Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing

Anusha Sumbal, Ramish Sumbal and Alina Amir

Dow University of Health Sciences, Karachi, Pakistan.

Journal of Medical Education and Curricular Development
Volume 11: 1–12
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23821205241238641
Sage

ABSTRACT

OBJECTIVE: We, therefore, aim to conduct a systematic review to assess the academic potential of ChatGPT-3.5, along with its strengths and limitations when giving medical exams.

METHOD: Following PRISMA guidelines, a systemic search of the literature was performed using electronic databases PUBMED/MEDLINE, Google Scholar, and Cochrane. Articles from their inception till April 4, 2023, were queried. A formal narrative analysis was conducted by systematically arranging similarities and differences between individual findings together.

RESULTS: After rigorous screening, 12 articles underwent this review. All the selected papers assessed the academic performance of ChatGPT-3.5. One study compared the performance of ChatGPT-3.5 with the performance of ChatGPT-4 when giving a medical exam. Overall, ChatGPT performed well in 4 tests, averaged in 4 tests, and performed badly in 4 tests. ChatGPT's performance was directly proportional to the level of the questions' difficulty but was unremarkable on whether the questions were binary, descriptive, or MCQ-based. ChatGPT's explanation, reasoning, memory, and accuracy were remarkably good, whereas it failed to understand image-based questions, and lacked insight and critical thinking.

CONCLUSION: ChatGPT-3.5 performed satisfactorily in the exams it took as an examinee. However, there is a need for future related studies to fully explore the potential of ChatGPT in medical education.

KEYWORDS: ChatGPT, academic performance, medical education, artificial intelligence, digital health, medicine

RECEIVED: October 31, 2023. **ACCEPTED:** February 25, 2024.

TYPE: Review

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Anusha Sumbal, Dow University of Health Sciences, Baba-E-Urdu Road, Karachi, 74200, Pakistan.
Email: anusha.sumbal@gmail.com

Introduction

In the modern era, artificial intelligence (AI) has taken the world by storm whereby in various fields, huge tasks that were solely performed by humans previously, are easily performed by AI-assisted software and robots expeditiously.[1] ChatGPT, is a prime example, being the latest tool among a class of large language models (LLMs) known as autoregressive language models that have been developed by open AI in November 2022.[2,3] ChatGPT is a freely accessible conversational AI tool designed on the concept of reinforcement learning from human feedback.[4] ChatGPT has vast applications in education, research writing, scientific advancement, diagnostics, disease management, treatment regimes, and safety measures.[5–8] One of its popular uses includes academic assessment and assistance among medical students.[1] Recently, ChatGPT successfully simulated clinical scenarios, enabling medical students to practice new elements of clinical curriculum along with therapeutic insights through patient interactions.[9]

The potential use of ChatGPT as a clinical tool is quite explicit as it has manifested the ability to answer clinical questions accurately in simple English, which can be understood by both healthcare providers and patients.[3] A study in 2023 by Gilson et al[2] found that ChatGPT is capable of correctly answering up to over 60% of questions representing topics covered in the USMLE Step 1 and Step 2 licensing exams. Also, ChatGPT accounts for the ability to generate new insights based on previously fed knowledge and past experiences.[10] These findings have opened an avenue for unveiling the credibility of ChatGPT when taking medical exams. However, the results of a recent study highlighted the poor performance of ChatGPT in the AHA ACLS exam.[11] ChatGPT failed to reach the passing threshold of the exam due to robustness and limited critical thinking.[10,11] A recent meta-analysis conducted by Levin et al[12] also tried to study the accuracy of ChatGPT when giving medical exams with multiple-choice questions. However, the study showed great variation in the result and it remained inconclusive. ChatGPT-3.5 is a very new groundbreaking platform among LLM and launched not before than November 2022. Although, recent articles exist that have studied the efficiency and application of ChatGPT and multiple ongoing research are being conducted on it, but still available information currently is limited. Existing studies show contradiction to each other and are inconclusive. At this point in time, we believe that the true potential of ChatGPT in academics especially in medical exams is still unclear.

With such an increasing trend in the dependency of students on ChatGPT and its use as a prospering educational tool in medicine, we believe that there is a dire need to weigh the efficiency and productivity of ChatGPT as an examinee itself.[5] Medical students and healthcare providers need to evaluate the accuracy of medical information generated by AI to ensure that valid information is available for patients and the public. Moreover, it is necessary to provide not just quantitative but also, qualitative feedback on the academic performance of ChatGPT to characterize it as a useful tool for medical students and clinicians. We need to evaluate how accurately ChatGPT can solve questions on medical examinations. This would help to ensure that via ChatGPT reliable information is accessible to patients and provide insights into how medical students could use ChatGPT for their learning.[13]

We, therefore, aim to conduct a systematic review to determine the overall academic performance of ChatGPT-3.5 when giving various medical exams. Additionally, we also aim to identify the academic strengths and limitations of ChatGPT-3.5 when attempting the medical exam.

Materials and method

Study design

This systematic review was performed following preferred reporting items for systematic review and meta-analysis (PRISMA) guidelines.[14]

Data sources and search strategy

A systemic and thorough search of the literature was performed by AS using electronic databases PUBMED/MEDLINE, Google Scholar, and Cochrane. Articles from their inception till April 29, 2023, were queried. Our search strategy mainly consisted of the following keywords in combination with Medical Subject Headings (MeSH) terms and text words ((chatGPT OR chatgpt OR CHATgpt OR Generative Pre-trained Transformer OR GPT-4 OR GPT-3)) AND ((academic performance OR academic* OR score OR exam* OR academic grad* OR test* OR licensing exam* OR USMLE* OR medical licensing exam OR medical exam* OR perform* OR reasoning OR question* OR response* OR evaluat* OR multiple choice exam OR exam* OR analysis OR choice exam OR scenerio based*)) (detailed search strategy and research question are provided Supplemental Tables 1 and 2, respectively). Protocol in PROSPERO could not be registered because the study is not directly related to human health.

