international medical education

MDPI

Article

Examining the Efficacy of ChatGPT in Marking Short-Answer Assessments in an Undergraduate Medical Program

Leo Morjaria [1], Levi Burns [1], Keyna Bracken [1,2], Anthony J. Levinson [1], Quang N. Ngo [1,2], Mark Lee [2] and Matthew Sibbald [1,2,*]

[1] Michael G. DeGroote School of Medicine, McMaster University, Hamilton, ON L8P 1H6, Canada; leo.morjaria@medportal.ca (L.M.); qngo@mcmaster.ca (Q.N.N.)
[2] McMaster Education Research, Innovation and Theory (MERIT) Program, McMaster University, Hamilton, ON L8P 1H6, Canada
* Correspondence: matthew.sibbald@medportal.ca

Abstract: Traditional approaches to marking short-answer questions face limitations in timeliness, scalability, inter-rater reliability, and faculty time costs. Harnessing generative artificial intelligence (AI) to address some of these shortcomings is attractive. This study aims to validate the use of ChatGPT for evaluating short-answer assessments in an undergraduate medical program. Ten questions from the pre-clerkship medical curriculum were randomly chosen, and for each, six previously marked student answers were collected. These sixty answers were evaluated by ChatGPT in July 2023 under four conditions: with both a rubric and standard, with only a standard, with only a rubric, and with neither. ChatGPT displayed good Spearman correlations with a single human assessor ($r = 0.6–0.7$, $p < 0.001$) across all conditions, with the absence of a standard or rubric yielding the best correlation. Scoring differences were common (65–80%), but score adjustments of more than one point were less frequent (20–38%). Notably, the absence of a rubric resulted in systematically higher scores ($p < 0.001$, partial $\eta^2 = 0.33$). Our findings demonstrate that ChatGPT is a viable, though imperfect, assistant to human assessment, performing comparably to a single expert assessor. This study serves as a foundation for future research on AI-based assessment techniques with potential for further optimization and increased reliability.

Keywords: ChatGPT; artificial intelligence; short-answer assessment; automated grading; generative AI; undergraduate medical education

Citation: Morjaria, L.; Burns, L.; Bracken, K.; Levinson, A.J.; Ngo, Q.N.; Lee, M.; Sibbald, M. Examining the Efficacy of ChatGPT in Marking Short-Answer Assessments in an Undergraduate Medical Program. *Int. Med. Educ.* **2024**, *3*, 32–43. https://doi.org/10.3390/ime3010004

Academic Editor: Hideki Kasuya

Received: 28 November 2023
Revised: 12 January 2024
Accepted: 17 January 2024
Published: 19 January 2024

1. Introduction

The growing influence of artificial intelligence (AI) in medicine is not confined to clinical practice or research techniques; it is also having an exciting and disruptive impact on medical education [1,2]. The use of large language models (LLM.s) has, in particular, drawn consideration for potential use in medical education since the public launch of ChatGPT (OpenAI, CA, USA) in November 2022 [3–5]. While the exact conceptual foundations of these AI-based tools are proprietary, functionally ChatGPT operates through a dialogue interface and responds to user prompts in a human-like manner. Less than a year following its release, ChatGPT progressed from accepting text-only prompts to prompts that can include both images and text, and other LLMs have been released including Google's Bard.

Shortly after the release of ChatGPT, it demonstrated the ability to pass well-known standardized multiple-choice assessments including the United States Medical Licensing Examination (USMLE) written by American medical students [6,7]. A large volume of research has since studied LLM performance on other examinations in specialized areas of medicine including but not limited to clinical informatics [8], ophthalmology [9], plastic surgery [10], urology [11], family medicine [12], obstetrics and gynecology [13], and a situational judgment, professionalism, and ethics exam [14]. It has also been studied for

performance on medical training examinations in languages including Chinese [15,16], German [17], Dutch [12], Hebrew [13], and Japanese [18], among others [19]. These early studies find mixed performance of ChatGPT across different subspecialist fields, problem difficulty levels, or non-English-speaking contexts, but collectively acknowledge the potential for LLMs to both augment and threaten traditional medical training at the global level.

While multiple-choice examination remains an important assessment method for standardized assessment in medical certification, and comprises the bulk of existing research for LLM performance, undergraduate medical training often includes short-answer response problems for formative assessment of student progress [20,21]. While short-answer assessment can evaluate skills beyond identifying correct answers from a list of options, such as communication, reasoning, and analysis, they are more resource-intensive to implement. Clinician educators in medical training are often balancing clinical service, teaching, and administrative responsibilities. The use of limited teaching hours on grading short-answer assessments provides useful feedback to students at the expense of faculty having less time to engage with students directly, particularly for students who may be struggling and require additional support. Moreover, delays in grading may result in feedback becoming less useful for learning purposes, and there is inherent inter-grader subjectivity in producing student feedback.

