Yihong Qiu and Chang Liu*

Capable exam-taker and question-generator: the dual role of generative AI in medical education assessment

https://doi.org/10.1515/gme-2024-0021
Received November 15, 2024; accepted December 10, 2024; published online January 14, 2025

Abstract

Objectives: Artificial intelligence (AI) is being increasingly used in medical education. This narrative review presents a comprehensive analysis of generative AI tools' performance in answering and generating medical exam questions, thereby providing a broader perspective on AI's strengths and limitations in the medical education context.
Methods: The Scopus database was searched for studies on generative AI in medical examinations from 2022 to 2024. Duplicates were removed, and relevant full texts were retrieved following inclusion and exclusion criteria. Narrative analysis and descriptive statistics were used to analyze the contents of the included studies.
Results: A total of 70 studies were included for analysis. The results showed that AI tools' performance varied when answering different types of questions and different specialty questions, with best average accuracy in psychiatry, and were influenced by prompts. With well-crafted prompts, AI models can efficiently produce high-quality examination questions.
Conclusion: Generative AI possesses the ability to answer and produce medical questions using carefully designed prompts. Its potential use in medical assessment is vast, ranging from detecting question error, aiding in exam preparation, facilitating formative assessments, to supporting personalized learning. However, it's crucial for educators to always double-check the AI's responses to maintain accuracy and prevent the spread of misinformation.

Keywords: generative artificial intelligence; large language model; medical examination; medical education

***Corresponding author: Chang Liu**, Department of Immunology and Microbiology, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China, E-mail: tiantianlc@sjtu.edu.cn. https://orcid.org/0000-0002-9996-9756
Yihong Qiu, Center for Teaching and Learning Development, Shanghai Jiao Tong University, Shanghai 200240, China

Introduction

The healthcare field is always quickly and deeply influenced by technology. Since the emergence of ChatGPT, many generative artificial intelligence (GAI) models, such as large language models, visual generation and video generation models have come into the public and is increasingly being used in healthcare field, including clinical decision support, management, medical research, and education. In medical education, GAI has been used for student selection and admission, augmenting teaching, generating teaching and learning materials, simulation, supporting personalized learning, and assessment, etc. [1–3].

Before AI tools can be integrated into medical education to assist medical students, they must possess extensive and accurate medical knowledge [4]. Just as exams are used to evaluate students' mastery of knowledge, researchers use various examinations to assess the medical knowledge of GAI models [5–49]. Studies reported that ChatGPT-4 can pass various medical exams [10, 12, 26, 44], even outperformed many medical students [10, 26, 44]. Although several review papers have evaluated AI competencies in taking medical examinations by their overall accuracy, particularly on multiple-choice questions [4, 50–52], some questions need to be answered. Do AI tools like ChatGPT-4 have a stronger foundation in some medical fields compared to others? In medical exams, single-best answer multiple-choice questions (MCQs) are the most common type of question, but there are other types of questions, such as open-ended questions. How does GAI perform on different question types? Regardless of the question type, what are the types of incorrect answers? These are the questions this narrative review aims to address.

Since exams are designed to assess the knowledge mastery of test-takers, the quality of exam questions is crucial. Creating exam questions is a time-consuming task that requires the question setter not only to have a deep understanding of the medical field, but also to have knowledge in evaluation. In formal exams, a team of assessment experts typically designs the questions. Studies have explored the possibility of using GAI for question setting [53–63]. Therefore, how is the quality of questions generated by GAI and

how to measure the quality? What prompts were used? These are also questions that this review will explore.

This narrative review aims to answer five research questions:

Q1: How did AI perform in different types of medical exam questions?

Q2: How did AI perform in different specialties?

Q3: What were the types of incorrect answers yielded by AI?

Q4: What were the qualities of AI-generated exam questions? How to measure?

Q5: What were the prompt strategies when using AI to answer or generate medical exam questions?

By investigating into the performances of AI tools as both exam-takers and exam-generators, we can uncover insights into AI tools' effectiveness and reliability in medical education assessments. This dual perspectives will allow us to better understand the potential that how AI can enhance evaluation processes, improve question quality, and contribute to personalized learning experiences.

Methods

Literature search

The literature search and screening process followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guideline [64]. Scopus database was searched in September 2024 with keywords of "Generative Artificial Intelligence," "GAI," "ChatGPT," "GPT," "Bard," "Bing," "Claude," "Gemini," "DALLE," "Midjourney," and "Stable Diffusion," as well as "medical examination," "medical exam," "medical assessment," "medical test" in title-abstract-keywords, from 2022 to 2024. The records obtained were examined to eliminate any duplicates. Once the duplicates were removed, the titles and abstracts of the retrieved studies were screened to identify those met the inclusion and exclusion criteria (Table 1). Subsequently, the full texts of the identified studies were retrieved, and those inaccessible to the full text were excluded from further analysis. When necessary, papers from reference were manually searched.

Table 1: Inclusion and exclusion criteria of the retrieved paper.

Criteria type	Description
Inclusion criteria	Peer-reviewed original studies and practical reports Testing generative artificial intelligence (AI) in any kinds of medical examinations Generative AI in generating any kinds of questions for medical examinations Literature published from 2022 to 2024 Literature in English
Exclusion criteria	Studies irrelevant to generative AI in medical examination Duplicate studies Letter to editor, editorial, correspondence, reply, conference paper, and book chapter

Data analysis

Data from the included studies were extracted into *Microsoft Excel* spreadsheets. The extracted characteristics of the studies included: title, authors, publication year, medical examination name, examination type, specialty, question type, country or region of the examination, AI model, prompt strategy, accuracy rate, passing score, error type, quality measure, and language interacting with AI, etc. Narrative analysis and descriptive statistics were used to analyze the contents of the included studies. When calculating the average accuracy of AI responses to exam questions in a specific medical specialty, only studies with at least 10 questions were included, provided there were at least five such studies. Studies that did not clearly specify the number of questions were excluded. The difficulty distributions of exam questions were assumed similar among studies. Average accuracy was calculated by dividing the total number of correctly answered questions in all exams by the total number of questions. The 95 % confidence interval of the accuracy was estimated using the binomial distribution.

Ethical review

This review was conducted based on published studies; therefore, no ethical review was required.

Results

Searched records

The searching strategy resulted in 119 studies. Then, 2 duplicate records, 49 irrelevant studies, and 2 inaccessible ones were removed; and 3 manually searched records were added. Ultimately, a total of 70 studies were included in this review (Figure 1).