We included all the available and related articles in the English language only. Animal studies, commentary, conference abstracts, and articles with no full text available were removed. After de-duplication, title screening of the acquired articles was performed. Eligible articles underwent full-text screening. This screening was independently performed by AA and AS, any discrepancy was resolved by consultation from RS.

Study selection and inclusion criteria:

Studies meeting our predefined inclusion criteria were selected. According to our inclusion criteria we included (a) all available published scientific research papers or preprint, (b) conducted on ChatGPT, (c) evaluating its academic performance in any manner, for example, marks obtained, whether passed or failed, etc and (d) particularly in medical examinations and tests.

We excluded records that used any other artificial intelligence platform than ChatGPT, studies assessing nonacademic or other capabilities of ChatGPT, examinations not related to medicine/and or healthcare, and studies not mentioning results of academic exams given by ChatGPT.

Quality assessment and risk of bias in studies

We could not perform quality assessment and risk of bias in the included studies because selected studies have no human subjects involved.

Data extraction

All the articles queried were exported to Endnote Reference Library software version 20 (Clarivate Analytics). After a rigorous screening process, articles meeting predefined inclusion criteria were selected. Desired data was extracted using data extraction form from each study. The following key information was extracted:

- Type of study, authors, time, and duration of the study
- Type of examination in which ChatGPT appeared (USMLE step1, USMLE step2, NBME, toxicology, parasitology, genetic test, etc)
- Exam-related data (total number of questions asked, no. of correct/wrong answers no. of attempts made, no. of subsets, etc)
- Model of examination (self-assessment, Multiple-choice questions, scenario-based questions, etc)
- Academic performance of ChatGPT (marks obtained, passed or failed, percentage, etc)
- Academic strengths (critical decision-making, logical reasoning, deep learning, etc)
- Academic limitations (automation bias, lack of insight, failure to interpret figures/tables, etc)
- Key result

Statistical analysis

We evaluated and summarized findings from all the studies, along with possible reasoning behind the obtained results. A formal narrative analysis was conducted by systematically arranging similarities and differences between individual findings together. Conclusions drawn from the obtained data were incorporated subsequently. Due to significant paucity in the study, pooling of the acquired data and its meta-analysis was not possible since it would have produced misleading results.

We will classify the performance of ChatGPT based on the conclusion stated by the individual study included in the review. We will consider its performance to be bad if a study reports that ChatGPT failed to reach the passing threshold, an average of a study states that ChatGPT's performance was near equal or equal to the passing threshold, and good when a study states that ChatGPT's performance was not only greater than passing threshold but it passed with distinction or considerably better than an average human score in the same exam.

Results

Literature search

A total of 375 articles were selected after the initial search. After removing duplicates and performing title/abstract screening, we included a total of 20 articles for full-text review. Finally, a total of 12 articles were included in this review[1,2,4,10,11,13,15–20] (Figure 1).

Study characteristics

A total of 12 studies evaluating ChatGPT performance in different exams were selected.[1,2,4,10,11,13,15–20] All the selected papers assessed the academic performance of ChatGPT-3.5. One study compared the performance of ChatGPT-3.5 with the performance of ChatGPT-4 when giving a medical exam. Out of these 12 studies, 5 studies were preprint,[13,15–17,19] 3 were letters to editors,[4,11,20] 2 were original studies,[1,2], 1 was a cross-sectional study,[10] and 1 study was descriptive.[18] These studies were conducted across different countries. Out of these 12 studies, 7 were conducted in the United States of America,[1,2,11,13,15–17] 2 were conducted in India,[4,10] while only 1 study was conducted in Belgium,[20] Ireland,[2] Canada,[19] and Korea.[18] The main characteristics of the included studies have been displayed in Table 1.

ChatGPT's overall performance in academic testing

Twelve studies have evaluated the potential of ChatGPT-3.5 in various academic tests.[1,2,4,10,11,13,15–20] Out of these tests, ChatGPT-3.5 performed very well in 4 tests,[4,10,13,17] average in 4 tests,[1,2,15,16] and performed badly in 4 tests.[11,18–20]