Using LLMs to evaluate student answers offers an attractive solution to facilitate timely feedback to students on academic progress and to allow educators to spend more time with students. To our knowledge, while there is a small but growing body of research that addresses ChatGPT performance on answering short-answer exam questions [22–24] and its ability to generate exam problems [15,25,26], there are presently no published studies that evaluate the ability of ChatGPT to score medical student answers to short-answer exam questions. The primary objective of this study is to fill this gap in the literature by examining the efficacy of ChatGPT in evaluating student assessments. Specifically, our research question is as follows: "How effective is ChatGPT in evaluating student responses to short-answer, formative assessment questions in an undergraduate medical program, compared to traditional human assessment methods?" To address this question, our study hypothesizes that ChatGPT will correlate strongly with human assessors in grading short-answer questions. Additionally, we posit that the grading effectiveness of ChatGPT will be positively influenced by the presence of supplemental information for grading, including standards and rubrics that are typically distributed to human assessors. These hypotheses will guide the analysis and provide a framework for evaluating the role of AI in educational assessment.

2. Materials and Methods

2.1. Context

Most undergraduate medical training programs begin with a period of foundational study ("pre-clerkship") before beginning clinical clerkship rotations in the latter period of training. This study focuses on the pre-clerkship stage of training. In our undergraduate medical program, pre-clerkship is divided into five Medical Foundation units, and students complete short-answer assessments known as Concept Application Exercises (CAEs) on an approximately monthly basis [27]. Each monthly CAE includes between three and five clinical vignettes based on the most recently covered curriculum material. An example of a vignette is provided in Table 1. CAEs are graded by faculty members who serve as tutors and problem-based learning group facilitators for seven to ten medical students at a time.

Table 1. Sample CAE problem vignette and grading rubric.

Vignette	You are seeing Henry Baker and Ali Khan in follow-up in your office. You review your notes from the previous visit. Mr. Baker is a 57-year-old man with chronic musculoskeletal pain, the causes of which are multifactorial. He has used a number of different medications in the past including both long-acting Oxycodone and short-acting Oxycodone for flare-ups. Ali is an 8-year-old boy with cerebral palsy who experiences pain with vomiting. Vomiting usually follows straining with passing a bowel motion, one to two times weekly. Ali recently began treatment for gastroesophageal reflux disease. Henry Baker and Ali Khan both experience pain, however they are experiencing different types of pain with different characteristics. Provide a brief description of the relevant nociceptive pathways that contribute to each patient's pain.
Rubric	Henry Baker: Chronic somatic pain arising from the musculoskeletal system with hyperalgesia; multiple contributing factors. Nociception via somatic A-delta and C-fibers, enters the CNS via the dorsal horn and synapses at second order neurons in Lamina II (Substantia Gelatinosa). There is enhanced transmission of nociceptive impulses due to nociplastic changes within the CNS arising from enhanced NMDA receptor activation, neurogenic inflammation, loss of segmental and supraspinal inhibitory control and probably opioid-induced hyperalgesia. Pain is then transmitted supraspinally via the spinothalamic tracts and spinoreticular tracts to the thalamus for both tracts and also to the parabrachial nuclei and amygdala for the spinoreticular tract. Ultimately projection onto the sensory cortex results in perception of pain, however this will be exaggerated and provoke significant additional distress in the case of this patient. Ali Khan: Acute bouts of visceral pain arising from distention or spasm of elements of the GI tract. This may involve the esophagus or stomach as GERD is present and gastric distension may be a factor in vomiting. Potentially the large bowel or rectum may be involved since constipation is suggested on history. Nociception is via thinly myelinated A-delta and C-fibers, but these travel with autonomic nerves towards the CNS. In the case of the esophagus and stomach this is shared with the vagus nerve (parasympathetic) and sympathetic fibers to the celiac plexus and then to the sympathetic chain with segmental input into the spinal cord from about T5-12. In the case of the large bowel this too can be via sympathetic pathways; Inferior mesenteric ganglion via sympathetic chain to L1-3 levels, also some towards celiac plexus via superior mesenteric ganglion. The rectum is innervated by the sacral parasympathetic fibers to S2-4. It is noteworthy that the sparse innervation by nociceptors on the viscera, and their inputs diverging widely into the CNS result in poorly localized pain and contribute to the phenomenon of referred pain. Novice: The student will be able to describe some of the major differences between somatic and visceral nociception, specifically the course of visceral nociception along autonomic fibers, and the poor localization of visceral pain. They will be able to describe some of the key elements of the nociceptive pathways. They will be able to identify that chronic pain and opioid use may result in hyperalgesia. Proficient: In addition to the components of a novice response, the student will be able to provide further details of the afferent pathways pertaining to the different portions of the GI tract, differentiating between sympathetic and parasympathetic input, and naming a reasonable portion of the nerves and structures involved, and correlate them to Ali Khan's clinical presentation. The students will be able to describe how A-delta and C-fibers synapse within the spinal cord, name the ascending spinal pathways, and their connection to the thalamus. There will be some understanding and description of nociplastic changes in the pain pathways resulting in enhanced nociception in the case of Henry Baker. Accomplished: Further to a proficient response, there will be a more detailed and complete account of somatic and visceral nociception and the processes involved in hyperalgesia, referred pain, and evidence of clinical correlation with the two cases. Though not expected of the students, reference to descending inhibition via the periaqueductal gray, and more detailed reference to structures in the brain making up the "Pain Matrix" beyond the thalamus can suggest a more accomplished response if some details are lacking otherwise.

