ChatGPT as a global doctor: a rapid review of its performance on national licensing medical examination

Javier Alejandro Flores Cohaila, Brayan Miranda-Chavez, Javier Flores-Arocutipa, Percy Mayta-Tristan

Submitted to: JMIR Medical Education
on: August 29, 2024

Table of Contents

ChatGPT as a global doctor: a rapid review of its performance on national licensing medical examination

Javier Alejandro Flores Cohaila[1] MD; Brayan Miranda-Chavez[2] MD; Javier Flores-Arocutipa[3] PhD; Percy Mayta-Tristan[4] MD

[1]Grupo de Investigación en Healthcare Simulation & Medical Education (HeSIM) Facultad de Ciencias de la Salud Universidad Científica del Sur Lima PE
[2]Centro de Investigación de Educación Médica y Bioética - EDUCAB-UPT Facultad de Ciencias de la Salud Universidad Privada de Tacna Tacna PE
[3]Universidad Nacional de Moquegua Moquegua PE
[4]Grupo de Investigación en Healthcare Simulation & Medical Education (HeSIM). Facultad de Ciencias de la Salud Universidad Científica del Sur Lima PE

Corresponding Author:
Javier Alejandro Flores Cohaila MD
Grupo de Investigación en Healthcare Simulation & Medical Education (HeSIM)
Facultad de Ciencias de la Salud
Universidad Científica del Sur
C. Cantuarias 385, Miraflores 15074
Lima
PE

Abstract

Background: The growth of studies evaluating ChatGPT's performance in exams swamped the medical education community. However, it has been proved from low to high-stakes examination, affecting the reliability and validity of findings. To ensure reliability and bring a final consensus, we opted to synthesize the evidence of ChatGPT's performance under high-stakes examinations, namely, National Licensing Medical Examinations (NLME).

Objective: To evaluate ChatGPT’s NLMEs performance and assess whether it could achieve a license to practice in various countries.

Methods: We searched the Pubmed and Scopus databases for studies that evaluated ChatGPT's performance in NLMEs. In addition to the reference list and in Google Scholar. Studies were screened, and the accuracy rate (performance) of ChatGPT was extracted, as well as other study characteristics.

Results: We identified 37 studies that evaluated ChatGPT's performance across 18 NLMEs. Most studies evaluated the performance of ChatGPT in the NLME of the United States, China, and Japan. While the majority of studies used official datasets, others used unofficial ones from third parties, and a scarce number of studies used prompting techniques. GPT-4 was superior to GPT-3.5 in all NLMEs and could pass all of them. GPT-4 overperformed the average performance of examinees' in most studies, except the Japan NLME.

Conclusions: Current evidence suggests that ChatGPT can pass 18 NLMEs, surpassing almost all candidates, and, if possible, receive a "global medical license." Further research should move towards using ChatGPT as GPT-4o in performance assessment and exploring the potential of ChatGPT for NLMEs development and validation. Moreover, our findings represent a call for reimagining assessment in medical education. Clinical Trial: Not applicable

(JMIR Preprints 29/08/2024:63194)
DOI: https://doi.org/10.2196/preprints.63194

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

Review

ChatGPT as a global doctor: a rapid review of its performance on national licensing medical examination

Abstract

Introduction: The growth of studies evaluating ChatGPT's performance in exams swamped the medical education community. However, it has been proved from low to high-stakes examination, affecting the reliability and validity of findings. To ensure reliability and bring a final consensus, we opted to synthesize the evidence of ChatGPT's performance under high-stakes examinations, namely, National Licensing Medical Examinations (NLME).
Objective: To evaluate ChatGPT’s NLMEs performance and assess whether it could achieve a license to practice in various countries.
Methods: We searched the Pubmed and Scopus databases for studies that evaluated ChatGPT's performance in NLMEs. In addition to the reference list and in Google Scholar. Studies were screened, and the accuracy rate (performance) of ChatGPT was extracted, as well as other study characteristics.
Results: We identified 37 studies that evaluated ChatGPT's performance across 18 NLMEs. Most studies evaluated the performance of ChatGPT in the NLME of the United States, China, and Japan. While the majority of studies used official datasets, others used unofficial ones from third parties, and a scarce number of studies used prompting techniques. GPT-4 was superior to GPT-3.5 in all NLMEs and could pass all of them. GPT-4 overperformed the average performance of examinees' in most studies, except the Japan NLME.
Conclusions: Current evidence suggests that ChatGPT can pass 18 NLMEs, surpassing almost all candidates, and, if possible, receive a "global medical license." Further research should move towards using ChatGPT as GPT-4o in performance assessment and exploring the potential of ChatGPT for NLMEs development and validation. Moreover, our findings represent a call for reimagining assessment in medical education.

Keywords: Medical education, assessment, artificial intelligence, ChatGPT, national licensing examination

Introduction

The performance of ChatGPT on medical exams has led to the claim that Artificial Intelligence (AI) could replace doctors. However, these exams differ in importance, as some were from low to high stakes [1]. While low-stakes exams cannot ensure a trainee's readiness for practice, high-stakes exams like National Licensing Medical Examinations (NLMEs) can [2]. NLMEs, mostly written or performance-based assessments, are currently the most reliable and valid tools for ensuring adequate patient care and readiness to practice [3].
Although previous studies have reported that ChatGPT can pass NLMEs [4–6], a concise summary of its performance across all available NLMEs is lacking. To fill this gap, we conducted a rapid review to evaluate ChatGPT’s NLMEs performance and assess whether it could achieve a license to practice in various countries.

Methods

A rapid review [7] was conducted to evaluate the performance of ChatGPT on NLME. This was in accordance with the Preferred Reporting Items for Systematic Review and Meta-analysis reporting guidelines [8].