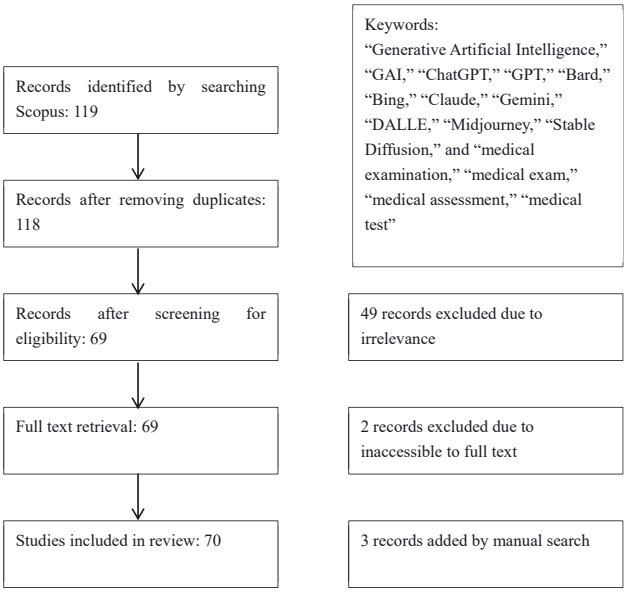

Figure 1: Literature screening diagram.

AI tools' performance in different types of medical exam questions

Single-best answer MCQs, choose-n-from-many, true or false, and open-ended questions are possible question types in medical examinations. Single-best answer MCQs are very popular in various medical exams, so AI's ability to answer this MCQs has attracted the interest of many researchers. Meta-analysis of the published studies shown that ChatGPT-3.5 had an overall accuracy of 61.1 % in Levin et al.'s study [51], and an overall accuracy of 58 % in Liu et al.'s study [4], which were quite similar, while ChatGPT-4 had a higher accuracy of 81 % [4] (Table 2).

Choose-n-from-many is a variant of single-best answer MCQs, which has two or more correct answers in answer options. In Haze et al.'s study, AI's ability to respond to this kind of question was inferior to answering single-best answer MCQs. For example, ChatGPT-4 had an accuracy of 69.8 % in answering choose-n-from-many questions, compared to an accuracy of 83.7 % in answering single-best answer MCQs [39]. However, in Hirano et al.'s study, ChatGPT-4 Turbo/ChatGPT-4 Turbo with vision had similar accuracy in answering single-best answer MCQs and choose-n-from-many questions [34] (Table 3).

True/false question is also a variant of MCQs, which has only two options. Sadeq et al. reported that AI's performance in true/false questions was lower than that in MCQs [6]. For example, ChatGPT-3.5 obtained an accuracy of 23.1 % in answering true/false questions, while it achieved an accuracy of 62.9 % in answering MCQs, and similar trends were observed in GTP-4, Bard, Bing, Claude, Claude Instant, and Perplexity [6]. However, in another study, Sood et al. found that GPT-4 had an accuracy of 83 % in answering true/false questions, better than answering MCQs, and so did GPT-3.5 (Table 3) [33].

For open-ended questions, ChatGPT-3.5 obtained 66.5 % accuracy in community medicine [65], 73.6 % in family medicine [66], and 77.4 % in psychiatry [9]. ChatGPT-4 achieved 75 % accuracy in otolaryngology-head and neck surgery [5], 81.0 % in family medicine [66] (Table 3). Both ChatGPT-3.5 and ChatGPT-4 seemed to have a better performance compared to answering MCQs, and both exceeded the common passing threshold of 60 %.

Table 2: Meta-analysis of AI's accuracy in multiple-choice questions.

Studies	Number of papers	AI tool	Accuracy with 95 % *CI*
Levin et al. 2024 [51]	19	ChatGPT-3.5	61.1 % (56.1 %–66.0 %)
Liu et al. 2024 [4]	25	ChatGPT-3.5	58 % (53 %–63 %)
	29	ChatGPT-4	81 % (78 %–84 %)

AI, artificial intelligence; *CI*, confidence interval.

Performance of AI in addressing MCQs across various specialties

Popular AI tools were employed in answering MCQs, and their performance varies across different specialties. Table 4 presents the results categorized by specialties. It reveals that ChatGPT-3.5 and ChatGPT-4 were most used AI tools in medical examinations. ChatGPT-3.5 performed best in psychiatry, with an average accuracy of 74.6 %. Its second-best performance was in general surgery, reached 70.6 % accuracy; then in neurology (61.8 %), internal medicine (61.6 %), and emergency medicine (54.9 %). Its worst performance was in pediatrics as well as gynecology and obstetrics, with an average accuracy of 53.6 %. While ChatGPT-4 performed better than ChatGPT-3.5, it also performed best in psychiatry, with an average accuracy of 90.1 %; followed by internal medicine (84.0 %), general surgery (81.2 %), neurology (78.9 %), pediatrics (78.7 %), emergency medicine (78.3 %), gynecology and obstetrics (76.8 %). ChatGPT-4 performed worst in osteology, with an average accuracy of 67.4 %. Detailed performance of AI tools across various specialties were in Supplementary Material 1.

Prompt strategies in answering questions

In many studies, the original examination questions were directly input to the AI tool [7, 8, 16, 17, 21], which simulated the humans taking the exams. However, a number of

Table 3: Performance of artificial intelligence (AI) in other question types except single-best answer multiple-choice questions (MCQs).

Studies	Question type	Number of questions	AI tool	Accuracy	Accuracy of MCQs as reference
Haze et al. 2023 [39]	Choose-n-from-many	129	ChatGPT-3.5	41.9 %	59.1 %
			ChatGPT-4	69.8 %	83.7 %
Hirano et al. 2024 [34]	Choose-n-from- many	16	ChatGPT-4 Turbo	44 %	41 %
			ChatGPT-4 Turbo with vision	44 %	41 %
Sadeq et al. 2024 [6]	True/false	13	ChatGPT-3.5	23.1 %	62.9 %
			ChatGPT-4	30.8 %	80.7 %
			Bard	15.4 %	61.0 %
			Bing	30.8 %	68.7 %
			Claude	7.7 %	67.4 %
			Claude instant	23.1 %	64.5 %
			Perplexity	0 %	58.7 %
Sood et al. 2023 [33]	True/false	182	ChatGPT-3.5	61 %	31.7 %
			ChatGPT-4	83 %	70.7 %
D'Souza et al. 2023 [9]	Open-ended question	100	ChatGPT-3.5	77.4 % (773.5 out of 1,000 points; 61 % 8.0–10.0 points; 31 % 5.0–7.9 points; 8 % 3.0–4.9 points; 0 % 0.0–2.9 points)	n.a
Gandhi et al. 2024 [65]	Open-ended question	40	ChatGPT-3.5	66.5 % (133 out of 200 points)	n. a.
Huang et al. 2023 [16]	Case	15	ChatGPT-4	87.5 % (correctness, 3.5 out of 4)[a]	78.8 %
Long et al. 2024 [5]	Open-ended question	21	ChatGPT-4	75 % (25.5 out of 34 points)	n. a.
Mousavi et al. 2024 [66]	Open-ended question	77	ChatGPT-3.5	73.6 %	n. a.
			ChatGPT-4	81.0 %	n. a.