To evaluate CAE answers, faculty tutors are given both a rubric and a standard. The rubrics are written for each specific CAE question outlining what level of detail is expected of students to attain each of the five score levels, with an example shown in Table 1. The standard is the same for all problems and is shown in Table 2. Evaluations are completed using a computer system that conceals student identity while grading. CAE scoring follows a 5-point scale, with 5 being the highest score achievable. A score of 3 is considered the minimum passing standard.

Table 2. Grading standard for CAE problems. "MF" refers to "Medical Foundation".

Score	Proficiency Level	Score Description
5	Accomplished	Student was able to describe a deep and complete understanding of the concepts/mechanisms and was able to explain how new information or concepts related to topics discussed in previous subunits or foundations, or encountered in other areas of the program. Mastered the learning objectives of the MF.
4		Student demonstrated a comprehensive understanding of the concepts/mechanisms central to the MF. Demonstrated excellent organization and integration of material. Demonstrated superior achievement of the learning objectives in the MF.
3	Proficient (passing score)	Student was able to describe the key concepts/mechanisms to a degree sufficient for the MF. Demonstrated an understanding of the importance or relevance of the concepts/mechanisms. Information was appropriately organized and prioritized. Demonstrated acceptable achievement of the learning objectives of the MF. Most students' responses on CAE questions are expected to be consistent with this level of achievement.
2	Novice	Student was able to describe most but not all of the key concepts/mechanisms. Understanding of some of the material was incomplete. Student was in the early stages of achieving the learning objectives in the MF.
1		Student was able to describe some but not all of the concepts/mechanisms. Seemed unclear/uncomfortable with at least some of the material. Understanding of some concepts was superficial. Difficulty organizing and prioritizing information. Achievement of the learning objectives of the MF was not yet adequate.

2.2. Data Collection

Ten CAE questions were selected from the bank of past CAE questions at our institution, ensuring representation across the entire pre-clerkship curriculum. For each of the ten questions, six past student responses were selected, representing two student responses at the novice level (score 1 out of 5), two at the proficient level (score 3 out of 5), and two at the accomplished level (score 5 out of 5). In total, this formed a dataset of sixty past student answers, with twenty answers at each of levels 1, 3, and 5. These past student answers contained no personally identifiable details. It is important to note that each of these sixty past student answers were marked by different tutors from our historic tutor roster.

ChatGPT was used to evaluate each of the sixty responses under four different grading conditions. Under the first condition, the problem vignette and past student answer were submitted to ChatGPT along with a prompt asking for a score from 1 to 5. The second condition also included the generic standard for grading shown in Table 2 in the ChatGPT prompt. The third condition omitted the generic standard, but included a grading rubric specific to the CAE question. The fourth and final condition included both the generic standard and the specific grading rubric in the prompt. Testing these four different conditions allows for some exploration into how the grading accuracy of ChatGPT is influenced by all combinations of available information currently given to our human tutors. Each of the four conditions included a statement that instructed ChatGPT to use a score of 3 as an anchor as this is the score achieved by the majority of medical students. Each submission produced a ChatGPT-assigned score from 1 to 5 to each student response as part of its textual output, resulting in 24 data items per CAE problem vignette. Ten vignettes were used in the study leading to a total of 240 ChatGPT-generated scores for analysis. A summary of the procedure for creating ChatGPT prompts is visualized below in Figure 1.

The study employed the most up-to-date version of ChatGPT at the time of the experiment (ChatGPT-4). For each CAE question, the set of six student responses was submitted to ChatGPT in July of 2023. The order of the six responses was randomized on each submission. A new session in ChatGPT was started for each of the four conditions to avoid any influence of previous evaluations on the scoring of the subsequent conditions.

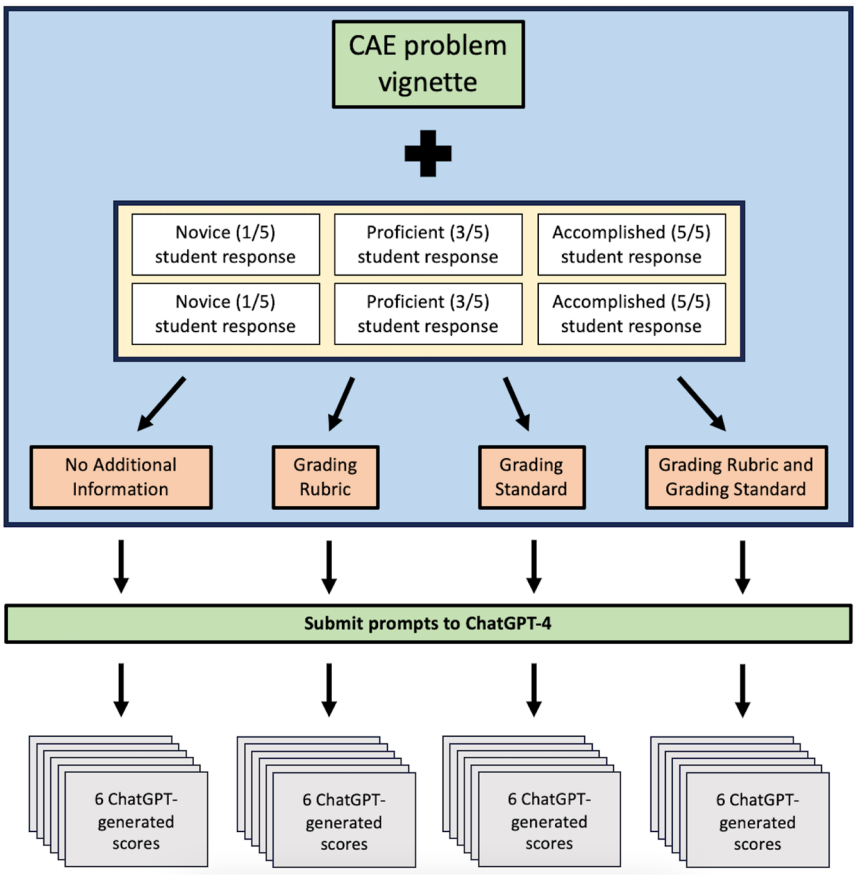

Figure 1. ChatGPT prompt design for each CAE question.