Research questions

The following question guided this review: What is the performance of ChatGPT in versions GPT-3.5 and GPT-4 in National Licensing Medical Examinations? Additionally, we seek to answer two secondary questions: Can ChatGPT pass NLMEs worldwide? and Is ChatGPT's performance superior to the average performance of examinees across NLMEs?

Eligibility criteria

Studies were included if: 1) reported ChatGPT's performance regarding the number of correct questions, and 2) evaluated ChatGPT performance in a written National Licensing Medical Examination. We define NLMEs as "large-scale exams that are required to practice or to continue training in a national jurisdiction". Hence, if a study described the exam as an NLME, it was included. The base list of NLMEs to be included was retrieved from the study of Pierce and colleagues [3]. However, this was not restrictive, as we considered including NLMEs that were not in their study. The list of NLMEs is available as supplementary material.

Search strategy

A systematic search on the PubMed and Scopus databases was conducted on June 1st, 2024. The search strategy included terms referencing ChatGPT and licensing examinations. There were no language restrictions. Additionally, we actively searched for NLMEs (previously described) in the first ten pages of Google Scholar and manually searched included studies' reference lists.
Retrieved studies were imported into RayyanAI [9], and duplicates were deleted. Two of us (JFC and BMC) did title and abstract screening. Potentially eligible studies were reviewed in full text by the same two authors. In case of discrepancies, these were solved through discussion until a consensus was reached.

Data extraction

The data was extracted by one of us (JFC) and revised by a second author (BMC). The extracted data included: the type of study, author, year, country, and examination, ChatGPT's version, dataset's characteristics, the prompt techniques used, ChatGPT's performance, the pass score of the NLME, and the examinees' performance.
For ChatGPT's performance, we extracted GPT-3.5 and GPT-4 scores if available. Due to expected differences in datasets (number of questions), we decided to extract the percentage of correct answers [1]. If a study reported two or more attempts per NLME, the average number of correct answers was extracted. This score was used to determine if ChatGPT passed or failed an exam. Lastly, to provide an estimate per NLMEs, the higher performance of ChatGPT per NLME was chosen.

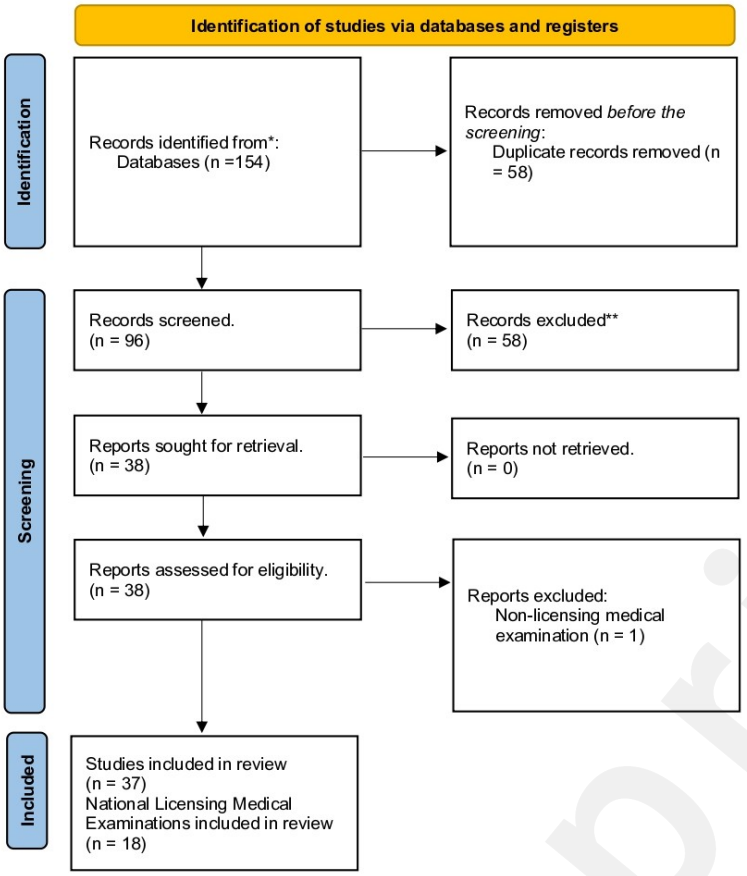
Data analysis

A meta-analysis was not performed due to expected high heterogeneity. The main reason for this decision was that NLMEs are designed for a specific country, which may impact the content covered and the format of items [10]. Moreover, under modern psychometrics, the same scores in an examination can not be treated as equal [11]. Hence, data was presented in tabular and narrative formats.

Results

The search yielded 154 studies, of which 37 studies and 18 NLMEs were included. The selection flowchart is shown in [Figure 1].

Figure 1. PRISMA flowchart



[Table 1] shows the key characteristics of the included studies. Most studies evaluated the ChatGPT's performance in the USMLE [4,12–18], the Japanese NLME [19–24], and the Chinese NLME [25–29]. Most studies employed an official dataset; for example, all studies on the Japanese and Peruvian exams used official items provided by the organizer [5,19–24,30]. Contrarily, all USMLE-related studies relied on unofficial/third-party items. Most studies reported that minimal prompts, such as zero-shot or no prompts, were used. When used, the most common prompting technique was "Persona'[5,19,30–33], where researchers asked ChatGPT to assume a role [34]. Other used prompt techniques were step-by-step prompts [5,30], where researchers gave a detailed list of tasks to perform[34], and self-consistency [35], an approach where researchers stimulate diverse reasoning paths in LLMs and select the most appropriate.

Table 1. Key characteristics of included studies

Author, year	Dataset: N (Official Exam) Language - Exclusions	Prompt	ChatGPT performance: Accuracy rate (%)	Pass score	Average examinee accuracy rate
Australia Licensing Examination					
Kleinig, 2023	50 (Unofficial) - English - None	Minimal or Zero Shot	GPT-3.5: 66% GPT-4: 79.30%	N.R.	N.R.