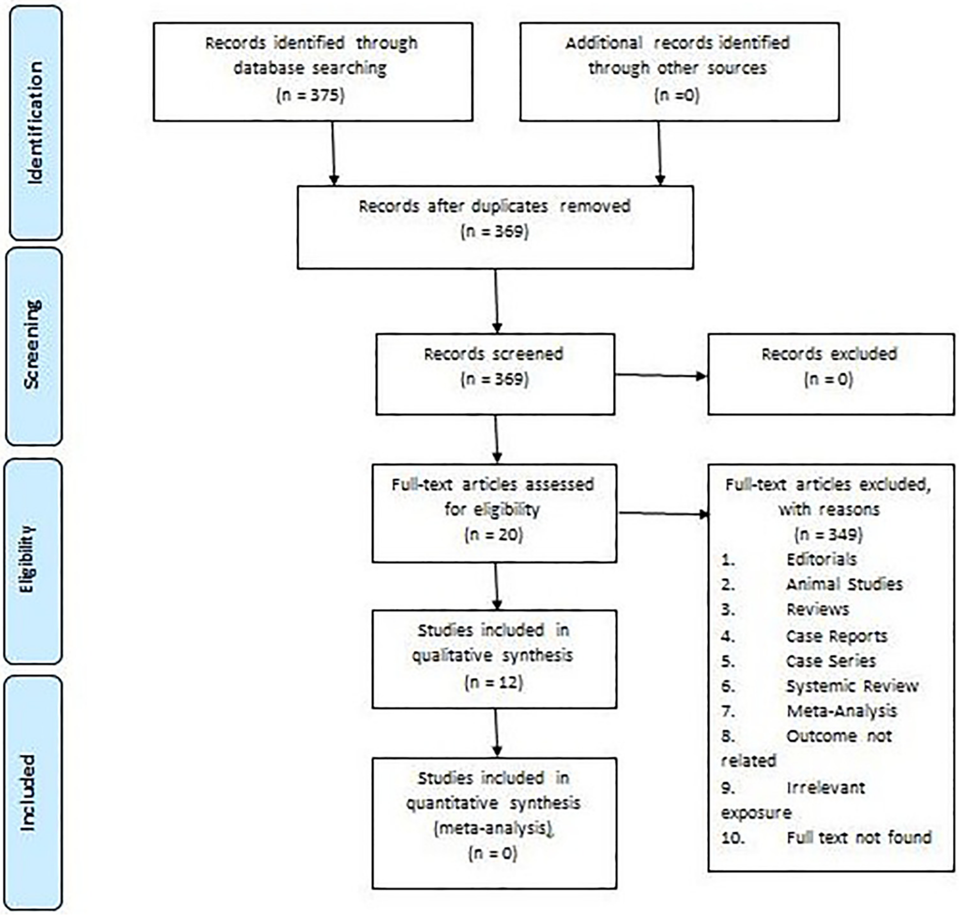
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 375)
Additional records identified through other sources (n =0)
Records after duplicates removed (n = 369)
Records screened (n = 369)
Records excluded (n = 0)
Full-text articles assessed for eligibility (n = 20)
Full-text articles excluded, with reasons (n = 349)
1. Editorials
2. Animal Studies
3. Reviews
4. Case Reports
5. Case Series
6. Systemic Review
7. Meta-Analysis
8. Outcome not related
9. Irrelevant exposure
10. Full text not found
Studies included in qualitative synthesis (n = 12)
Studies included in quantitative synthesis (meta-analysis) (n = 0)

Figure 1. Preferred reporting items for systematic review and meta-analysis (PRISMA) flowchart.

Table 1. Main characteristics of included studies.

NAME OF STUDY	YEAR OF STUDY	COUNTRY	STUDY DESIGN	FIELD OF EXAMINATION	MODEL OF EXAMINATION	ASSESSED FOR	COMPARED WITH	RESULT PRESENTATION
Johnson et al[13]	2023	USA	Preprint	Medical, surgical, and pediatrics specialties	Descriptive questions, binary questions	Accuracy (correct answer) SS, completeness	N/A	Median, mean, interquartile range, standard deviation
Ali et al[15]	2023	USA	Preprint	Self-Assessment Neurosurgery Exams (SANS) American Board of Neurological Surgery (ABNS) Self-Assessment Exam 1	Self-assessment, single best answer MCQ format, imaging-based questions	First-order questions (simple fact recall), Higher-order questions (diagnosis along with evaluative or analytical tasks)	Question bank users, and GPT-4	Percentage with 95% confidence interval
Subramani et al[4]	2023	India	Letter to editor	Medical Physiology Examination of Phase I MBBS, Physiology University Examination	Theory questions, multiple choice questions (MCQs) and essay (case-based question)	Correct answers, Case diagnoses with subset case-related questions	N/A	Score and percentage obtained
Sinha et al[10]	2023	India	Cross-sectional study	One hundred structure pathology questions from a free website	Evaluated on a 0-5 scale, the structure of the observed learning outcome taxonomy for evaluating the answers.	higher-order reasoning, rote memorization, underlying concepts and principles,	N/A	Number of correct answers, mean, median, standard deviation, and first and third quartile.
Duong et al[16]	2023	USA	Preprint	Genetics and genomics	Multiple choice questions	General knowledge of genetics, clinical genetics, diagnosis, management, risk calculations, memorization, critical thinking	Human respondents on 2 social media platforms, Twitter and Mastodon.	Number of correct answers and their percentage
Hou and Ji[17]	2023	USA	Preprint	Genetics and genomics,	GeneTuring, a comprehensive question-and-answer (Q&A) database	Gene name extraction, gene alias, gene name conversion, gene location, SNP location and association, a protein-coding gene, genetic diseases association, gene ontology, TF regulation, DNA sequence alignment to human and multiple species,	Other versions of ChatGPT	Score (average score in all modules)
Huh[18]	2023 January 1, 2023	Korea	Descriptive study	Parasitology examination	Multiple-choice question, case-based questions	Correct answers along with the item's knowledge level and acceptability of explanation	77 medical students on the parasitology examination and ChatGPT were compared	Correct answer and percentage
Gilson et al[2]	2023	Ireland	Original article	USMLE[a] Step 1 and Step 2, AMBOSS,	Question text and direct questions along with	Logical reasoning, internal information (from question),	Compared with	Correct answer and percentage

(continued)

Table 1. Continued.