2.3. Analysis

The scores that were originally assigned to the pool of sixty student responses by human tutors were compared to the scores assigned by ChatGPT under each of the four described assessment conditions. That is, sixty human-assigned scores were compared to a total of two hundred and forty ChatGPT-assigned scores.

We used basic descriptive statistics to analyze and visualize the distribution of ChatGPT-generated responses under each condition and to examine trends in ChatGPT-assigned scores based on the original human-assigned score for each student response. Spearman's correlation coefficient was calculated to quantify the strength and direction of the relationship between ChatGPT-assigned scores and human-assigned scores under each of the four assessment conditions. Additionally, partial η^2 was calculated to measure effect size where appropriate.

We sought to understand the frequency and size of changes between human-assigned and ChatGPT-assigned scores. Descriptive statistics were used to summarize answer-by-answer changes in score. We considered the absolute difference in score as well as whether the change in score would have resulted in a change in scoring category, that is, a change between the categories Novice, Proficient, and Accomplished as outlined in Table 2. Further statistical comparisons were made using the repeated measures analysis of variance (ANOVA) to evaluate the consistency and reliability of ChatGPT's scoring relative to an expert human assessor [28], including calculations of partial η^2.

All statistical analysis was performed using SPSS (v26, IBM, Redmond, WA, USA).

3. Results

A summary of the ChatGPT-assigned scores across the pool of sixty student CAE answers is visualized in Figure 2. Conditions where the problem-specific grading rubrics are included in the ChatGPT prompt (the third and fourth conditions) do not have a right-sided skew, showing that ChatGPT-assigned scores tend to be lower when a rubric is not provided.

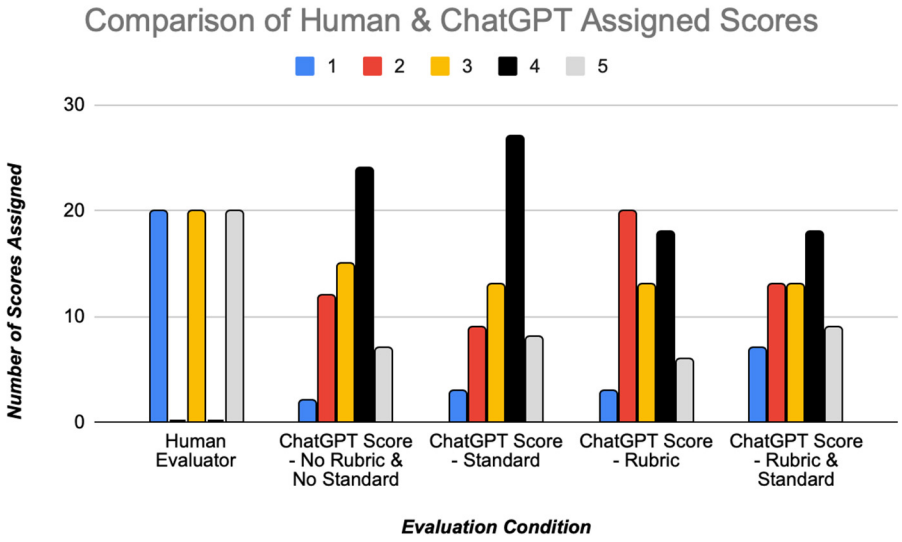

Figure 2. Score comparison between human assessors and ChatGPT.

Mean scores for ChatGPT-scored responses are summarized in quantitative terms in Table 3 for each assessment condition including standard error values, with comparison to the human-assigned scores. The average and median score of selected human-generated responses was 3.0 in keeping with the study design. The difference in mean scores between human assessors and ChatGPT was statistically significant in the two conditions where rubrics were not included in the prompt to ChatGPT, with ChatGPT giving higher average scores in these grading conditions.

Table 3. Mean scores of human-assigned scores as well as ChatGPT-assigned scores under each assessment condition, with standard error, 95% confidence intervals, and comparisons to the human evaluator for each condition. The differences between the mean scores of human assessors and ChatGPT were only statistically significant for the two assessment criteria where question-specific rubrics were not provided to the ChatGPT prompt.