Kleinig, 2023	50 (Unofficial) - English - None	Minimal or Zero Shot.	GPT-3.5: 58%	N.R.	N.R.
Belgium Medical Examination					
Morrell, 2024	97 (Official) - Translated to English - Tables and Images	Minimal or Zero Shot	GPT-3.5: 67% GPT-4: 76%	60%	61%
Chile Single National Medical Knowledge Examination					
Rojas, 2024	540 (Unofficial) - N.R. - None	Minimal or Zero Shot	GPT-3.5: 57.53% GPT-4: 79.32%	51%	N.R.
Chinese National Licensing Medical Examination					
Fang, 2023	600 (Unofficial) - Chinese - None	Minimal or Zero Shot	GPT-4: 73.67%	60%	68.7%
Shang, 2023	600 (Unofficial) - Chinese - None	Minimal or Zero Shot	GPT-3.5: 57%	60%	N.R.
Tong, 2023	160 (Official) - English and Chinese - N.R.	Minimal or Zero Shot	GPT-4: 81.25% in Chinese GPT-4: 86.25% in English	60%	N.R.
Wang, 2023	1800 (Official) - English and Chinese - None	Minimal or Zero Shot	GPT-3.5: 34.1%	60%	N.R.
Zong, 2024	2400 (Official) - N.R. - None	Minimal or Zero Shot	GPT-3.5: 53.05%	60%	N.R.
Peruvian National Medical Examination					
Flores-Cohaila, 2024	180 (Official) - Spanish - None	Persona and step-by-step	GPT-3.5: 77% GPT-4: 86%	55%	55%
Torres-Zegarra, 2023	180 (Official) - Spanish - None	Persona and step-by-step	GPT-3.5: 66.7% GPT-4: 86.7%	55%	N.R.
French medical licensing examination					
Alfertshofer, 2024	300 (Unofficial) - N.R. - N.R.	Minimal or Zero Shot	GPT-3.5: 22%	N.R.	N.R.
Germany Staatsexamen					

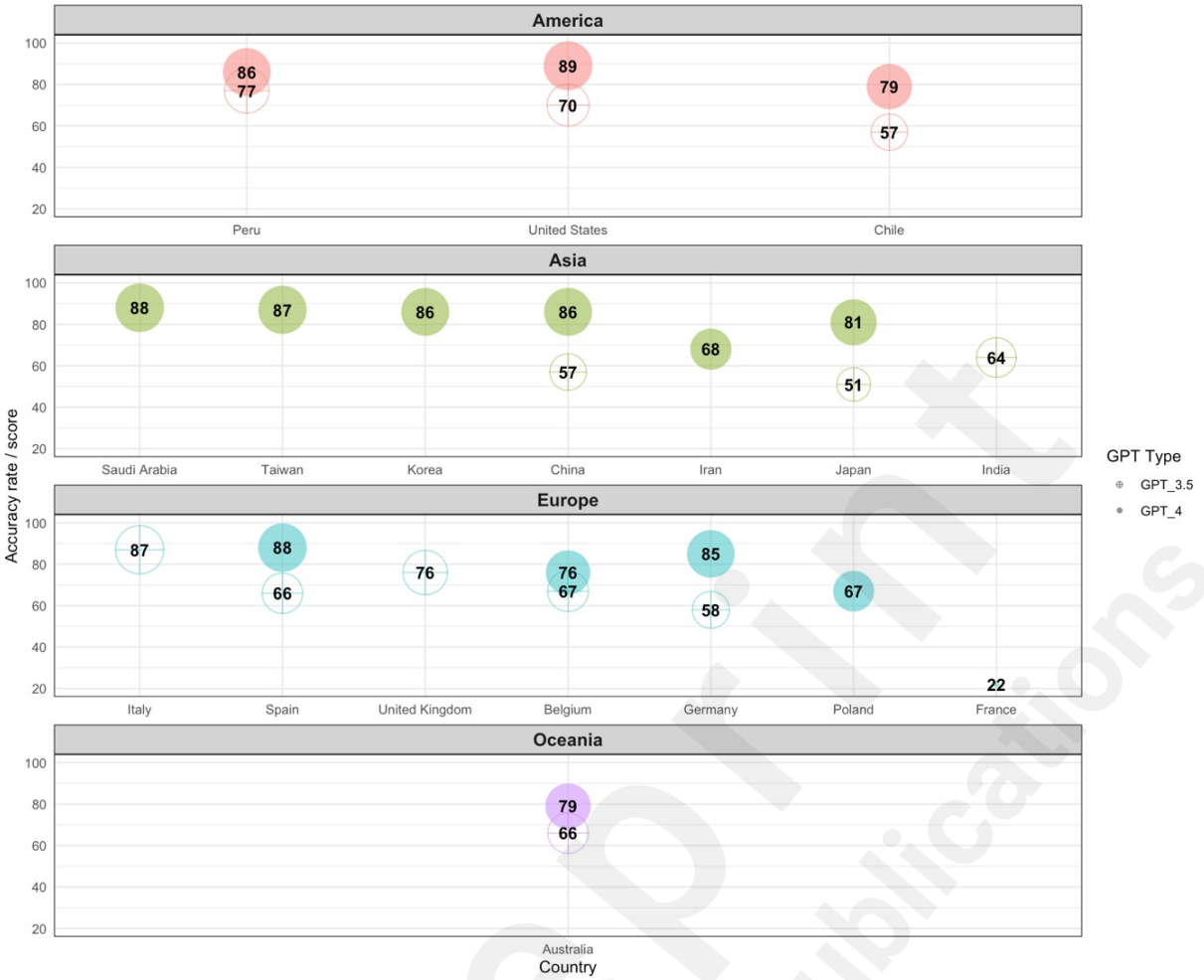
Meyer, 2024	835 (Unofficial) - N.R. - Tables and images	Minimal or Zero Shot	GPT-3.5: 58% GPT-4: 85%	60%	77%
Indian Licensing Examination					
Alfertshofer, 2024	300 (Unofficial) - N.R. - N.R.	Minimal or Zero Shot	GPT-3.5: 64%	N.R.	N.R.
Iranian Pre-Internship Medical License Examination					
Alfertshofer, 2024	200 (Official) - English - None	Minimal or Zero Shot	GPT-4: 68.5%	45%	57%
Italy National Licensing Medical Examination					
Alfertshofer, 2024	300 (Unofficial) - N.R. - N.R.	Minimal or Zero Shot	GPT-3.5: 73%	N.R.	N.R.
Bonetti, 2023	140 (Official) - N.R - None	Minimal or Zero Shot	GPT-3.5: 87%	N.R.	N.R.
Japan National Licensing Medical Examination					
Kawahara, 2024	3532 (Official) - Japanese - None	Persona	GPT-4/4VV: 67.4%	80%	N.R.
Nakao, 2024	108 Image Questions (Official) - Japanese	Minimal or Zero Shot	GPT-4: 68%	N.A.	N.R.
Takagi, 2023	254 (Official) - Japanese	Minimal or Zero Shot	GPT-3.5: 50.8% GPT-4: 79.9%	80%	84.9%
Tanaka, 2024	262 (Official) - English/Japanese - None	Optimized prompt	GPT-4: 79.8%	80%*	N.R.
Yanagita, 2023	292 (Official) - Japanese - Tables and images	Minimal or Zero Shot	GPT-3.5: 42.8% GPT-4: 81.5%	80%	N.R.
Kataoka, 2023	281 (Official) - Japanese - Tables and Images	Minimal or Zero Shot	GPT-3.5: 38%	80%	N.R.
Korea Medical Licensing Examination					
Jang, 2023	340 (Official) - English/Korean - None	Self-consistency	GPT-4: 86.18%	60%	76.7%