[a]Other index including comprehensiveness (3.1 out of 4), novelty (80 %), and hallucination (13.3 %).

Table 4: Performance of AI across specialties.

Specialty	AI tool	Average accuracy with 95 % *CI*	References
Emergency medicine	ChatGPT-3.5	54.9 % (50.6–59.3 %)	[6, 7, 24, 39, 67, 68]
	ChatGPT-4	78.3 % (74.6–82.0 %)	[6, 32, 39, 67, 68]
General surgery	ChatGPT-3.5	70.6 % (65.9–75.3 %)	[6, 19, 22, 24, 26, 31, 50, 67, 69]
	ChatGPT-4	81.2 % (78.0–84.3 %)	[6, 19–21, 26, 31, 32, 50, 67, 69]
Gynecology and obstetrics	ChatGPT-3.5	53.6 % (49.6–57.7 %)	[6, 22, 24, 26, 31, 38, 39, 46, 67, 69]
	ChatGPT-4	76.8 % (72.6–80.9 %)	[6, 26, 31, 32, 39, 67, 69]
Internal medicine	ChatGPT-3.5	61.6 % (57.9–65.3 %)	[6, 24, 26, 31, 67, 69]
	ChatGPT-4	84.0 % (81.3–86.7 %)	[6, 21, 26, 31, 32, 67, 69]
Neurology	ChatGPT-3.5	61.8 % (59.3–64.4 %)	[17, 26, 29, 38, 39, 42, 44]
	ChatGPT-4	78.9 % (76.5–81.2 %)	[26, 28, 29, 32, 39, 42, 44]
Osteology	ChatGPT-4	67.4 % (63.4–71.4 %)	[11, 23, 28, 32, 39, 49]
Pediatrics	CharGPT-3.5	53.6 % (49.0–58.1 %)	[6, 24, 26, 31, 39, 67, 69, 70]
	ChatGPT-4	78.7 % (75.4–81.9 %)	[6, 21, 26, 28, 31, 32, 39, 67, 69, 70]
Psychiatry	ChatGPT-3.5	74.6 % (69.1–80.0 %)	[24, 26, 31, 38, 39]
	ChatGPT-4	90.1 % (87.5–92.7 %)	[12, 26, 31, 32, 39]

AI, artificial intelligence; *CI*, confidence interval.

studies employed lead-in prompts from simple to complex. The components in these lead-in prompts could be classified as basic components and advanced ones (Figure 2). The most commonly used basic component was requiring AI to select one correct answer for MCQs [15, 34, 41, 44, 71]. The other basic components were to specify specialty

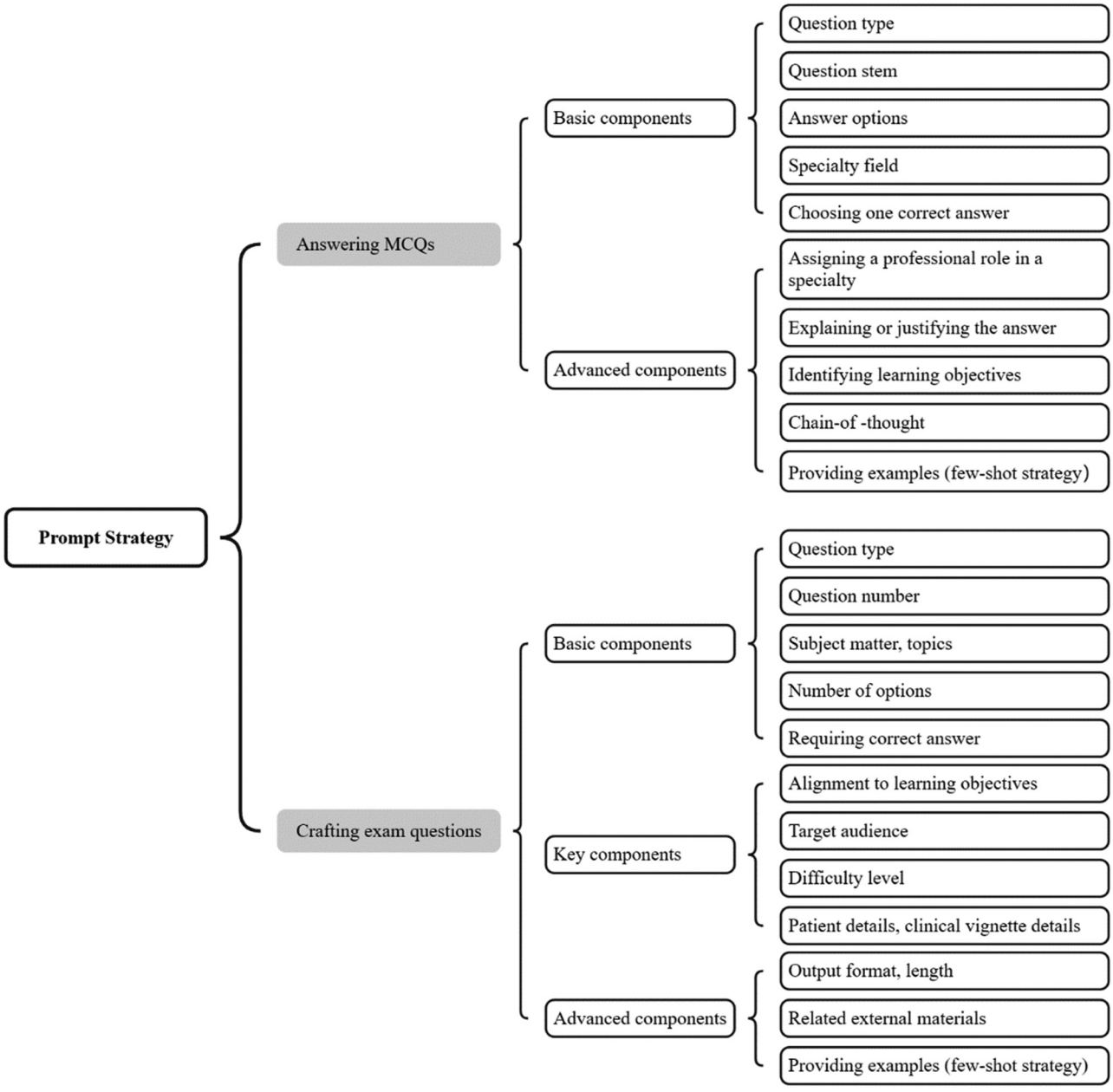

Figure 2: Components of prompt in answering and crafting medical exam questions. MCQs, multiple-choice questions.

field [5, 10, 12, 13, 40, 41, 66, 67] and question type [12, 28, 36, 41, 47, 66]. The advanced components included assigning a professional role in an expertise field [12, 32, 33, 35, 41], explaining and/or justifying its answer [12, 33, 67, 69] or not explaining or justifying its answer [13, 71, 72], identifying learning objective [67, 69], chain-of-thought strategy [32, 34, 41], and few-shot strategy [40]. Roos et al. [10], Wu et al. [40] and Torres-Zegarra et al. [69] employed structured prompts that compiled basic and advanced components in a clearer way (Figure 2). More detailed analysis of prompts in response to medical exam questions were in Supplementary Material 2.