	Evaluation Condition	Mean	Std. Error	95% CI	Mean Difference	Std. Error	*p*-Value
Human-Assigned Score	N/A	3.00	0.21	(2.58, 3.43)	N/A	N/A	N/A
ChatGPT-Assigned Score	No Rubric and No Standard	3.37	0.13	(3.10, 3.64)	−0.37	0.15	0.015
	Standard Only	3.47	0.48	(3.19, 3.74)	−0.47	0.17	0.008
	Rubric Only	3.07	0.14	(2.78, 3.36)	−0.07	0.16	0.67
	Rubric and Standard	3.15	0.16	(2.83, 3.48)	−0.15	0.17	0.37

To evaluate the ChatGPT-assigned scores at a more granular level, we summarize in Table 4 the distribution of ChatGPT-assigned scores based on the corresponding human-assigned score to each problem. We also measured Spearman correlation coefficients for each of the four assessment conditions between human-assigned and ChatGPT-assigned scores. Good correlation with statistical significance was observed in each of the four conditions (0.599–0.732). There was no statistically significant difference in the strength of the correlations across the four conditions.

Table 4. Distribution of ChatGPT-assigned scores under each of the four assessment conditions, grouped by original human assessor score.

Average ChatGPT-Assigned Scores		No Rubric and No Standard		Standard Only		Rubric Only		Rubric and Standard	
		Mean	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**
Tutor Rating	1 (Novice)	2.30	0.66	2.55	0.94	2.00	0.65	2.00	0.86
	3 (Proficient)	3.65	0.81	3.75	0.85	3.35	1.04	3.50	1.10
	5 (Accomplished)	4.15	0.59	4.10	0.72	3.85	0.67	3.95	0.89
Correlation coefficient (with 95% CI)		0.732 (0.587, 0.831) $p < 0.001$		0.599 (0.407, 0.740) $p < 0.001$		0.681 (0.517, 0.797) $p < 0.001$		0.637 (0.457, 0.767) $p < 0.001$	

We subsequently compared, under each assessment condition, human-assigned and ChatGPT-assigned scores based on the size of the difference between the two assigned scores. While some difference between the human-assigned and ChatGPT-assigned scores was very common, and in fact occurred in a majority of cases (65–80%), score changes by more than one point out of five were more rare (20–38%). Score changes constituting a change in marking category between Novice, Proficient, and Accomplished occurred more frequently than changes by more than two points, with more than half of marking categories differing between ChatGPT and human graders in the grading condition with only a generic standard (34–57%). This is an expected finding given that a change in grading condition may occur with a change of only one point depending on the original assigned score. These findings are summarized in Table 5. Multivariate testing produced a partial η^2 value of 0.332, meaning that 33.2% of the variance between the four assessment conditions is related to the conditions under which ChatGPT is asked to grade the responses.

Table 5. Differences between human-assigned and ChatGPT-assigned scores. These values represent percentages of the sixty ChatGPT-assigned answers for each condition. A positive scoring difference indicates that ChatGPT assigned a greater score than the human tutor, and vice versa. Marking categories include Novice (scores of 1 or 2), Proficient (3), or Accomplished (4 or 5). For changes by two or more points, sum of individual score change frequencies may not match exact totals due to rounding.

Scoring Difference	No Rubric and No Standard	Standard Only	Rubric Only	Rubric and Standard
−4	0	0	0	0
−3	0	7	10	2
−2	3	22	27	10
−1	25	17	20	18
0	20	33	35	27
+1	35	17	7	30
+2	17	5	2	12
+3	0	0	0	2
+4	0	0	0	0
Change by 1 or more points (%)	80	67	65	73
Change by 2 or more points (%)	20	33	38	25
Change by 1 or more marking categories (%)	42	57	34	45

Although not explicitly prompted for feedback, ChatGPT autonomously justified each score given to the medical student assessments with brief, elucidative comments. There are themes and recurring phrases observed in these outputs. ChatGPT consistently underscored the "depth of conceptual understanding", illustrated by frequent mentions of students' ability or inability to demonstrate a "nuanced understanding" and "in-depth grasp of medical concepts". Emphasis on "relevance and application to clinical cases" was

also prevalent. The theme of "clarity and logical organization" emerged, with ChatGPT frequently mentioning whether answers were "clear", "logically sequenced", or "coherent". The "completeness of the response" appeared to be an important consideration, marked by comments like "covers essential aspects". These findings are summarized in Table 6.

Table 6. Thematic analysis of ChatGPT grading justifications.

Evaluation Condition	Concept	Illustrative Quotes from ChatGPT Comments
All Conditions	Depth of understanding	"This student thoroughly considers both central and peripheral causes of vertigo" "This response provides an in-depth explanation of both the thyroid uptake and thyroid scan but contains a few errors and confusions in the explanations" "However, the explanation lacks depth in describing the clinical relevance in Maria's case".
	Accuracy	"The response correctly acknowledges that genetics can play a role in cancer, but the assertion that 'His cancer risk is high because 3 of his first relatives were diagnosed with cancer' might be an oversimplification". "Accuracy: Correct information regarding the ascending and descending pathways and opioid effects". "The mention of BRCA genes in relation to Mr. Steel's family history is accurate, but the connection between BRCA and melanoma is not clear".
	Completeness	"Overall, it's a strong response with a slight lack of completeness in the second part". "Completeness: The student emphasizes the importance of UV exposure and the need for shared decision-making".
No Rubric and No Standard	Organization	"The response may be slightly less organized, and there's some ambiguity in Ali's description, with the student noting a lack of case information". "The answer might benefit from a more structured and clear presentation of the physiological mechanisms involved".
Standard Only	Thoroughness of knowledge	"The response is thorough and emphasizes the biological and social factors contributing to the pain experienced by Henry Baker and Ali Khan".
Rubric Only	Meeting of criteria	"The student correctly identifies Angiotensin II's role in the RAAS system and its impact on fluid retention. However, the explanation regarding its effect on the afferent and efferent arterioles is not as detailed as in the guide, and the student doesn't mention ACE inhibitors' impact on GFR or provide the expected side effect". "This answer is thorough and matches much of the evaluation guide"
Rubric and Standard	Combination of success on both evaluation frameworks	"Demonstrates an understanding of the potential benefits of genetic testing for family members and not just the patient. Recognizes the importance of genetics but also acknowledges the multifactorial causes of cancer. Provides a balanced perspective of risks and benefits. Rating: 3 (Proficient)—This student sufficiently describes the key concepts and prioritizes the importance of genetics and other factors but doesn't dive deep into specifics or the broader understanding of cancer genetics".