NAME OF STUDY	YEAR OF STUDY	COUNTRY	STUDY DESIGN	FIELD OF EXAMINATION	MODEL OF EXAMINATION	ASSESSED FOR	COMPARED WITH	RESULT PRESENTATION
				NBME[b]-Free-Step1 and NBME-Free-Step2	multiple-choice answers, clinical knowledge questions	external information (outside question)	GPT-3 and InstructGPT	
Kung et al[1]	2023	USA	Original article	USMLE step1, step2 CK,[c] step 3	Open-ended, MCQ, MCQ with justification (no difference in prompt types)	Accuracy, Concordance, and Insight (ACI)	N/A	Correct answer and percentage
Fijačko et al[11]	2023	USA	Letter to editor	AHA BLS[d] Exams AHA ACLS[e] exam	scenario-based question,	Correct answer and level of correctness	N/A	The correct answer, percentage with a confidence interval
Antaki et al[19]	2023	Canada	Preprint	BCSC[f] Self-Assessment Program and Ophtho Questions, OKAP exam	Self-assessment, multiple-choice format	Low cognitive level (recall of facts and concepts), high cognitive level, (interpreting data, making calculations, and managing patients, best treatment)	N/A	Correct answer and percentage
Morreel et al[20]	2023	Belgium	Letter to editor	Family Medicine in the third year MBBS	One multiple-choice exam	The Natural Prompt (“Give one single answer”) and the Rank Prompt (“Rank the possible answers”; the first ranked answer was used).	N/A	Correct answer and percentage

[a] United States Medical Licensing Examination.
[b] National Board of Medical Examiners.
[c] Clinical Knowledge.
[d] American Heart Association Basic Life Support.
[e] American Heart Association Advance Cardiovascular Life Support.
[f] Basic and Clinical Sciences Courses.

In the study by Sinha et al,[10] ChatGPT performed reasonably achieving a median score of 4 out of 5. Johnson et al[13] evaluated ChatGPT's performance across 17 different specialties and found that ChatGPT performed well both in accuracy and completeness. When ChatGPT was asked to take a first-year physiology exam, ChatGPT cleared it with distinction.[4] However, there were a few studies in which ChatGPT did not perform well. In a parasitology exam, ChatGPT scored 60.8% as compared to the 90.8% score by first-year medical students.[18] Similarly, ChatGPT failed the American Heart Association BLS and ACS examinations.[11] In other studies, ChatGPT significantly underperformed as compared to their human counterpart.[15,20] Along with these studies, there were a few studies in which ChatGPT performed satisfactorily.[1,2,15,16] Two studies evaluated ChatGPT's performance in USMLE examinations.[1,2] ChatGPT barely reached the passing threshold in both studies. Similarly, ChatGPT scored 68.2% as compared to 69.3% by the human respondents in a genetic examination.[16] In another study, ChatGPT scored nearly the same as examinees giving neurosurgical board examinations.[15] Some studies compared ChatGPT's performance with other AI models.[15,17] ChatGPT outperformed its previous models like GPT-1, GPT-2, and GPT-3. However, it could not score as well as GPT-4 or New Bing models. The academic performance of ChatGPT is displayed in Figure 2. The academic potential and limitations of ChatGPT have been comprehensively stated in Figures 3 and 4, respectively. The results of the systematic review have been mentioned in Table 2. Also, a comparison of

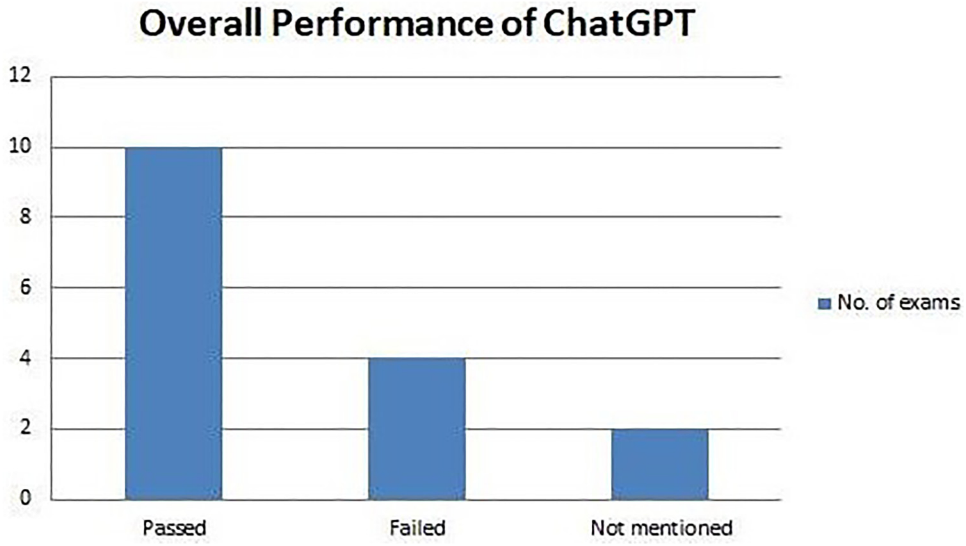

Figure 2. Overall performance of ChatGPT.