Analysis of incorrect AI answers

Several studies analyzed in detail the types of incorrect answers yielded by AI (Table 5). Guillen-Grima et al. [28] analyzed wrong answers based on Taxonomy of Medication Errors by National Coordinating Council for Medication Error Reporting and Prevention [73], which has 9 categories.

Table 5: Analysis of incorrect answers.

Studies	AI tool	Criteria	Type and number of incorrect answers
Guillen-Grima et al. 2023 [28]	GPT-4	NCC MERP classification	Total 24 Category A-capacity to cause error (n=10) Category B-error did not reach the patient (n=1) Category C-error reached patient but did not cause harm (n=3) Category D-error reached the patient and required monitoring (n=4) Category E-error caused temporary harm and required intervention (n=2) Category F-error lead to initial or prolonged hospitalization (n=2) Category G-error resulted in permanent patient harm (n=2) Category H-error necessitated intervention to sustain life (n=0) Category I-error contributed to or resulted in the death (n=0)
Herrmann-Werner et al. 2024 [12]	GPT-4	Bloom's taxonomy	Total 68 Remember (n=29) Understand (n=23) Apply (n=15) Analyze (n=0) Evaluate (n=1) Create (n=0)
Maitland et al. 2024 [18]	GPT-4	Clinical thinking and reasoning	Total 51 Assumption error (n=1) Base-rate neglect (n=5) Confabulation error (n=1) Confirmation biases (n=1) Context error (n=8) Factual error (n=27) Misinterpretation of question (n=5) Omission error (n=12)
Wang et al. 2023 [41]	GPT-3.5	Hallucination analysis	Total 106 Open-domain error (n=66) Closed-domain error (n=40)
	GPT-4		Total 48 Open-domain error (n=30) Closed-domain error (n=18)

AI, artificial intelligence; NCC MERP, National Coordinating Council for Medication Error Reporting and Prevention.

They identified that in a total of 24 incorrect answers, 10 could cause medication errors (category A), and 8 could not cause harm to patient (category B–D), while 6 could cause harm to patient (category E–H), and none would cause death (category I).

Herrmann-Werner et al. [12] categorized incorrect answers according to the revised Bloom's taxonomy, which has six levels regarding to cognition challenge: "remember, understand, apply, analyze, evaluate and create". In a total of 68 wrong responses, the most incorrect answers were at the "remember" and "understand" level, with 29 and 23 incorrect answers respectively. Maitland et al. [18] classified wrong answers into 8 types in accordance with clinical thinking and reasoning. In 51 incorrect answers, the factual error was the most common one, followed by omission error [18]. Wang et al. [41] divided incorrect answers into open-domain and closed-domain hallucination. They found that GPT-4 had less open-domain and closed-domain errors.

Prompt strategies in generating questions

Prompts used to generate medical exam questions were various (Figure 2). The basic components were the question types [53–63], subject matter or topics [53–55, 57–59, 62, 63], number of questions [53–63], number of answer options [59, 62, 63], and requirement to provide correct answers [53, 55, 60–63]. The key components included aligning to learning objective [56, 57, 63], targeted at specific audiences [53, 55, 56, 60, 63], specifying question difficulty level, such as easy

or difficult [58], knowledge-based [55, 60] or clinical case-based [55–57]; for generating clinical cases or case-based questions, the patient details [59] or clinical vignette details were required [57, 58]. The advanced components included providing examples (few-shot strategy) [55], referring to uploaded file as the question source [54, 60, 61], and specifying the output format of questions and answers [55, 57, 59, 61]. Kıyak et al. [58] employed a well-structured prompt framework to generate MCQs, which consisted of the above-mentioned basic, key and advanced components. Detailed analysis of prompts in generating medical exam questions were in Supplementary Material 2.

Quality assessment of AI-generated medical questions

While AI's performance in answering medical questions could be evaluated by comparing its answers to reference answers, there are no standard criteria for assessing AI's performance in generating medical questions. Thus, researchers proposed their own quality measures to evaluate the quality of AI-generated questions (Table 6).

The commonly used quality measures were clarity [54, 58, 60, 62] or ambiguity [61], and correctness [55, 58], accuracy [54] or appropriateness [60]. Measures such as appropriateness [55, 58], suitability [60], validity [56] or instructional alignment [61] were used to judge the degree a question aligned to a topic, content or intended learning objective. For the difficulty level of the questions, some researchers used Likert scale to measure the question difficulty [54, 56, 61, 62], while some other researchers put AI-generated questions in real exams to measure the difficulty [57, 63]. Besides, AI-generated questions in real exams were also assessed by discrimination index [57, 58, 63]. When comparing AI-generated questions to those created by humans, the quality of the AI-generated questions were almost as good as human-generated ones, either by human judgement [60, 63] or by test results [57].

Discussion

Principal findings

This narrative review highlights that generative AI tools, particularly large language models, demonstrated capabilities in answering and creating medical examination questions. AIs' performance varied when answering different types of questions, and probably performed best when answering open-ended questions. AIs' performance also varied when answering different specialty questions, with the best achievement in psychiatry for both ChatGPT-3.5 and ChatGPT-4 [9, 12, 26], and the worst achievement in osteology for ChatGPT-4 [11, 23, 49] and in pediatrics as well as gynecology and obstetrics for ChatGPT-3.5 [39, 46, 49, 70]. When guided by appropriate prompts, AI tools could generate suitable medical exam questions [53], which were comparable to questions created by humans [57, 59, 60, 63].

AI tools' performance is influenced by question types, specialty knowledge and prompts [20, 39, 67]. MCQs, choose-n-form-many and true/false questions are objective questions, and open-ended questions are subjective questions. Objective questions have question stem and answer options where the clue to the answer is hidden. Essentially, answering objective questions is a kind of finding the best match. There are only two possibilities for an answer: either it is correct or it is wrong; there is no middle ground. While open-ended questions, especially clinical vignette-based questions, require exam taker to apply and synthesize their knowledge, and is not an easy task for humans. However, it seemed not a hard task for AI. AI tools achieved high scores in answering open-ended questions. This might be due to the grading mechanism. Even if the final decision is wrong, each key point can get a score. Since AI was trained on large scale data set, it is quite knowledgeable and easy to generate clinical vignette-related content that may contain key points, thus it performs quite well in answering open-ended questions, not necessary to have real clinical thinking and reasoning skills.