4. Discussion

The primary objective of this study was to evaluate the effectiveness of ChatGPT in grading short-answer, formative assessments within an undergraduate medical program. Our findings demonstrate that ChatGPT has potential to perform well as an assisting tool for grading. Furthermore, when benchmarked against expert human assessors, ChatGPT shows favorable Spearman correlations. This has the potential to have a significant impact

on medical education as administrators continue to search for increasingly reliable, timely, and cost-effective grading solutions.

It was instructive to understand that, while the scores assigned by ChatGPT exhibit significant differences between grading conditions, the strength of correlation between ChatGPT-assigned and human-assigned scores was not significantly different between groups. The inclusion of a rubric yielded lower ChatGPT-assigned scores compared to grading conditions where ChatGPT was not given a rubric, but these lower scores did not significantly differ from the scores assigned by human tutors. There was no obvious benefit from our analysis to include the generic grading standard to the ChatGPT prompts, suggesting that the standard does not add substantial value or information in terms of ChatGPT being able to reason and assign scores between 1 and 5 without this information. However, the inclusion of the rubric may potentially enhance the accuracy and meaningfulness of the feedback given, but this assumption warrants further investigation. Moreover, it is possible that these differences between results across grading conditions may be attributed to our current rubrics and standard either being insufficiently clear, which would suggest a need for further refinement of these materials, or the fact that the rubric and standards were designed for a human assessor rather than designed for the purpose of being inserted into a ChatGPT prompt.

Most responses (65–80%) did show a change in score by at least one point between ChatGPT-assigned and human-assigned scores, but score adjustments by more than one point were less frequent (20–38%), indicating that while ChatGPT may not always match the specific score given by a human grader, it generally falls within an acceptable range to avoid changing the overall scoring category. In general, ChatGPT tended to assign higher scores than human assessors. In the event that such a tool was implemented for grading assistance, this may falsely increase scores assigned to lower quality responses. However, given that each CAE is comprised of multiple vignettes, and that ChatGPT assigning lower scores than human tutors was still quite common (as shown in Table 5), we would not necessarily expect the final results for a given student on a CAE to change. As a result, we are not concerned based on the results of this study that using an LLM to assist in student evaluation would have negative implications in being able to identify students who may require additional support, although this represents a worthwhile future area of investigation. The frequency of changes between scoring categories (34–57%) suggests that relying solely on AI-based grading could sometimes overlook nuances that would be critical in a medical educational context. However, while not formally tested in this study, there also exists some level of inter-rater variability with independent human tutors; our group formally investigated in a previous work and found a Cronbach alpha value of 0.816 for a team of six human assessors on past student-generated CAE responses [29].

Our study is a continuation of prior work in understanding how LLMs such as ChatGPT can augment the medical student learning experience [29]. In particular, it has been suggested and observed that these tools can augment the self-directed and case-based student learning process [30–32], and we are interested in studying this for our particular institution in the pre-clerkship context. Integrating the use of AI technology in any context requires comprehensive ethical considerations and transparency. There is an ongoing need for further study into best practices for ethical uses of such tools and concerns for training medical learners to recognize medical misinformation or AI-generated hallucinations [33,34]. For example, using an LLM to evaluate students in summative settings would appropriately be subject to scrutiny, as these results would then ultimately impact career progression for students, such as application to residency or graduate training programs. At our institution, CAEs are used only for formative evaluation as learning aids, and results do not impact final evaluations that appear on transcripts. Nevertheless, difficulty in the CAE may prompt the initiation of a targeted learning enhancement plan to provide additional support to a struggling student. As a result, although there is no threat that an erroneous ChatGPT-assigned score would influence overall medical student

progression at our institution, it may impact the allocation of supplemental education resources.

This work serves as a stepping stone for future research in AI-based assessment techniques and also raises further questions. The variations in ChatGPT's performance under different conditions, such as the inclusion or omission of grading rubrics or standards, suggests that there may be other input models that may align more closely with human-assigned scores. A combination workflow of an LLM and a human tutor, rather than selecting only one option, may also yield an optimal outcome for student assessment; for example, a tutor could review ChatGPT-graded student answers for significant recategorizations before the results are returned to students, or tutors could only mark certain students either based on an opt-in basis or for students who have previously been identified to be missing certain learning objectives. Similarly, multiple LLM outputs could be pooled into one combined response to test the accuracy of a multi-rater method.