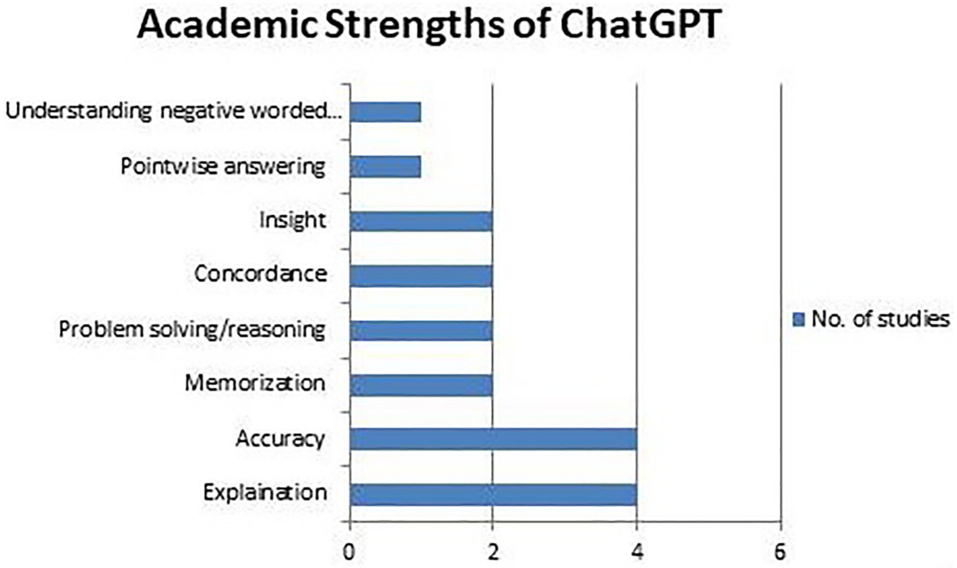

Figure 3. Academic strengths of ChatGPT.

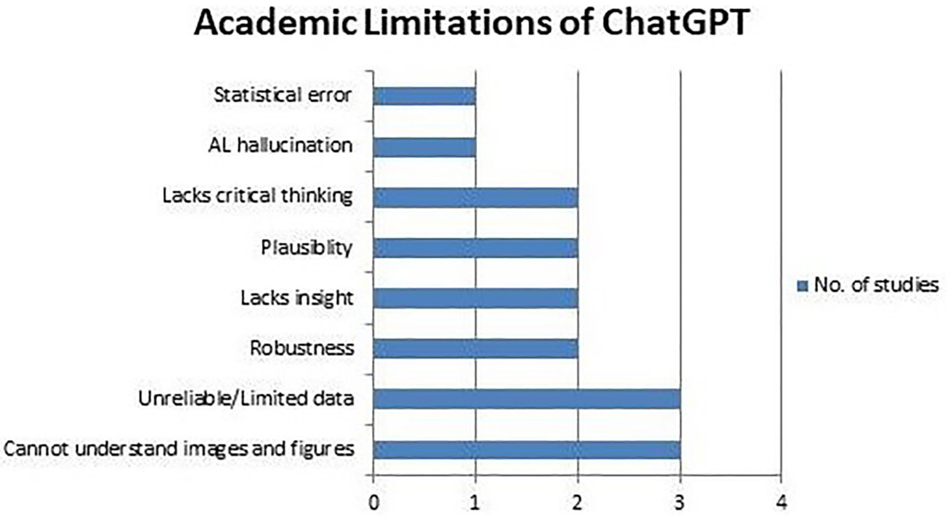

Figure 4. Academic limitations of ChatGPT.

ChatGPT's performance with a human counterpart is mentioned in detail in Table 3.

ChatGPT's performance based on question difficulty

Three studies[2,4,19] have studied the impact of the question's difficulty on ChatGPT's performance. Gilson et al[2] and Antaki et al[19] found that ChatGPT's performance was directly proportional to the level of the questions' difficulty. On the contrary, Johnson et al[13] found that there was no significant difference in ChatGPT's performance based on difficulty levels. Details of ChatGPT's performance based on question difficulty are mentioned in Table 4.

ChatGPT's performance across various question types

A total of 5 studies[1,10,16,18,20] evaluated the impact of various types of questions on ChatGPT's performance. Johnson et al[13] found that there was no difference in ChatGPT's performance whether the questions were binary, descriptive, or MCQ-based. Similarly, Kung et al[1] found that there was no difference in ChatGPT's performance either on open-ended questions, MCQS, or MCQS with justifications. Huh[18] concludes that there was no difference in performance when ChatGPT answered questions based on recall, interpretation, and problem-solving. However, Duong and Solomon[16] found that ChatGPT performed poorly on critical thinking questions compared to questions requiring memorization. On the other hand, Morreel et al[20] found that ChatGPT performed badly on straight questions and scored well on negative or complex questions. Details of ChatGPT's performance based on question type are mentioned in Table 5.

Efficacy of ChatGPT's explanations

ChatGPT's explanation was assessed in 4 studies,[4,11,18,20] and all the authors found that they were effective. Fijačko et al[11] found that the rationale in ChatGPT's explanation was more detailed than the AHA answer key. Similarly, Morreel et al[20] found that ChatGPT gave confident and clear explanations even for wrong answers. Subramani et al[4] identified that along with giving a correct diagnosis, ChatGPT can structure its explanation in a point-wise manner. Also, it provides relevant examples and important background information to make its explanation more comprehensible to a human learner. Similarly, Huh[18] highlights that ChatGPT provides a deep understanding and practical interpretation in the explanation it provides.

Discussion

Through this review, we found that there is a mixed result regarding ChatGPT's performance in various medical tests. ChatGPT performed better in some tests and average in others while it failed to reach the passing threshold in the remaining exams. We further analyzed ChatGPT's performance based on the question's structure and difficulty level. We found that ChatGPT performed poorly when answering difficult questions. However, there was no difference in the results when the questions were either MCQ-type or open-ended questions.