The performance differences among specialties likely stem from a combination of factors such as the types of data available for training the AI, the complexity of clinical reasoning, and the diagnostic process specific to each field. Haze et al. [39] investigated the relationship between the ChatGPT's accuracy in different specialties and the number of related documents in the Web of Science Core Collection. They found significant positive correlation between the accuracy of ChatGPT-4 and the number of all-type documents. In specialties like psychiatric, where standardized questionnaires and diagnostic criteria are well-documented in textual form, AI can easily process the data, leading to better performance. In contrast, in fields like orthopedics, where diagnostic decisions often rely on interpreting medical imaging, current language models like ChatGPT have limitations, resulting in weaker performance. Pediatrics and obstetrics/gynecology involve more case-by-case variability, where factors like age, medical history, and developmental stage matter significantly. AI model might struggle with the complexities of clinical decision-making, thus leading to lower performance. Additionally, AI accuracy

Table 6: Quality assessment of AI-generated items.

Studies	Subject	AI tool	Quantity of items	Quality metrics	Metric value
Agarwal et al. 2023 [56]	Physiology	ChatGPT	110	Validity	3 (3–3)[a]
				Difficulty	1 (0–1)
				Reasoning effort	1 (1–2)
		Bard	110	Validity	3 (1.5–3)
				Difficulty	1 (1–2)
				Reasoning effort	1 (1–2)
		Bing	100	Validity	3 (1.5–3)
				Difficulty	1 (1–2)
				Reasoning effort	1 (1–2)
Ayub et al. 2023 [54]	Dermatology	ChatPDF	40	Accuracy	87.5 %
				Complexity	75 %
				Clarity	77.5 %
					40 % questions were accurate and appropriate
Cheung et al. 2023 [60]	Internal medicine and surgery	ChatGPT plus	50	Appropriateness of the question	7.72[b]
				Clarity and specificity	7.56
				Relevance	7.56
				Discriminative power of alternatives	7.26
				Suitability	7.25
				Compared with human-generated questions	No significant difference except for humans got a slightly higher score in relevance
Coşkun et al. 2024 [59]	Evidence-based medicine	ChatGPT-3.5	15	Discrimination index	6 items greater than 0.3; 5 items greater than 0.25
Grévisse et al. 2024 [61]	Endocrinology	API (gpt 4-1106-preview)	80	Pertinence	79 %
				Difficulty	36 %
				Level of specificity	68 %
				Ambiguity	21 %
				Instructional alignment	84 %
	Neurology		20	Pertinence	5 %
				Difficulty	20 %
				Level of specificity	5 %
				Ambiguity	0 %
				Instructional alignment	5 %
Klang et al. 2023 [55]	Internal medicine, general surgery, obstetrics and gynecology, psychiatry and pediatric	GPT-4	210	Correctness	n.a.
				Appropriateness	n.a.
					0.5 % false; 15 % needed revisions
Kıyak et al. 2024 [58]	Rational pharmacotherapy	ChatGPT-3.5	10	Correctness	100 %
				Clarity	100 %
				Appropriateness	20 %
				Discrimination index	Greater than 0.3

Table 6: (continued)

Studies	Subject	AI tool	Quantity of items	Quality metrics	Metric value
Laupichler et al. 2024 [63]	Neurophysiology	ChatGPT-3.5	25	Difficulty Discrimination index Compared to human-generated questions	0.69 0.24 57 % of question sources were identified correctly
Rivera-Rosas et al. 2024 [62]	Anatomy and kinesiology	ChatGPT-3.5	55	Concise and comprehensible of questions Clarity Simpleness of language Difficult of questions	89 % 91 % 91 % 24 %
Zuckerman et al. 2023 [57]	Reproductive system	ChatGPT	29	Difficulty Discrimination index Compared to human-generated questions	0.71 0.23 No significant difference

[a]Median with interquartile range. [b]Likert scale of 1–10.

tends to decline when questions involve country- or region-specific knowledge, likely due to limited training on such localized data [67].

Prompt may also influence AI's accuracy in answering medical questions. In Herrmann-Werner et al.'s study, detailed prompt resulted in a higher accuracy than the short prompt did but without significance [12], because the key components in detailed prompt and short component functioned the same, except that detailed one specified the answer format. When chain-of-though prompt was employed, ChatGPT could correctly answer more than half of the originally wrongly answered questions [32]. When few-shot technology was used to enhance AI models' in-context learning, their performance were improved; and AI models' performance were even better when few-shot technology and external knowledge were combined [40]. However, when the context information "CFPC exam" were removed from the prompt, it resulted in an improved accuracy [66], probably because AI did not understand the acronym CFPC correctly. Besides, by highlighting errors in AI's answers through prompt engineering, the AI might arrive at the correct response. However, the studies included in this review did not address the situation of identifying AI mistakes and then re-evaluating its subsequent answers.

When AI gave an incorrect answer, a close look at it could reveal valuable insights into the limitations of AI. From the viewpoint of outcomes induced by wrong answers [28], it could remind medical users to always keep in mind the importance of human oversight and critical evaluation. From the viewpoint of thinking process to identity where and why the AI's reasoning went wrong [12, 18, 41], it could enhance our understanding of the difference between human judgment and AI reckoning.

Using AI to generate medical exam questions could save medical educators' time [60]. To ensure the quality of AI-generated questions, it is crucial to carefully craft the prompts as well as critically review the generated questions. Clarity of the questions is not a problem [54, 58, 60, 62], but appropriateness can be an issue. Medical educators often instructed AI to generate questions in specific field or topic [53–55, 57–59, 62, 63], rather than aligning them with intended learning objectives [56, 57, 63]. This approach can lead to questions that are correct but not suitable for assessment [58, 61], whereas focusing on learning objectives can ensure the validity of the examination [56, 57, 63]. Thus, instruction to align learning objective in the prompt is key to generate suitable exam questions.

Critically review AI-generated questions with predefined criteria before putting the questions in an exam is a good practice [54–63]. Although these indexes that measure question quality seemed different, the key measures should focus on fact correctness and alignment with learning objectives. For MCQs, possibility of the options is also a key measure. It could ensure that the questions are not only correct and relevant but also effectively measure the intended competencies and knowledge areas. Check the AI-generated questions in an exam with difficulty level and

discrimination index [57, 58, 63] could help identify questions that are either too easy or too hard, as well as those that do not effectively differentiate between high and low performers. This analysis can lead to decision on whether and how to use these questions in future assessment.