Poland National Licensing Medical Examination					
Wójcik, 2023	120 (Official) - English - None	Minimal or Zero Shot	GPT-4: 67.1%	60%	N.R.
Saudi Arabia National Licensing Medical Examination					
Aljindan, 2023	220 (Unofficial) - N.R. - None	Persona	GPT-4: 88.6%	N.R.	N.R.
Spain Medical Intern-Resident Exam					
Cerame, 2024	210 (Official) - N.R. - None	Minimal or Zero Shot	GPT-3.5 51.4% GPT-4: 82.4%	N.R.	N.R.
Guillen-Crima, 2024	182 (Official) - Spanish/English - Tables and images	Minimal or Zero Shot	GPT-3.5: 66.48% (English) and 63.18% (Spanish) GPT-4: 87.91% (English) and 86.81% (Spanish)	N.R.	N.R.
Taiwanese National Licensing Medical Examination					
Huang, 2024	600 (Official) - English/Chinese) - None	Persona	GPT-4: 87.5%	60%	N.R.
Lin, 2024	1280 (Unofficial) - Chinese - None	Persona	GPT-4: 82.5%	60%	N.R.
United Kingdom Licensing Examination: Professional and Linguistic Assessments Board					
Alfertshofer, 2024	300 (Unofficial) - English - None	Minimal or Zero Shot	GPT-3.5: 68%	N.R.	N.R.
Lai, 2023	91 (Unofficial) - English - Tables and images	Minimal or Zero Shot	GPT-3.5: 76.3%	70%	N.R.
United States Medical Licensing Examination					
Alfertshofer, 2024	300 Step 2 CK (Unofficial) - English - None	Minimal or Zero Shot	GPT-3.5: 53.50%	60%	N.R.
Garabet, 2024	1300 Step 1 (Unofficial) -	Minimal or Zero Shot	GPT-4: 86%	60%	N.R.

	English - None				
Gilson, 2023	389 (Unofficial) - English - Tables and images	Minimal or Zero Shot	GPT-3.5: 50% (Step 1) and 50.5% (Step 2 CK)	60%	N.R.
Kung, 2023	350 (Unofficial) - English - None	Minimal or Zero Shot	GPT-3.5: 75% (Step 1), 61.5% (Step 2 CK), 68.8% (Step 3[1])	60%	N.R.
Mackey, 2024	900 Step 2CK (Unofficial) - English - None	Minimal or Zero Shot	GPT-4: 89%	60%	N.R.
Mihalache, 2024	319 (Unofficial) - English - None	Minimal or Zero Shot	GPT-4: 88% (Step 1), 86% (Step 2 CK), and 90% (Step 3[1])	60%	N.R.
Shieh, 2024	109 Step 2CK (Unofficial) - English - Tables and Images	Minimal or Zero Shot	GPT-3.5: 47.7% GPT-4: 87.2%	60%	N.R.
Yaneva, 2024	377 (Unofficial) - English - None	Minimal or Zero Shot	GPT-3.5: 63.06% (Step 1), 70% (Step 2 CK), and 60.34% (Step 3[1])	60%	N.R.

N.R. - Not Reported; N.A. - Not applicable; GPT- Generative Pretrained Transformer
Green cell - GPT did pass the exam. Red cell - GPT did not pass the exam.
*GPT-4 passed the examination, scoring 82.7% in essential questions and 77.2% in basic clinical. The pass score is 80% on essential and 74.6% on basic clinical.
CK - Clinical Knowledge
[1] Step 3 only comprised the written part

[Figure 2] depicts ChatGPT's performance in all NLMEs. GPT-3.5 accuracy rate ranged from 22% [14] to 87% [36], averaging around 50%, while GPT-4 ranged from 75% to 89% [15], except for the Iranian [37] and Polish NLMEs [38]. Hence, as shown in Table 1, GPT-4 passes all NLMEs, while GPT-3.5 does not.
Figure 2. Performance of ChatGPT in versions GPT-3.5 and GPT-4



There were examinees' scores for seven NLMEs. Compared to the examinees' performance, GPT-3.5 outperformed them in the Belgian, Peruvian, and Iranian NLMEs [5,37,39]. However, GPT-4 surpassed all candidates except for Japan [21], where the average performance of examinees was 84.9% versus 81.5% of ChatGPT.