Tutors typically provide narrative feedback on CAE responses to the student, but this study did not focus on the content of ChatGPT responses beyond their numerical score from 1 to 5; it remains to be considered how the loss of this comprehensive feedback from tutors would be received by students when balanced against the fact that the scores would become available much closer to the time of the initial assessment. Providing feedback that is more prompt is a recognized area of opportunity for AI in medical education [35]. Since tutors both interact with students and evaluate CAEs, automating this process may give them more time to interact directly with students and to explore any learning deficits independent of a student's test-taking ability. However, it is important to note that taking this responsibility may be difficult for tutors in a group setting regardless of any saved time. It is for this reason that our institution is currently considering the use of a "learning director" to isolate test-taking ability versus the application of knowledge-related issues.

This proof-of-concept study, while insightful, has several limitations. First, its findings are limited to a relatively small number of short-answer formative assessment questions specific to a Canadian undergraduate medical program, which may not be applicable to other educational settings. Second, results are dependent on the specific version of ChatGPT and our chosen prompting strategy, suggesting that any changes to these factors could alter the outcomes. Finally, the study does not address the potential grade inflation by ChatGPT or the quality of its narrative feedback, both of which are crucial for effective student evaluation and learning.

Planning is underway for a series of studies aimed at assessing the impact of ChatGPT on team problem-based learning behaviors. This future research will also explore the ability of generative AI to provide effective feedback on student responses and to develop new question stems and rubrics specifically for medical student assessment. Future studies should explore the potential application of AI grading systems in various disciplines within the health sciences. This includes examining the adaptability of AI grading across different educational contexts, integrating AI with human assessment to achieve balanced grading, and evaluating the impact of AI grading on student learning outcomes and feedback quality. Future research should also explore the use of different prompting strategies, as well as different AI models.

5. Conclusions

Our study demonstrates that ChatGPT is a viable, though imperfect, alternative to expert human graders of short-answer formative assessments in a medical education program. However, its performance raises concerns of possible grade inflation and, without the use of supplementary grading criteria, may miss nuanced details resulting in scoring a student's response. In terms of incorporating such a system into a workflow, using ChatGPT as an assistant rather than an alternative in an LLM–human hybrid solution is an intuitive option to address the limitations of an LLM-only grading system. These hybrid solutions have yet to be formally explored, as well as potential opportunities for time and

cost savings and scalability. Our findings open avenues for future research aimed at refining AI-based grading systems through multi-rater comparisons and algorithmic improvements.

Author Contributions: Conceptualization, L.M., M.L. and M.S.; methodology, L.M., L.B., K.B., A.J.L., Q.N.N., M.L. and M.S.; data curation, M.L.; formal analysis, M.S.; investigation, L.M. and M.L.; writing—original draft preparation, L.M. and L.B.; writing—review and editing, L.M., L.B., K.B., A.J.L., Q.N.N., M.L. and M.S.; visualization, L.M. and M.S.; supervision, K.B., A.J.L., Q.N.N. and M.S.; project administration, L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data due to privacy concerns. Requests for data will need to be approved by the Michael G. DeGroote School of Medicine UGME office.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wartman, S.; Combs, C.D. Reimagining Medical Education in the Age of AI. *AMA J. Ethics* **2019**, *21*, E146–E152. [CrossRef] [PubMed]
2. Masters, K. Artificial intelligence in medical education. *Med. Teach.* **2019**, *41*, 976–980. [CrossRef] [PubMed]
3. Khan, R.A.; Jawaid, M.; Khan, A.R.; Sajjad, M. ChatGPT—Reshaping medical education and clinical management. *Pak. J. Med. Sci.* **2023**, *39*, 605. [CrossRef] [PubMed]
4. Lee, H. The rise of ChatGPT: Exploring its potential in medical education. *Anat. Sci. Educ.* **2023**, ase.2270. [CrossRef] [PubMed]
5. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef]
6. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. 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. [CrossRef]
7. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. [CrossRef]
8. Kumah-Crystal, Y.; Mankowitz, S.; Embi, P.; Lehmann, C.U. ChatGPT and the clinical informatics board examination: The end of unproctored maintenance of certification? *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1558–1560. [CrossRef]
9. Antaki, F.; Touma, S.; Milad, D.; El-Khoury, J.; Duval, R. Evaluating the Performance of ChatGPT in Ophthalmology. *Ophthalmol. Sci.* **2023**, *3*, 100324. [CrossRef]
10. Humar, P.; Asaad, M.; Bengur, F.B.; Nguyen, V. ChatGPT Is Equivalent to First-Year Plastic Surgery Residents: Evaluation of ChatGPT on the Plastic Surgery In-Service Examination. *Aesthet. Surg. J.* **2023**, *43*, NP1085–NP1089. [CrossRef]
11. Huynh, L.M.; Bonebrake, B.T.; Schultis, K.; Quach, A.; Deibert, C.M. New Artificial Intelligence ChatGPT Performs Poorly on the 2022 Self-assessment Study Program for Urology. *Urol. Pract.* **2023**, *10*, 409–415. [CrossRef]
12. Morreel, S.; Mathysen, D.; Verhoeven, V. Aye, AI! ChatGPT passes multiple-choice family medicine exam. *Med. Teach.* **2023**, *45*, 665–666. [CrossRef] [PubMed]
13. 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*, 1797–1802. [CrossRef] [PubMed]
14. Borchert, R.J.; Hickman, C.R.; Pepys, J.; Sadler, T.J. Performance of ChatGPT on the Situational Judgement Test—A Professional Dilemmas–Based Examination for Doctors in the United Kingdom. *JMIR Med. Educ.* **2023**, *9*, e48978. [CrossRef] [PubMed]
15. Cheung, B.H.H.; Lau, G.K.K.; Wong, G.T.C.; Lee, E.Y.P.; Kulkarni, D.; Seow, C.S.; Wong, R.; Co, M.T.H. 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. [CrossRef]
16. Kao, Y.S.; Chuang, W.K.; Yang, J. Use of ChatGPT on Taiwan's Examination for Medical Doctors. *Ann. Biomed. Eng.* **2023**. [CrossRef] [PubMed]
17. Friederichs, H.; Friederichs, W.J.; März, M. ChatGPT in medical school: How successful is AI in progress testing? *Med. Educ. Online* **2023**, *28*, 2220920. [CrossRef]
18. Takagi, S.; Watari, T.; Erabi, A.; Sakaguchi, K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Med. Educ.* **2023**, *9*, e48002. [CrossRef]