The relationship between AI's ability to answer questions and generate them is an interesting yet underexplored area, but direct evidence on this topic is scarce right now. Previous studies have shown that students who engaged in question generation activities tended to have better academic performance [52, 74–76], suggesting that generating questions can enhance learning. This implies that strong performance in answering questions may be linked to the ability to generate high-quality questions. However, since AI models like ChatGPT have been trained on vast datasets and do not "learn" from the process of generating questions, their ability to generate and answer questions is likely correlated to the quality of the specific knowledge embedded in those datasets.

Implications of AI in medical education

Medical educators can employ AI to verify whether the questions created by humans for examination contain any ambiguities or errors [10, 53]. As mistakes are sometimes discovered after formal exams [10, 15, 24], it's beneficial to have AI check the quality of the questions, while ensuring that the exam questions are not leaked. By using AI to answer these questions and asking it to explain its reasoning, medical educators can quickly spot potential issues in the exam questions.

AI can serve as a tool for medical students in preparing for exams. While some argue that AI tools are not yet perfect in accuracy and thus cannot be considered as learning tools [4, 54], we, along with some researchers [16, 22, 77], hold a different perspective. For medical students or residents taking licensing exams or specialty exams, the passing score is typically around 60 %–70 % [13, 17, 24, 26, 29, 35, 38], and they are not required to achieve a very high accuracy rate. They can use AI as a peer to assist in exam preparation. As they are not beginners, they should have developed medical thinking and reasoning skills, enabling them to judge the quality of AI response, especially when AI provides explanations for its answers. Although AI's accuracy is not top-notch, this can also be an advantage, as it forces users to maintain critical thinking rather than relying on AI blindly. If a medical student detects an AI error, pointing out its mistakes might sometimes lead to the correct answer. This kind of human-AI collaboration is happening in the real workplace. The studies included in this review did not mention the scenario of pointing out AI errors and then looking at its answers again. However, for beginners who are just learning the new knowledge, AI is not an ideal authoritative source for learning [6], as they lack comprehensive judgment capabilities.

With well-crafted prompts, AI can efficiently produce high-quality examination questions [78]. Clear requirements, providing context, alignment with learning objectives, describing clinical scenarios and the provision of examples [79–81] are all good practices for ensuring the quality of the questions. Additionally, specifying output formats can significantly reduce the workload of editing. Medical educators should learn about prompt engineering or follow guidelines for crafting prompts to create excellent ones [78–81]. Of course, due to the risk of hallucination, human review of AI-generated questions is always essential [61].

The knowledge-based questions generated by AI can serve as an effective tool for formative assessment in the classroom [57, 59, 63]. Medical teachers can use the questions to gain real-time insights into students' mastery of previous knowledge and progress in learning new concepts, thereby adjust teaching content and pace if necessary. AI-generated medical cases can be used as material for classroom discussions [59], fostering students' clinical judgment and decision-making skills. Additionally, educators can also teach students on how to utilize AI for creating questions, thus, students can use AI for self-assessment to check their understanding of the knowledge, thus support personalized learning [82].

The strong capabilities of AI in answering and creating medical exam questions undoubtedly challenge the traditional modes of examination [17, 22, 26, 38, 83]. In the near future, medical exams may more closely mirror real-life medical practice. For instance, it could involve simulating scenarios where patients describe their physical discomfort to doctors, with these descriptions potentially being ambiguous or conflicting. Doctors need to make preliminary judgments and gather key information for decision-making through questioning, laboratory tests, and other methods, continuously adjusting and refining their decisions based on new information. Accordingly, exam questions could be presented in a step-by-step adaptive manner to simulate the actual diagnostic and treatment process. Current clinical case-based questions, though seemingly complex, essentially provide necessary and consistent information in advance, subtly offering exam-takers clues to find the answers.

As AI technology continues to break new ground, the capabilities of AI are becoming even powerful. It is transforming the way we teach and learn. Educators should

always maintain a vigilant and cautious approach when utilizing AI in teaching, to ensure that AI tools are used responsibly and ethically to enhance student learning experiences without compromising the integrity of the educational process [17, 25, 58, 78].

Limitations

The included studies were only from Scopus database, which could introduce selection bias and potentially exclude studies with alternative findings or perspectives on the topic. Some studies that addressed questions spanning multiple specialties did not report the AI's performance within each specialty, which could introduce bias to the findings. Additionally, the uneven categorization of specialties, and limited number of non-MCQ questions might also influenced the results. Furthermore, with the rapid advancement of AI technology, sophisticated models like ChatGPT-4o are now available for free use. Consequently, findings based on earlier models may vary from those obtained using the latest models.

Conclusions

This narrative review analyzed 70 studies using AI in the field of medical examination. AI tools performed quite well in answering open-ended questions. Their performance varied across different specialty questions, with the highest accuracy in psychiatry for both ChatGPT-3.5 and ChatGPT-4, while ChatGPT-4 performed the worst in osteology and ChatGPT-3.5 in pediatrics and gynecology/obstetrics. With well-crafted prompts, AI models can efficiently produce high-quality examination questions. By investigating into the performances of AI tools as both exam-takers and exam-generators, we suggest their usage in question error checking, exam preparation, question generation, formative assessment, and personalized learning. In the same time, critical judgment should always be applied when checking AI-yielded answers and AI-generated questions as these models can produce plausible but inaccurate information. Educators must always verify AI outputs to ensure accuracy and avoid the risk of misinformation in medical education.

Acknowledgments: The authors thanked the reviewers for their valuable comments and suggestions.
Research ethics: This review was conducted based on published studies; therefore, no ethical review was required.
Informed consent: Not applicable.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The authors state no conflict of interest.
Research funding: This work was supported by Medical Education Research Project of Medical Education Branch of Chinese Medical Association (2023B344).
Data availability: The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

References

1. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach 2024;46:446–70.
2. Benítez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, Phuoc LV, et al. Harnessing the potential of large language models in medical education: promise and pitfalls. J Am Med Inf Assoc 2024;31:776–83.
3. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ 2024;58:1276–85.
4. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. J Med Internet Res 2024;26:e60807. https://doi.org/10.2196/60807.
5. Long C, Lowe K, Zhang J, dos Santos A, Alanazi A, O'Brien D, et al. A novel evaluation model for assessing ChatGPT on otolaryngology-head and neck surgery certification examinations: performance study. JMIR Med Educ 2024;10:e49970. https://doi.org/10.2196/49970.
6. Sadeq MA, Ghorab RMF, Ashry MH, Abozaid AM, Banihani HA, Salem M, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. Sci Rep 2024;14:18859. https://doi.org/10.1038/s41598-024-68996-2.
7. Akhter M. Accuracy of GPT's artificial intelligence on emergency medicine board recertification exam. Am J Emerg Med 2024;76:254–5.
8. Kelloniemi M, Koljonen V. AI did not pass finnish plastic surgery written board examination. J Plast Reconstr Aesthetic Surg 2023;87:172–9.
9. D'Souza FR, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatry 2023;89:103770.
10. 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. https://doi.org/10.2196/46482.
11. Saad A, Iyengar KP, Kurisunkal V, Botchu R. Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. Surgeon 2023;21:263–6.

12. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res 2024;26:e52113. https://doi.org/10.2196/52113.
13. Kufel J, Bielówka M, Rojek M, Mitręga A, Czogalik Ł, Kaczyńska D, et al. Assessing ChatGPT's performance in national nuclear medicine specialty examination: an evaluative analysis. Iran J Nucl Med 2024;32:60—5.
14. Surapaneni KM. Assessing the performance of ChatGPT in medical biochemistry using clinical case vignettes: observational study. JMIR Med Educ 2023;9:e47191. https://doi.org/10.2196/47191.
15. Siebielec J, Ordak M, Oskroba A, Dworakowska A, Bujalska-Zadrozny M. Assessment study of ChatGPT-3.5's performance on the final polish medical examination: accuracy in answering 980 questions. Healthcare Switz 2024;12:1637.
16. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology. Front Oncol 2023;13:1265024. https://doi.org/10.3389/fonc.2023.1265024.
17. Stengel FC, Stienen MN, Ivanov M, Gandía-González ML, Raffa G, Ganau M, et al. Can AI pass the written European board examination in neurological surgery? — Ethical and practical issues. Brain Spine 2024;4:102765. https://doi.org/10.1016/j.bas.2024.102765.
18. Maitland A, Fowkes R, Maitland S. Can ChatGPT pass the MRCP (UK) written examinations? Analysis of performance and errors using a clinical decision-reasoning framework. BMJ Open 2024;14:e080558. https://doi.org/10.1136/bmjopen-2023-080558.
19. Gencer A, Aydin S. Can ChatGPT pass the thoracic surgery exam? Am J Med Sci 2023;366:291—5.
20. Ghanem D, Nassar JE, El Bachour J, Hanna T. ChatGPT earns American board certification in hand surgery. Hand Surg Rehabil 2024;43:101688. https://doi.org/10.1016/j.hansur.2024.101688.
21. 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.
22. Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare Switz 2023;11:2046.
23. Fiedler B, Azua EN, Phillips T, Ahmed AS. ChatGPT performance on the American Shoulder and Elbow Surgeons maintenance of certification exam. J Shoulder Elbow Surg 2024;33:1888—93.
24. Suwała S, Szulc P, Guzowski C, Kamińska B, Dorobiała J, Wojciechowska K, et al. ChatGPT-3.5 passes Poland's medical final examination—is it possible for ChatGPT to become a doctor in Poland? SAGE Open Med 2024;12:1—7.
25. Funk PF, Hoch CC, Knoedler S, Knoedler L, Cotofana S, Sofo G, et al. ChatGPT's response consistency: a study on repeated queries of medical examination questions. Eur J Investig Health Psychol Educ 2024;14:657—68.
26. 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. https://doi.org/10.2196/50965.
27. Farhat F, Chaudhry BM, Nadeem M, Sohail SS, Madsen DØ. Evaluating large language models for the national premedical exam in India: comparative analysis of GPT-3.5, GPT-4, and bard. JMIR Med Educ 2024;10:e51523. https://doi.org/10.2196/51523.
28. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, 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;13:1460—87.
29. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK neurology specialty certificate examination. BMJ Neurol Open 2023;5:e000451. https://doi.org/10.1136/bmjno-2023-000451.
30. Tsoutsanis P, Tsoutsanis A. Evaluation of large language model performance on the multi-specialty recruitment assessment (MSRA) exam. Comput Biol Med 2024;168:107794. https://doi.org/10.1016/j.compbiomed.2023.107794.
31. Rojas M, Rojas M, Burgess V, Toro-Pérez J, Salehi S. Exploring the proficiency of ChatGPT 3.5, 4, and 4 with vision in the Chilean medical licensing exam: an observational study. JMIR Med Educ 2024;10:e55048. https://doi.org/10.2196/55048.
32. 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. Digital Health 2024;10:1—11.
33. Sood A, Mansoor N, Memmi C, Lynch M, Lynch J. Generative pretrained transformer-4, an artificial intelligence text predictive model, has a high capability for passing novel written radiology exam questions. Int J Comput Assist Radiol Surg 2024;19:645—53.
34. Hirano Y, Hanaoka S, Nakao T, Miki S, Kikuchi T, Nakamura Y, et al. GPT-4 Turbo with vision fails to outperform text-only GPT-4 Turbo in the Japan diagnostic radiology board examination. Jpn J Radiol 2024;42:918—26.
35. 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.
36. Kollitsch L, Eredics K, Marszalek M, Rauchenwald M, Brookman-May SD, Burger M, et al. How does artificial intelligence master urological board examinations? A comparative analysis of different Large Language Models' accuracy and reliability in the 2022 In-Service Assessment of the European Board of Urology. World J Urol 2024;42:20.
37. 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.
38. Knoedler L, Knoedler S, Hoch CC, Prantl L, Frank K, Soiderer L, et al. In-depth analysis of ChatGPT's performance based on specific signaling words and phrases in the question stem of 2377 USMLE step 1 style questions. Sci Rep 2024;14:13553. https://doi.org/10.1038/s41598-024-63997-7.
39. 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 Inf 2023;180:105283. https://doi.org/10.1016/j.ijmedinf.2023.105283.