Discussion

Summary of findings

Here, we synthesized the evidence on the performance of ChatGPT across 18 NLMEs. Our major findings are as follows: 1) GPT-4 was superior to GPT-3.5 in all NLMEs, with a performance above 75%; 2) GPT-4 passed all NLMEs; hence, if possible, it gained a license to practice worldwide; and 3) GPT-4 was superior to the average examinees' performance of all NLMEs, except for the Japan NLME.

Limitations and strengths

The main limitation of this review was the heterogeneity among studies. First, the datasets used vary even at the NLME level. Some studies employed official and unofficial datasets, which may reduce the reliability of the tests. Exams from different years and different numbers of items were used. Second, while most studies comprised similar constructs, such as medical knowledge and clinical reasoning, there was little to no reporting on the validation of this large-scale examination. Third, the strategies to obtain the

responses from ChatGPT vary. Some studies employed advanced prompting techniques such as self-consistency, while others did not use prompts. This leaves open the possibility that ChatGPT's performance may be underestimated as prompting techniques to enhance ChatGPT's output [40].

Strengths in this review include the concise research questions, the inclusion of only NLMEs as they are more rigorous than low-to-medium-stakes exams, the use of a guidance document to identify NLMEs, and the peer-review process during the review.

Implications

Our study builds upon the findings of a previous review that evaluates ChatGPT performance among different medical examinations [1]. However, in their seminal study, the authors combined the results from classroom-based examinations with licensing examinations, and from a validity perspective, this is not feasible. To offer a more valid stance and evidence on ChatGPT's performance, we decided only to include NLMEs. Therefore, the most proximal implication is the approach to categorize the performance of ChatGPT under different stakes, specialties, and formats, as this comparison is more feasible than mixing oranges with apples.

Only 8 out of the 21 NLMEs described in 2018 were found [3], with the addition of 10 NLMEs not previously described. This suggests a growing expansion of NLMEs and reflects the increasing focus on quality assurance [41]. The lack of studies on some NLMEs may indicate that the exam items are not publicly available. This conjecture is supported by the fact that almost half of the studies included in our review were conducted on unofficial datasets. Although this poses challenges for researchers, evidence suggests that not disclosing NLME items is important for maintaining test reliability [42,43].

The stable performance of GPT-4 across all NLMEs opens the space for hypotheses. While NLMEs vary in their development process, their intended purpose is the same: to ensure adequate competency for patient care [3]. Hence, under a validity stance, NLMEs may share similarities [44]. Most NLMEs use multiple-choice question format to assess medical knowledge, clinical reasoning, and specific knowledge for the country. Moreover, studies outside the US reported that specific knowledge of the country accounted for more mistakes [5,19,25]. The reason for this remains in ChatGPT nature, as it was trained with human-generated data, and most of it is in the English language [45]. However, it also highlights similarities in the contents assessed in the medical knowledge and clinical reasoning sections. The latter may suggest the existence of a shared mental model on what constitutes readiness to practice across countries.

Almost all examinees were surpassed by ChatGPT, except the Japanese ones. At first glance, it may suggest that ChatGPT is superior to doctors, but on deeper examination, it appears to be more of an educational problem [46]. In NLMEs where ChatGPT outperformed examinees, a similarity was observed: a low cut-off score ranging from 50-60%. In those NLMEs, the examinees' scores were slightly above the passing mark. Previous studies have indicated that examinees' motivation to pass summative assessments results in a low effort and that high-stakes exams promote surface approaches to learning [46]. In the case of Japan NLME, the cut-off score was 80%, and the examinee's average score was 84%. This points toward basic human behavior, as if the passing score for an exam is relatively low, it may seem unnecessary for all examinees to strive for top scores. This requires us to re-look at how we establish standard

settings [47], and for NLMEs, we may need to integrate the public perspective in this process. Although with contrasting results, recent surveys on public perspective on AI support this claim, with some suggesting that the public had more confidence in a diagnosis given by AI than of a doctor [48,49]. This may be augmented with more research suggesting a triumph of AI over doctors. However, as we have shown, it appears to be a problem with cut-off scores. Hence, we need to revisit how we establish passing scores in NLMEs.

This review offers several directions for future research. We encourage researchers to move forward with ChatGPT's performance on written examinations toward performance assessment, as it has shown promising findings but requires further exploration [50]. With the recent release of GPT-4o with voice and vision, this seems feasible. The differences and stability between GPT-3.5 and GPT-4 performance may be helpful for assessment purposes. This can serve as a benchmarking tool to define if an NLME is suitable for its purpose. It can also be used for validation purposes, as different versions can be used to take examinations and then, with their scores, calculate the reliability, difficulty, and discrimination indexes. Hence ensuring the privacy of items. Outside the realm of ChatGPT, we suggest that researchers explore the similarities among NLMEs under a validity framework such as Rusell's [44]. For example, the declared purpose, the time when they are conducted, the relevance toward continuing practicing, or the format of items, as if ChatGPT performance remains similar, there may be hidden similarities across NLMEs.

Lastly, the remarkable performance of ChatGPT reveals more about the limitations of our current assessment methods than the capabilities of the assessed individuals. Therefore, this calls for reimagining assessment in medical education. Embrace assessment as a system through authentic and programmatic assessment, leaving behind summative and knowledge-oriented traditions. Integrating competency-based assessment, such as objective structure clinical examination or workplace-based assessment, into NLMEs can capture essential competencies for this new era, such as communication, teamwork, professionalism, or clinical reasoning. This may require changing towards a programmatic assessment through the employment of milestones, as it allows stakeholders to witness the growth and development of trainees' competencies over time. However, our current assessment methods go far beyond these recommendations. Like the famous quote, "Don't judge a fish by its inability to climb a tree," you cannot judge a doctor by his ability to pass a one-time test.