A recent meta-analysis conducted by Levin et al[12] assessed ChatGPT's performance in medical examinations with multiple-choice questions. They found that ChatGPT's academic knowledge ranged from 40% in the biomedical admission test to 100% in a diabetes knowledge questionnaire.[12] The inconsistency in ChatGPT's performance raised concerns

Table 2. Results of systematic review.

NAME OF STUDY	OVERALL SCORE	KEY RESULT	ACADEMIC STRENGTHS	ACADEMIC LIMITATION
Johnson et al[13]	Median accuracy score: 5.5 (median 4.8, SD 1.6, and IQR 2), Median completeness score: 3 (mean 2.5, SD 0.7, and IQR 1)	ChatGPT passed the exam, with high accuracy and completeness scores	High accuracy and completeness scores across various specialties, question types, and difficulty levels	Lack of reliability and robustness in clinical integration
Ali et al[15]	73.5% (95% CI 69.3%-77.2%)	ChatGPT passed the exam, with GPT-4 significantly outperforming ChatGPT.	Good at functional neurosurgery, peripheral nerve surgery, and first-order recall.	Poor performance in imaging-based questions
Subramani et al[4]	> 75%	ChatGPT passed exam	Correct diagnosis, answers in a point-wise manner, appropriate examples, and important information that examiners look for.	Could not generate relevant diagrams for the answers simultaneously
Sinha et al[10]	Median score 4.08 (Q1-Q3: 4-4.33).	ChatGPT passed the exam with a score of 80%	Good at recognizing patterns, classifying data, and generating new information based on existing data	Lacks true insight, lack subjective judgment, plausible answers, lacks the ability to understand personal values and biases, cannot create ideas without human input
Duong et al[16]	68.2%	ChatGPT passed the exam but did not perform significantly differently than human respondents	Good at memorization-type questions	Poor at critical thinking, plausible answers, and explanations
Hou and Ji[17]	N/A	ChatGPT performed better than all the other AI platforms, except the New Bing Model	Overall better performance than GPT-2, BioGPT, BioMedLM, GPT-3, and ChatGPT	AI hallucination
Huh[18]	60.8%	ChatGPT's performance was lower than that of medical students.	Good recall, problem-solving, and interpretation	Plausible answers, unable to interpret figures, graphs, and tables some epidemiological data unique to Korea were outside ChatGPT's knowledge
Gilson et al[2]	44% (AMBOSS-STEP1) 42% (AMBOSS-STEP2) 64.4% (NBME[a]-FREE-STEP1) 57.8% (NBME-FREE-STEP2)	ChatGPT failed AMBOSS Steps 1 and 2 but passed NMBE-free-step 1 and 2 exam. Overall, ChatGPT performs at a level expected of a third-year medical student	Gives a logical explanation of its answer selection, has the ability to understand the context and carry on a conversation, and shows coherence	Frequent logical and information errors. Few statistical errors.
Kung et al[1]	For open-ended questions: 45.4% USMLE[b] STEP-1 54.1% USMLE STEP-2CK[c] 61.5% USMLE STEP-3 For MCQ with no forced justification: 36.1% USMLE STEP-1 56.9% USMLE STEP-2CK 55.7% USMLE STEP-3 For MCQs with forced justification: 41.2%	ChatGPT passed the exams with a near-threshold of 60% accuracy	Good accuracy and analyses. Gives prompt tuning, high concordance, role-modeling deductive reasoning	Missing information, leading to diminished insight and indecision

(continued)

Table 2. Continued.

NAME OF STUDY	OVERALL SCORE	KEY RESULT	ACADEMIC STRENGTHS	ACADEMIC LIMITATION
	USMLE STEP-1 49.5% USMLE STEP-2CK 59.8% USMLE STEP-3			
Fijačko et al[11]	66.0% AHA BLS[d] exam 72.37% AHA ACLS[e] exam	ChatGPT did not reach the passing threshold (> 75%) for any of the exams.	Provides insightful explanations to support the given answer, high relevancy, and accuracy showed significantly better concordance	The reference provided by ChatGPT for each answer was very general the
Antaki et al[19]	55.8% BCSC[f] 42.7% Optho questions	Passed BCSC exam but failed Optho exam	Concordance and insight	Lacks the capability to process images
Morreel et al[20]	Scored 8/20; with Rank Prompt this increased to 10/20	ChatGPT passed the family medicine exam	ChatGPT performed better on negative-worded questions with confidence and clear explanations	N/A

[a] National Board of Medical Examiners.
[b] United States Medical Licensing Examination.
[c] Clinical Knowledge.
[d] American Heart Association Basic Life Support.
[e] American Heart Association Advance Cardiovascular Life Support.
[f] Basic and Clinical Sciences Courses.

about the overall reliability of the data. Such a variation could be possible because this study was limited to multiple-choice exams only. Also, the study failed to appreciate an in-depth qualitative analysis of the performance of ChatGPT in individual exams.[12]

ChatGPT is a large language model (LLM) that is trained on a specific dataset to generate human-like responses.[21] Although ChatGPT didn't perform well in some tests, its overall performance was much better than previous models. This is seen as a huge achievement in LLM development since ChatGPT can also generate plausible explanations for its answers.[1,16,19] Another feature that differentiates ChatGPT from its predecessors is that users can ask follow-up questions. This will further enhance the quality of answers generated. We also found that ChatGPT's performance was comparable to its human counterpart on various exams.