40. Wu J, Wu X, Qiu Z, Li M, Lin S, Zhang Y, et al. Large language models leverage external knowledge to extend clinical insight beyond language boundaries. J Am Med Inf Assoc 2024;31:2054—64.
41. 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.
42. Cheong RCT, Pang KP, Unadkat S, Mcneillis V, Williamson A, Joseph J, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol 2024;281:2137—43.
43. Noda R, Izaki Y, Kitano F, Komatsu J, Ichikawa D, Shibagaki Y. Performance of ChatGPT and Bard in self-assessment questions for nephrology board renewal. Clin Exp Nephrol 2024;28:465—9.
44. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023;93:1353—65.
45. Ozeri DJ, Cohen A, Bacharach N, Ukashi O, Oppenheim A. Performance of ChatGPT in Israeli Hebrew internal medicine national residency exam. Isr Med Assoc J 2024;26:86—8.
46. Cohen A, Alter R, Lessans N, Meyer R, Brezinov Y, Levin G. Performance of ChatGPT in Israeli Hebrew OBGYN national residency examinations. Arch Gynecol Obstet 2023;308:1797—802.
47. 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.
48. van Nuland M, Erdogan A, Açar C, Contrucci R, Hilbrants S, Maanach L, et al. Performance of ChatGPT on factual knowledge questions regarding clinical pharmacy. J Clin Pharmacol 2024;64:1095—100.
49. Kim SE, Lee JH, Choi BS, Han HS, Lee MC, Ro DH. Performance of ChatGPT on solving orthopedic board-style questions: a comparative analysis of ChatGPT 3.5 and ChatGPT 4. CIOS Clin Orthop Surg 2024;16:669—73.
50. Moglia A, Georgiou K, Cerveri P, Mainardi L, Satava RM, Cuschieri A. Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. Artif Intell Rev 2024;57:231.
51. 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.
52. Vij O, Calver H, Myall N, Dey M, Kouranloo K. Evaluating the competency of ChatGPT in MRCP Part 1 and a systematic literature review of its capabilities in postgraduate medical assessments. PLOS ONE 2024. https://doi.org/10.1371/journal.pone.0307372.
53. Bartoli A, May AT, Al-Awadhi A, Schaller K. Probing artificial intelligence in neurosurgical training: ChatGPT takes a neurosurgical residents written exam. Brain Spine 2024;4:102715. https://doi.org/10.1016/j.bas.2023.102715.
54. Ayub I, Hamann D, Hamann CR, Davis MJ. Exploring the potential and limitations of chat generative pre-trained transformer (ChatGPT) in generating board-style dermatology questions: a qualitative analysis. Cureus 2023;15:e43717. https://doi.org/10.7759/cureus.43717.
55. Klang E, Portugez S, Gross R, Kassif Lerner R, Brenner A, Gilboa M, et al. Advantages and pitfalls in utilizing artificial intelligence for crafting medical examinations: a medical education pilot study with GPT-4. BMC Med Educ 2023;23:772.
56. Agarwal M, Sharma P, Goswami A. Analysing the applicability of ChatGPT, bard, and bing to generate reasoning-based Multiple_Choice questions in medical physiology. Cureus 2023;15:e40977. https://doi.org/10.7759/cureus.40977.
57. Zuckerman M, Flood R, Tan RJB, Kelp N, Ecker DJ, Menke J, et al. ChatGPT for assessment writing. Med Teach 2023;45:1224—7.
58. Kıyak YS, Coşkun Ö, Budakoğlu Iİ, Uluoğlu C. ChatGPT for generating multiple-choice questions: evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. Eur J Clin Pharmacol 2024;80:729—35.
59. Coşkun Ö, Kıyak YS, Budakoğlu Iİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach 2024:1—7. https://doi.org/10.1080/0142159X.2024.2327477.
60. Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S. A.R., Singapore, Ireland, and the United Kingdom). PLoS One 2023;18:e0290691. https://doi.org/10.1371/journal.pone.0290691.
61. Grévisse C, Pavlou MAS, Schneider JG. Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. SN Comput Sci 2024;5:636.
62. Rivera-Rosas CN, Calleja-López JRT, Ruibal-Tavares E, Villanueva-Neri A, Flores-Felix CM, Trujillo-López S. Exploring the potential of ChatGPT to create multiple-choice question exams. Educ Méd 2024;25:100930. https://doi.org/10.1016/j.edumed.2024.100930.
63. Laupichler MC, Rother JF, Grunwald Kadow IC, Ahmadi S, Raupach T. Large Language models in medical education: comparing ChatGPT- to human-generated exam questions. Acad Med 2024;99:508—12.
64. 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. PLoS Med 2021;18:e1003583. https://doi.org/10.1371/journal.pmed.1003583.
65. Gandhi AP, Joesph FK, Rajagopal V, Aparnavi P, Katkuri S, Dayama S, et al. Performance of ChatGPT on the India undergraduate community medicine examination: cross-sectional study. JMIR Form Res 2024;8:e49964. https://doi.org/10.2196/49964.
66. Mousavi M, Shafiee S, Harley JM, Cheung JCK, Abbasgholizadeh Rahimi S. Performance of generative pre-trained transformers (GPTs) in certification examination of the college of family physicians of Canada. Fam Med Community Health 2024;12:e002626. https://doi.org/10.1136/fmch-2023-002626.
67. 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.
68. Smith J, Choi PMC, Buntine P. Will code one day run a code? Performance of language models on ACEM primary examinations and implications. EMA Emerg Med Australas 2023;35:876—8.
69. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Benavente Chalco XC, Bustamante Ordoñez MA, 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.
70. Gritti MN, AlTurki H, Farid P, Morgan CT. Progression of an artificial intelligence chatbot (ChatGPT) for pediatric cardiology educational knowledge assessment. Pediatr Cardiol 2024;45:309—13.
71. Nicikowski J, Szczepański M, Miedziaszczyk M, Kudliński B. The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland. Clin Kidney J 2024;17:sfae193. https://doi.org/10.1093/ckj/sfae193.
72. Kufel J, Paszkiewicz I, Bielówka M, Bartnikowska W, Janik M, Stencel M, et al. Will ChatGPT pass the polish specialty exam in radiology and diagnostic imaging? Insights into strengths and limitations. Pol J Radiol 2023;88:e430—4.
73. National Coordinating Council for Medication Error Reporting and Prevention (NCC MERP). NCC MERP taxonomy of medication errors; 1998. Available from: https://www.nccmerp.org/sites/default/files/taxonomy2001-07-31.pdf.
74. Shakurnia A, Aslami M, Bijanzadeh M. The effect of question generation activity on students' learning and perception. J Adv Med Educ Prof 2018;6:70—7.
75. Hutchinson D, Wells J. An inquiry into the effectiveness of student generated MCQs as a method of assessment to improve teaching and learning. Creat Educ 2013;4:117—25.
76. Sanchez-Elez M, Pardines I, Garcia P, Miñana G, Roman S, Sanchez M, et al. Enhancing students' learning process through self-generated tests. J Sci Educ Technol 2014;23:15—25.
77. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: a novel approach to medical knowledge assessment. J Fr Ophtalmol 2023;46:706—11.
78. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ 2024;24:354.
79. Stadler M, Horrer A, Fischer MR. Crafting medical MCQs with generative AI: a how-to guide on leveraging ChatGPT. GMS J Med Educ 2024;41:1—5.
80. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638. https://doi.org/10.2196/50638.
81. Indran IR, Paranthaman P, Gupta N, Mustafa N. Twelve tips to leverage AI for efficient and effective medical question generation: a guide for educators using Chat GPT. Med Teach 2024;46:1021—6.
82. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. Digital Health 2024;10:1—8.
83. 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 2024;31:442—50.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/gme-2024-0021).