Conclusions

In conclusion, current evidence suggests that ChatGPT can pass 18 NLMEs, surpassing almost all candidates. Further research should move towards using ChatGPT as GPT-4o in performance assessment and exploring ChatGPT's potential for NLMEs development and validation. Moreover, our findings represent a call for reimagining assessment in medical education.

Acknowledgments

None

Conflict of interest

None declared

Multimedia appendix

Multimedia Appendix 1: List of NLMEs and search strategy
Multimedia Appendix 2: Excel extraction sheet

References

[1] Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: A systematic review and a meta-analysis. BJOG Int J Obstet Gynaecol 2024;131:378–80. https://doi.org/10.1111/1471-0528.17641.
[2] Swanson DB, Roberts TE. Trends in national licensing examinations in medicine. Med Educ 2016;50:101–14. https://doi.org/10.1111/medu.12810.
[3] Price T, Lynn N, Coombes L, Roberts M, Gale T, Regan De Bere S, et al. The International Landscape of Medical Licensing Examinations: A Typology Derived From a Systematic Review. Int J Health Policy Manag 2018;7:782–90. https://doi.org/10.15171/ijhpm.2018.32.
[4] 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.
[5] 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.
[6] 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. https://doi.org/10.2196/55048.
[7] Haby MM, Barreto JOM, Kim JYH, Peiris S, Mansilla C, Torres M, et al. What are the best methods for rapid reviews of the research evidence? A systematic review of reviews and primary studies. Res Synth Methods 2024;15:2–20. https://doi.org/10.1002/jrsm.1664.
[8] 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.
[9] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016;5:1–10. https://doi.org/10.1186/s13643-016-0384-4.
[10] Schurter T, Escher M, Gachoud D, Bednarski P, Hug B, Kropf R, et al. Essential steps in the development, implementation, evaluation and quality assurance of the written part of the Swiss federal licensing examination for human medicine. GMS J Med Educ 2022;39:Doc43. https://doi.org/10.3205/zma001564.
[11] De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ 2010;44:109–17. https://doi.org/10.1111/j.1365-2923.2009.03425.x.
[12] 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 2024;34:145–52. https://doi.org/10.1007/s40670-023-01956-z.
[13] 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.
[14] Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the Seven Seas: A Multinational Comparison of ChatGPT's Performance on Medical Licensing Examinations. Ann Biomed Eng 2024;52:1542–5. https://doi.org/10.1007/s10439-023-03338-3.
[15] Mackey BP, Garabet R, Maule L, Tadesse A, Cross J, Weingarten M. Evaluating ChatGPT-4

in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. Discov Artif Intell 2024;4:38. https://doi.org/10.1007/s44163-024-00135-2.
[16] 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;46:366–72. https://doi.org/10.1080/0142159X.2023.2249588.
[17] Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep 2024;14:9330. https://doi.org/10.1038/s41598-024-58760-x.
[18] Yaneva V, Baldwin P, Jurich DP, Swygert K, Clauser BE. Examining ChatGPT Performance on USMLE Sample Items and Implications for Assessment. Acad Med 2024;99:192. https://doi.org/10.1097/ACM.0000000000005549.
[19] Kawahara T, Sumi Y. GPT-4/4V's performance on the Japanese National Medical Licensing Examination. Med Teach 2024:1–8. https://doi.org/10.1080/0142159X.2024.2342545.
[20] Nakao T, Miki S, Nakamura Y, Kikuchi T, Nomura Y, Hanaoka S, et al. Capability of GPT-4V(ision) in the Japanese National Medical Licensing Examination: Evaluation Study. JMIR Med Educ 2024;10:e54393. https://doi.org/10.2196/54393.
[21] 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.
[22] 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:e0000433. https://doi.org/10.1371/journal.pdig.0000433.
[23] 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.
[24] 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;6:536–8. https://doi.org/10.31662/jmaj.2023-0043.
[25] 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:e0000397. https://doi.org/10.1371/journal.pdig.0000397.
[26] Shang L, Xue M, Hou Y, Tang B. Can ChatGPT pass China's national medical licensing examination? Asian J Surg 2023;46:6112–3. https://doi.org/10.1016/j.asjsur.2023.09.089.
[27] 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 2023;10:1237432. https://doi.org/10.3389/fmed.2023.1237432.
[28] 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.ijmedinf.2023.105173.
[29] 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;24:143. https://doi.org/10.1186/s12909-024-05125-7.
[30] Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, 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.