Apart from ChatGPT's overall performance in these tests, we also evaluated the impact of question types on these results. We found that ChatGPT's performance wasn't affected by whether questions were in MCQ format or descriptive type. This means that questions of all types will yield similar results. Therefore there will be no need for any special prompts to get desired results. In addition, we found that explanations generated by ChatGPT were very concordant and succinct. A human learner can easily understand the logic and rationale behind every explanation. Furthermore, Kung et al[1] found that nearly 90% of ChatGPT's answers consists of valuable insight. All of this makes ChatGPT an important learning tool.

Although ChatGPT showed promising results in these tests, our review highlighted a few weaknesses of ChatGPT. ChatGPT's accuracy has improved from its earlier version;

Table 3. ChatGPT's performance compared with a human counterpart.

STUDY NAME	CHATGPT'S PERFORMANCE COMPARED WITH A HUMAN COUNTERPART
Johnson et al[13]	N/A
Ali et al[15]	ChatGPT obtained 73.4% nearly the same as a human user obtaining 73.7%. ChatGPT's score was better than last year passing the threshold that was 69%
Subramani et al[4]	Cleared with distinction
Sinha et al[10]	N/A
Duong et al[16]	ChatGPT scored 68.2% which was poorer than human respondents 69.3%
Hou and Ji[17]	ChatGPT performed better than other AI models but poorer compared to the new Bing. Comparison with human counterpart was not available
Huh[18]	ChatGPT's score was 60.8% as compared to 90% secured by human examinee
Gilson et al[2]	N/A
Kung et al[1]	Near equal passing threshold
Fijačko et al[11]	Could not reach the average passing threshold
Antaki et al[19]	ChatGPT's performance was poorer than a human counterpart in both of the exams. ChatGPT scored 55.8% on the BCSC set, was a human scored 74% ChatGPT scored 42.7% on the Ophtho Questions, whereas average humans scored 61%
Morreel et al[20]	ChatGPt's performance was poorer than its human counterpart. It scored 8/20 on the exam whereas a normal student scored 15/20

Table 4. ChatGPT's performance is based on difficulty level.

NAME OF STUDY	CHATGPT-3.5 PERFORMANCE BASED ON DIFFICULTY LEVEL			*P* VALUE
	EASY	MEDIUM	DIFFICULT	
Johnson et al[13]	Median accuracy score: 5 (mean 4.6, SD 1.7, and IQR 3), Median completeness score: 3 (mean 2.6, SD 0.7, and IQR 1)	Median accuracy score: 5 (mean 4.3, SD 1.7, and IQR 3) Median completeness score: 3 (mean 2.4, SD 0.7, and IQR 1)	Median accuracy score: 5 (mean 4.2, SD 1.8, and IQR 3.8) Median completeness score: 2.5 (mean 2.4, 0.7, and IQR 1)	For accuracy Kruskal Wallis *P* = .4 For completeness Kruskal Wallis *P* = .3
Gilson et al[2]	Step 1			Step 1
	Level 1[a]: 64.3% Level 2: 59.3%	Level 3: 40.6%	Level 4: 33.3% Level 5: 0%	.01
	Step 2			Step 2
	Level 1: 60% Level 2: 43.5%	Level 3: 40.7%	Level 4: 18.8% Level 5: 13.3%	.13
Antaki et al[19]	BSCS exam			N/A
	65% to 77%	40% to 52%	18% to 28%	
	Optho exam			
	56% to 76%	23% to 39%	7% to 12%	

[a]Levels 1 and 2 were assumed to be easy, level 3 was assumed to be medium, and levels 4 and 5 were assumed to be difficult in level by the author.

however, it is still far from being completely accurate. ChatGPT generated wrong answers on various occasions and that too in a convincing manner. Researchers believed that ChatGPT could not currently be used unsupervised.[22] ChatGPT-3.5 doesn't support visual input.[15,22] This means that ChatGPT has limited knowledge regarding information present on visual diagrams. However, ChatGPT-4 currently includes support for image capabilities, although they are in the initial stages, and the performance may not be deemed impressive.[23] Our review highlighted that ChatGPT performed considerably poorly on questions that were associated with visual input. Similarly, the ChatGPT's performance was highly affected by the type of examination conducted. ChatGPT scored good marks in a first-year physiology test as compared to the USMLE examination. Since the latter requires a more complex approach.

Along with academic tests, researchers have also explored ChatGPT's performance in other academic fields.[24,25] A lot of studies have been conducted to evaluate the role of ChatGPT in academic research.[24,25] It is found that ChatGPT can conceptualize and propose a research topic.[26] Furthermore, ChatGPT can also write the manuscript. With all these advantages, there were certain flaws found during the process.[24,27,28] There were certain ethical considerations attached to ChatGPT being an author. The manuscript prepared by ChatGPT has plagiarism issues.[27,28] The dataset used by ChatGPT was not updated after 2021.[29] In addition, citations generated by ChatGPT were also fake.[27] With all these points many people now recommend the careful and supervised use of ChatGPT in research.[22]

Our research opens up a lot of prospects. Through our research, we propose that ChatGPT could be incorporated into the educational system. But before this, we require clear policy-making in this regard. Along with this, careful consideration should be taken regarding its merits and demerits. ChatGPT is constantly learning and is expected to get better with time. ChatGPT could be used by students for learning in small groups. Evidence shows small group peer learning is equal to teaching by a teacher.[30] Students could use ChatGPT to answer their queries. Explanations provided with these answers will further solidify their concepts. In conclusion, our study provides evidence regarding the correct position of ChatGPT in the academic world and provides a layout of how it can be used in the future.