19. Alfertshofer, M.; Hoch, C.C.; Funk, P.F.; Hollmann, K.; Wollenberg, B.; Knoedler, S.; Knoedler, L. Sailing the Seven Seas: A Multinational Comparison of ChatGPT's Performance on Medical Licensing Examinations. *Ann. Biomed. Eng.* **2023**. [CrossRef]
20. Bird, J.B.; Olvet, D.M.; Willey, J.M.; Brenner, J. Patients don't come with multiple choice options: Essay-based assessment in UME. *Med. Educ. Online* **2019**, *24*, 1649959. [CrossRef]
21. Tabish, S.A. Assessment methods in medical education. *Int. J. Health Sci.* **2008**, *2*, 3–7.
22. Sinha, R.K.; Deb Roy, A.; Kumar, N.; Mondal, H. Applicability of ChatGPT in Assisting to Solve Higher Order Problems in Pathology. *Cureus* **2023**, *15*, e35237. [CrossRef] [PubMed]
23. Das, D.; Kumar, N.; Longjam, L.A.; Sinha, R.; Roy, A.D.; Mondal, H.; Gupta, P. Assessing the Capability of ChatGPT in Answering First- and Second-Order Knowledge Questions on Microbiology as per Competency-Based Medical Education Curriculum. *Cureus* **2023**, *15*, e36034. [CrossRef] [PubMed]
24. Ghosh, A.; Bir, A. Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry. *Cureus* **2023**, *15*, e37023. [CrossRef] [PubMed]
25. 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. [CrossRef] [PubMed]
26. Ayub, I.; Hamann, D.; Hamann, C.R.; Davis, M.J. 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. [CrossRef]
27. Neville, A.J.; Cunnington, J.; Norman, G.R. Development of clinical reasoning exercises in a problem-based curriculum. *Acad. Med.* **1996**, *71*, S105–S107. [CrossRef]
28. Norman, G. Likert scales, levels of measurement and the "laws" of statistics. *Adv. Health Sci. Educ.* **2010**, *15*, 625–632. [CrossRef]
29. Morjaria, L.; Burns, L.; Bracken, K.; Ngo, Q.N.; Lee, M.; Levinson, A.J.; Smith, J.; Thompson, P.; Sibbald, M. Examining the Threat of ChatGPT to the Validity of Short Answer Assessments in an Undergraduate Medical Program. *J. Med. Educ. Curric. Dev.* **2023**, *10*, 23821205231204178. [CrossRef]
30. Xie, Y.; Seth, I.; Hunter-Smith, D.J.; Rozen, W.M.; Seifman, M.A. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: A comprehensive analysis. *ANZ J. Surg.* **2023**, ans.18666. [CrossRef]
31. 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. [CrossRef] [PubMed]
32. Karabacak, M.; Ozkara, B.B.; Margetis, K.; Wintermark, M.; Bisdas, S. The Advent of Generative Language Models in Medical Education. *JMIR Med. Educ.* **2023**, *9*, e48163. [CrossRef] [PubMed]
33. Fischetti, C.; Bhatter, P.; Frisch, E.; Sidhu, A.; Helmy, M.; Lungren, M.; Duhaime, E. The Evolving Importance of Artificial Intelligence and Radiology in Medical Trainee Education. *Acad. Radiol.* **2022**, *29*, S70–S75. [CrossRef] [PubMed]
34. Winkler-Schwartz, A.; Bissonnette, V.; Mirchi, N.; Ponnudurai, N.; Yilmaz, R.; Ledwos, N.; Siyar, S.; Azarnoush, H.; Karlik, B.; Del Maestro, R.F. Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *J. Surg. Educ.* **2019**, *76*, 1681–1690. [CrossRef]
35. Abd-Alrazaq, A.; AlSaad, R.; Alhuwail, D.; Ahmed, A.; Healy, P.M.; Latifi, S.; Aziz, S.; Damseh, R.; Alrazak, S.A.; Sheikh, J. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med. Educ.* **2023**, *9*, e48291. [CrossRef]