https://doi.org/10.3352/jeehp.2023.20.30.
[31] 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:e45043. https://doi.org/10.7759/cureus.45043.
[32] Huang C-H, Hsiao H-J, Yeh P-C, Wu K-C, Kao C-H. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. Digit Health 2024;10:20552076241233144. https://doi.org/10.1177/20552076241233144.
[33] Lin S-Y, Chan PK, Hsu W-H, Kao C-H. Exploring the proficiency of ChatGPT-4: An evaluation of its performance in the Taiwan advanced medical licensing examination. Digit Health 2024;10:20552076241237678. https://doi.org/10.1177/20552076241237678.
[34] White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT 2023. https://doi.org/10.48550/ARXIV.2302.11382.
[35] Jang D, Yun T-R, Lee C-Y, Kwon Y-K, Kim C-E. GPT-4 can pass the Korean National Licensing Examination for Korean Medicine Doctors. PLOS Digit Health 2023;2:e0000416. https://doi.org/10.1371/journal.pdig.0000416.
[36] 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;52:745–9. https://doi.org/10.1007/s10439-023-03318-7.
[37] 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:e100815. https://doi.org/10.1136/bmjhci-2023-100815.
[38] Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Reshaping medical education: Performance of ChatGPT on a PES medical examination. Cardiol J 2023. https://doi.org/10.5603/cj.97517.
[39] 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:e0000349. https://doi.org/10.1371/journal.pdig.0000349.
[40] Kıyak YS, Emekli E. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad Med J 2024:qgae065. https://doi.org/10.1093/postmj/qgae065.
[41] Amaral E, Norcini J. Quality assurance in health professions education: Role of accreditation and licensure. Med Educ 2023;57:40–8. https://doi.org/10.1111/medu.14880.
[42] Park YS, Yang EB. Three controversies over item disclosure in medical licensure examinations. Med Educ Online 2015;20:10.3402/meo.v20.28821. https://doi.org/10.3402/meo.v20.28821.
[43] Appelhaus S, Werner S, Grosse P, Kämmer JE. Feedback, fairness, and validity: effects of disclosing and reusing multiple-choice questions in medical schools. Med Educ Online 2023;28:2143298. https://doi.org/10.1080/10872981.2022.2143298.
[44] Carrillo-Avalos BA, Leenen I, Trejo-Mejía JA, Sánchez-Mendiola M. Bridging Validity Frameworks in Assessment: Beyond Traditional Approaches in Health Professions Education. Teach Learn Med 2023:1–10. https://doi.org/10.1080/10401334.2023.2293871.
[45] Yenduri G, Ramalingam M, Selvi GC, Supriya Y, Srivastava G, Maddikunta PKR, et al. GPT (Generative Pre-Trained Transformer)— A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. IEEE Access 2024;12:54608–49. https://doi.org/10.1109/ACCESS.2024.3389497.

[46] Kusurkar RA, Orsini C, Somra S, Artino AR, Daelmans HEM, Schoonmade LJ, et al. The Effect of Assessments on Student Motivation for Learning and Its Outcomes in Health Professions Education: A Review and Realist Synthesis. Acad Med J Assoc Am Med Coll 2023;98:1083–92. https://doi.org/10.1097/ACM.0000000000005263.
[47] Gonullu I, Dogan CD, Erden S, Gokmen D. A study on the standard setting, validity, and reliability of a standardized patient performance rating scale – student version. Ann Med 2023;55:490–501. https://doi.org/10.1080/07853890.2023.2168744.
[48] Beets B, Newman TP, Howell EL, Bao L, Yang S. Surveying Public Perceptions of Artificial Intelligence in Health Care in the United States: Systematic Review. J Med Internet Res 2023;25:e40337. https://doi.org/10.2196/40337.
[49] Stai B, Heller N, McSweeney S, Rickman J, Blake P, Vasdev R, et al. Public Perceptions of Artificial Intelligence and Robotics in Medicine. J Endourol 2020;34:1041–8. https://doi.org/10.1089/end.2020.0137.
[50] Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023;229:172.e1-172.e12. https://doi.org/10.1016/j.ajog.2023.04.020.