Our study should be viewed in light of the following limitations. First of all our review included studies from different medical fields but there are still several fields in which ChatGPT's performance is yet to be tested. These studies were conducted in different periods. Also, our review was conducted in a specific time duration; all those studies that met our inclusion criteria in that duration are included in our review. However, we are unable to include studies done after our literature search that were beyond our scope of review. Since ChatGPT is constantly learning and evolving, there is a

Table 5. ChatGPT's performance is based on the type of questions.

NAME OF STUDY	TYPE OF QUESTIONS	SCORE OF CHATGPT-3.5
Johnson et al[13]	Descriptive questions	median accuracy score 5 (mean 4.3, SD 1.7, and IQR 3)
	Binary questions	median accuracy score of 5 (mean 4.5, SD 1.7, and IQR 3),
Huh[18]	Recall question	17/32 (53.0%)
	Interpretation question	20/32 (62.5%)
	Problem-solving	11/15 (73.0%)
Kung et al[1]	Step 1	
	Open-ended	42.0%
	Multiple-choice question	35.0%
	Step 2 CK	
	Open-ended	52%
	Multiple-choice question	51%
	Step 3	
	Open-ended	60%
	Multiple-choice question	50%
Duong et al[16]	Memorization question	80.3%
	Critical thinking	26.3%
Morreel et al[20]	Negative-worded questions	80%
	Regular search engine questions	Failed

chance that the time of the conducted study may have impacted the results. Furthermore, ChatGPT's results depend on the prompt given and there is a risk that results may not be constant when ChatGPT is tested again. However, in our research, we found that ChatGPT's answers are unaffected by the type of questions. But the risk remains. Currently, ChatGPT cannot process visual information so it could not answer questions whose data is given in images or graphs. Lastly, ChatGPT is trained on the dataset before 2021 therefore it is unaware of recent advances in the field.[29]

Conclusion

ChatGPT-3.5 passed most of the exams it took as an examinee. Medical students, trainees, and doctors can safely rely on ChatGPT-3.5 when it comes to academic aid and clinical tools. In light of the academic strength identified by our study, we recommend such similar features should be added to this version of ChatGPT-3.5. Also, the academic limitation of ChatGPT identified by our study should be taken into account and those features must be modified. Such changes would help in the development of a better version in the future, increasing ChatGPT's reliability and efficiency in our healthcare.

Author's Contribution

Miss Anusha Sumbal: conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing-original draft, and writing–review and editing. Dr Ramish Sumbal: formal analysis, investigation, methodology, resources, software, validation, visualization, writing–original draft, and writing–review and editing. Miss Alina Amir: resources, validation, visualization, writing–original draft, and writing–review and editing.

ORCID iD

Anusha Sumbal https://orcid.org/0000-0003-0685-9767

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198.
2. Gilson A, Safranek CW, Huang T, 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(2023):e45312.
3. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023;228(6): 696-705.
4. Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. *Adv Physiol Educ*. 2023;47(2):270-271.
5. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2023.
6. Rahad K, Martin K, Amugo I, et al. ChatGPT to enhance learning in dental education at a historically black medical college. *Res Sq*. 2023.
7. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care*. 2023;27(1):75.
8. Wong RS, Ming LC, Raja Ali RA. The intersection of ChatGPT, clinical medicine, and medical education. *JMIR Med Educ*. 2023;9(2023):e47274.
9. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. *JMIR Med Educ*. 2023;9(2023): e49877.
10. Sinha RK, Deb Roy A, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus*. 2023;15(2):e35237.
11. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American Heart Association course? *Resuscitation*. 2023;185(2023):109732.
12. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. *Bjog*. 2024;131(3):378-380.
13. Johnson D, Goodman R, Patrinely J, et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. *Res Sq*. 2023.
14. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *J Clin Epidemiol*. 2021;134(2021):178-189.
15. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023;93(6):1353-1365.

16. Duong D, Solomon BD. Analysis of large-language model versus human performance for genetics questions. *Eur J Hum Genet*. 2023;31(5):1-3.
17. Hou W, Ji Z. GeneTuring tests GPT models in genomics. bioRxiv. 2023.
18. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*. 2023;20(2023):1.
19. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmol Sci*. 2023;3(4):100324.
20. Morreel S, Mathysen D, Aye VV. AI! ChatGPT passes multiple-choice family medicine exam. *Med Teach*. 2023;45(6):665-666.
21. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307(2):e230163.
22. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):1-20.
23. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *J Am Acad Orthop Surg*. 2023;31(23):1173-1179.
24. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit Health*. 2023;5(3):e105-e1e6.
25. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224-226.
26. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023.
27. ChatGPT LJ. An artificial intelligence chatbot, is impacting medical literature. *Arthroscopy*. 2023;39(5):1121-1122. doi:10.1016/j.arthro.2023.01.015
28. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature*. 2023;614(7947):214-216.
29. Kim S-G. Using ChatGPT for language editing in scientific articles. *Maxillofac Plast Reconstr Surg*. 2023;45(1):13.
30. Rees EL, Quinn PJ, Davies B, Fotheringham V. How does peer teaching compare to faculty teaching? A systematic review and meta-analysis (.). *Med Teach*. 2016; 38(8):829-837.