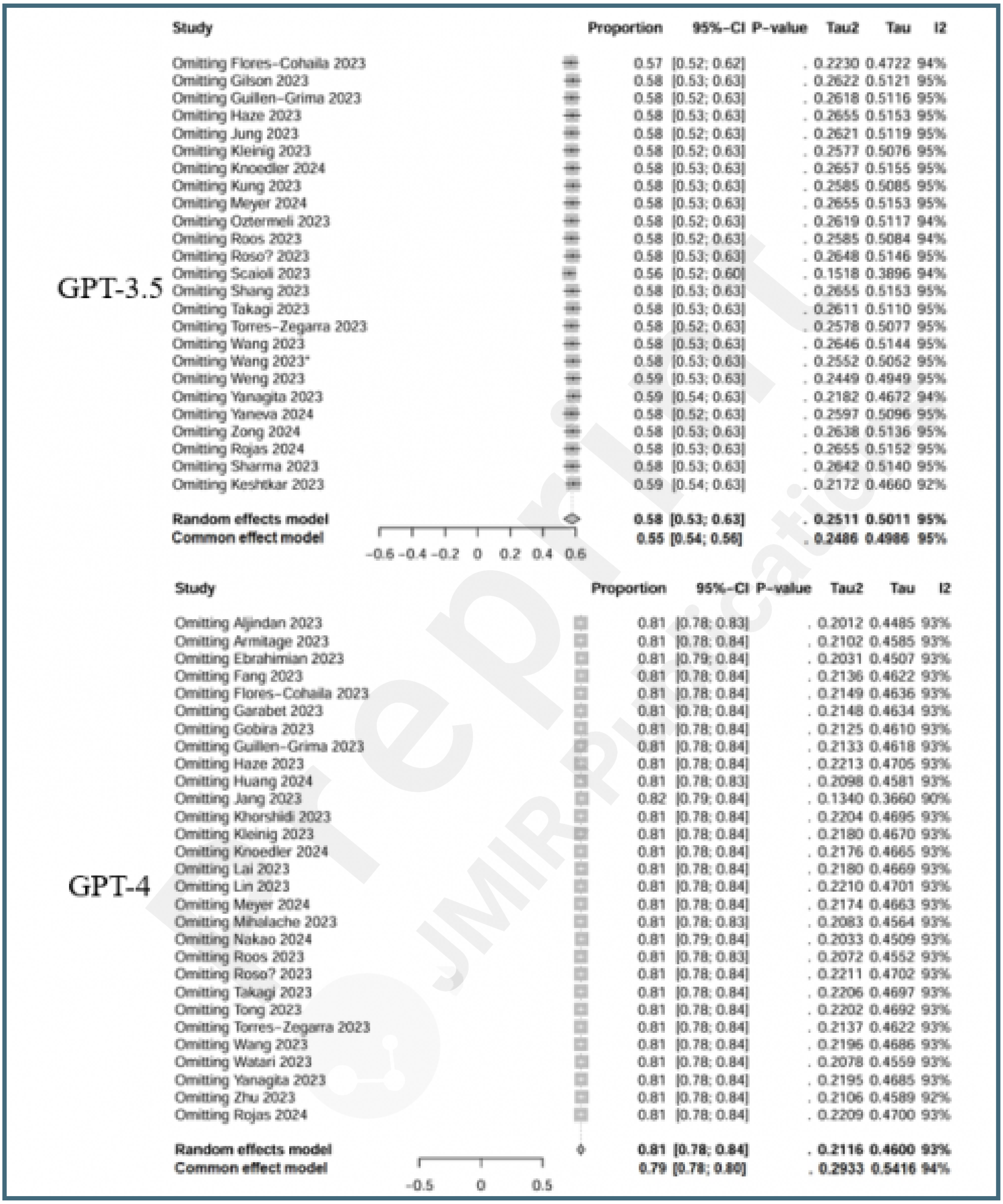
Subgroup 3: Performance of GPT-4 regarding “Flow and Timing”.



Funnel plot of included studies reported GPT-3.5 and GPT-4 performance.



Sensitivity analyses result of performance of GPT-3.5 and GPT-4.



Multimedia Appendixes

Query strings of WOS, Scopus, and PubMed.
URL: http://asset.jmir.pub/assets/3929fea8fc61b9a1d411a8bb6c49adc5.docx

Evaluation framework used in this systematic review.
URL: http://asset.jmir.pub/assets/45feba22d89d78d3bf743aa95a337f01.docx

General characteristics of included studies.
URL: http://asset.jmir.pub/assets/f25145071a78f515a24c8ca00032c1a4.xlsx

PRISMA 2020 checklist.
URL: http://asset.jmir.pub/assets/efc63803abf7bebbdfc8ed8ad506eaa7.docx