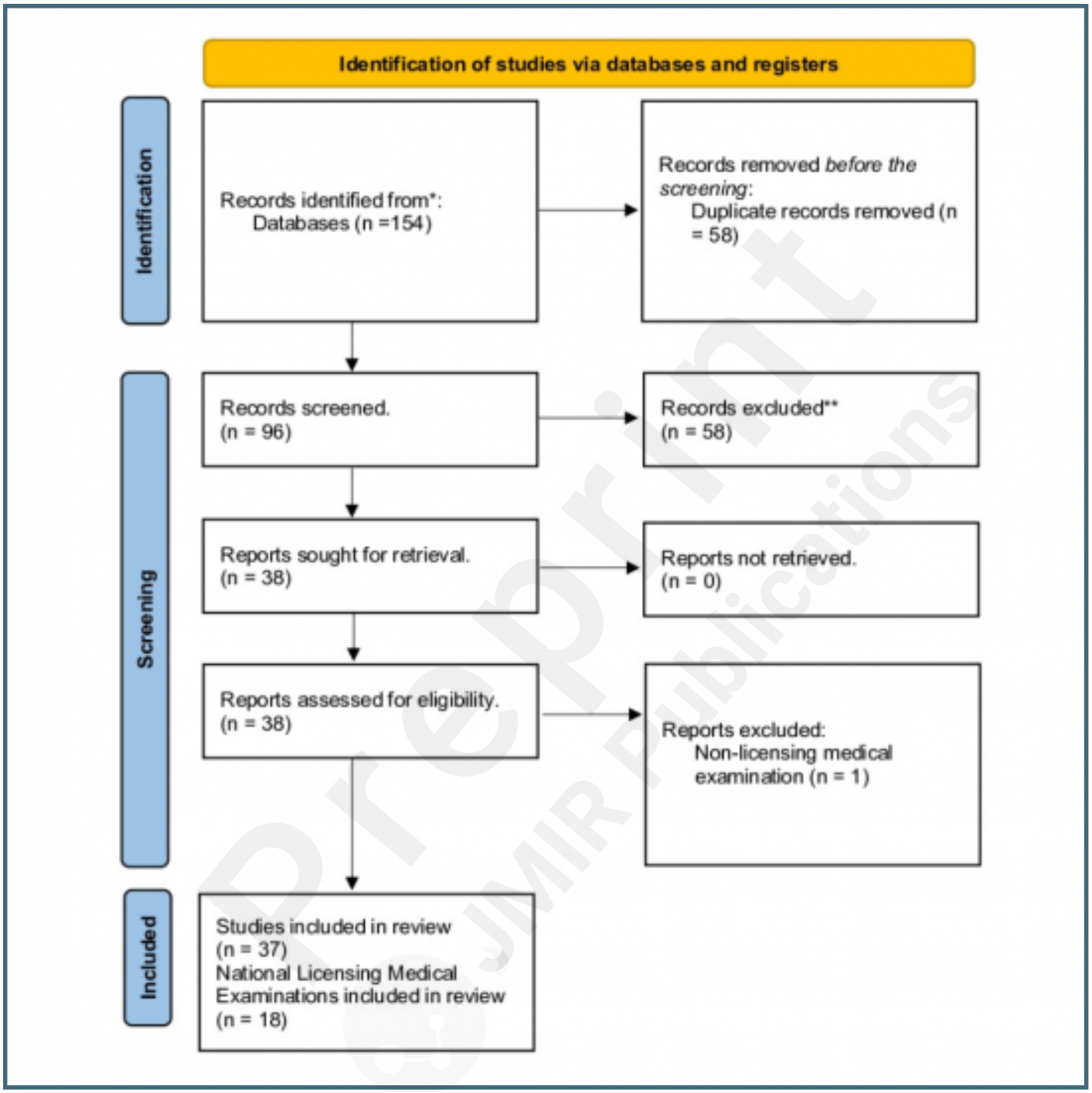
Abbreviations

AI - Artificial Intelligence
GPT - Generative Pre-trained Transformer
NLME - National Licensing Medical Examination
USMLE - United States Medical Licensing Examination
PRISMA - Preferred Reporting Items for Systematic Review and Meta-analysis
LLM - Large Language Model

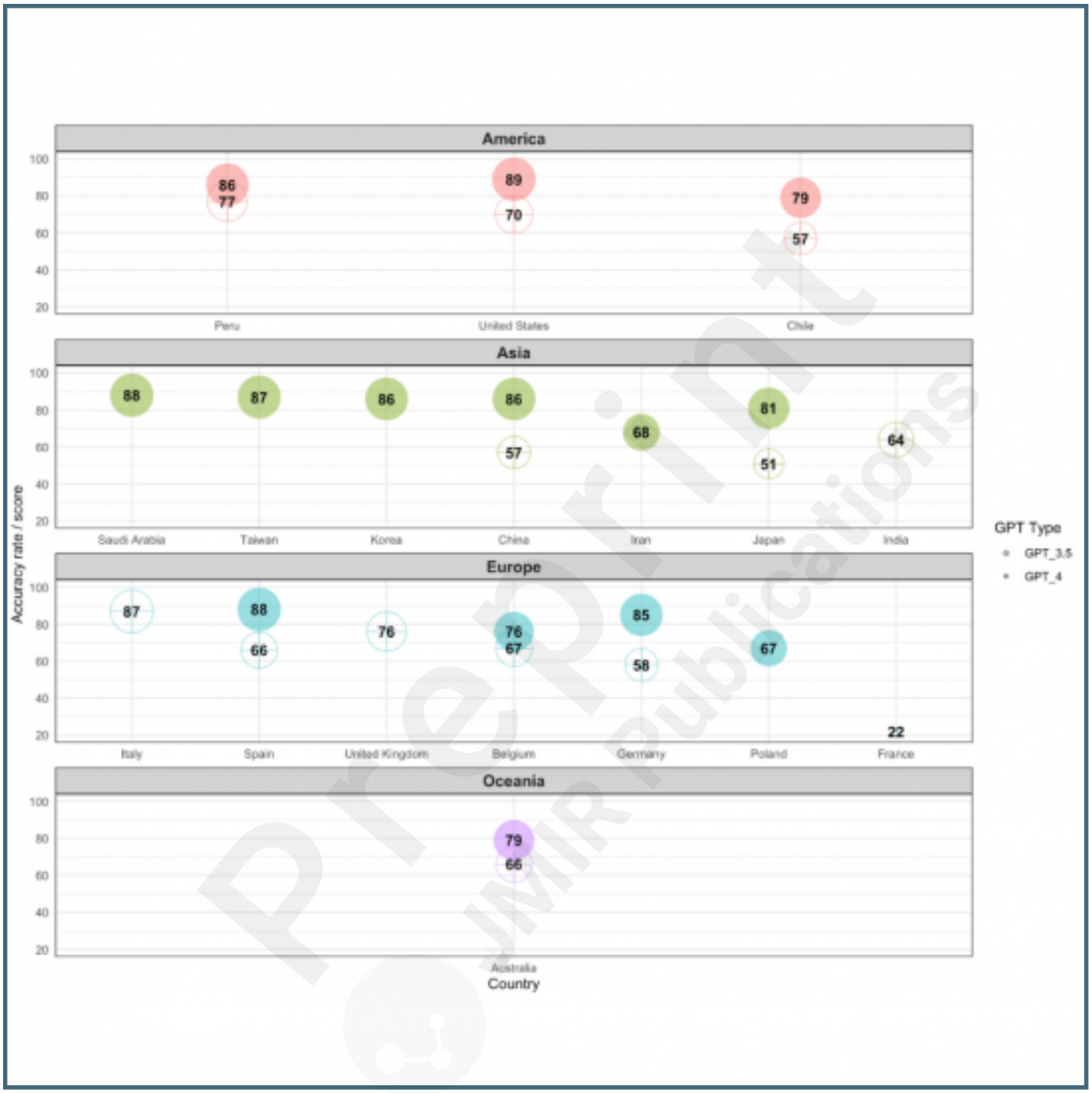
Supplementary Files

Figures

Prisma flowchart.



Performance of ChatGPT in versions GPT-3.5 and GPT-4.



Multimedia Appendixes

List of NLMEs and search strategy.
URL: http://asset.jmir.pub/assets/440b988e29c1f4e7b96a9d0f9210c7a7.docx

Excel extraction sheet.
URL: http://asset.jmir.pub/assets/ac3328cc02224b0a3d8021069ee67efb.xlsx

TOC/Feature image for homepages

AI-generated image, in response to the request, "Create a ChatGPT acquiring a license across the world as a doctor." (Generator: DALL-E June 12, 2024; Requestor: Javier Flores Cohaila).