2025, Volume 57, Issue 10, 727-731, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Evaluating the Effectiveness of ChatGPT Versus Human Proctors in Grading Medical Students' Post-OSCE Notes

Kirstyn Thomas, MD[a]; Laura Szalacha, EdD[b]; Karim Hanna, MD[c]; James Anibal[d]; John Petrilli, MD[c]

AUTHOR AFFILIATIONS:

[a] Family Medicine, University of South Florida/BayCare Health System, Tampa, FL, United States

[b] Morsani College of Medicine, University of South Florida, Tampa, FL

[c] Department of Family Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL

[d] Computational Health Informatics Lab, Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Oxford, UK

CORRESPONDING AUTHOR:

Kirstyn Thomas, Family Medicine, University of South Florida/BayCare Health System, Tampa, FL, United States, kirstynthomasmd@gmail.com

HOW TO CITE: Thomas K, Szalacha L, Hanna K, et al. Evaluating the Effectiveness of ChatGPT Versus Human Proctors in Grading Medical Students' Post-OSCE Notes. *Fam Med.* 2025;57(10):727-731. doi: 10.22454/FamMed.2025.954255

FIRST PUBLISHED: November 20, 2025

ABSTRACT

Background and Objectives: Artificial intelligence (AI) tools have potential utility in multiple domains, including medical education. However, educators have yet to evaluate AI's assessment of medical students' clinical reasoning as evidenced in note-writing. This study compares ChatGPT with a human proctor's grading of medical students' notes.

Methods: A total of 127 subjective, objective, assessment, and plan notes, derived from an objective structured clinical examination, were previously graded by a physician proctor across four categories: history, physical exam, differential diagnosis/thought process, and treatment plan. ChatGPT-4, using the same rubric, was tasked with evaluating these 127 notes. We compared AI-generated scores with proctors' scores using *t* tests and χ^2 analysis.

Results: The grades assigned by ChatGPT were significantly different than those assigned by proctors in history ($P<.001$), differential diagnosis/thought process ($P<.001$), and treatment plan ($P<.001$). Cohen's *d* was the largest for treatment plan at 1.25. The differences led to a significant difference in students' mean cumulative grade (proctor 23.13 [SD=2.84], ChatGPT 24.11 [SD 1.27], $P<.001$), affecting final grade distribution ($P<.001$). With proctor grading, 81 of the 127 (63.8%) notes were honors and 46 of the 127 (36.2%) were pass. ChatGPT gave significantly more honors (118/127 [92.9%]) than pass (9/127 [7.1%]).

Conclusions: When compared to a human proctor, ChatGPT-4 assigned statistically different grades to students' SOAP notes, although the practical difference was small. The most substantial grading discrepancy occurred in the treatment plan. Despite the slight numerical difference, ChatGPT assigned significantly more honors grades. Medical educators should therefore investigate a large language model's performance characteristics in their local grading framework before using AI to augment grading of summative, written assessments.

INTRODUCTION

Large language models (LLMs), a subset of artificial intelligence (AI), are advanced algorithms trained on vast amounts of unstructured text data. With this knowledge, LLMs like the generative pretrained transformer (GPT) can generate human-like language and complete complex tasks.

LLMs, and AI in general, continue to gain new uses in many fields. AI is being studied for its potential in a broad number of applications, though consensus is lacking on how AI might best be used in medical education.[1] This lack of consensus is particularly true for how medical educators can use AI to augment their educational duties and curricular development.

Past studies have investigated AI's ability to assess medical knowledge and respond to medical questions. Ilgaz and Zahra found that ChatGPT (OpenAI LP) and Google Bard (Alphabet Inc) were able to generate anatomy-related multiple-choice questions with a high degree of

accuracy, despite one question having an incorrect answer.[2] Moreover, LLMs may be able to augment problem-based learning by acting as a "virtual patient,"[3] though this area is still being explored. Multiple studies also have examined the performance of AI in medical education examinations with good results. On the United States Medical Licensing Exam, LLMs performed at or near the passing threshold of 60%.[4] On the family medicine in-training exam, ChatGPT-4 passed with 86.5% accuracy.[5] GPT answered 46% of ophthalmology board preparation practice questions correctly[6] and scored 60.2% on neurosurgery written boards.[7] These results indicate that LLMs, ChatGPT in particular, demonstrate at least rudamentary ability to assess medical knowledge and respond appropriately; so AI has potential as a tool for medical educators.

Grading is a potential application for LLMs, with possible benefits for both educators and students. Automated grading tools may give educators more time that can be used for curricular development, innovation, student support, and other endeavors. Students may benefit from immediate, standardized, and in-depth feedback. However, the enthusiasm to use AI more broadly must be tempered by legitimate concerns about reliability, bias, and privacy.[8] As such, medical educators are exploring its use in grading. One study assessed this capability by examining AI's potential for grading anatomy laboratory assessments, finding that the software reduced grading time by half, reduced grading bias, and improved grading consistency and transparency.[9] UT Southwestern used AI to grade medical students' objective structured clinical examination (OSCE) notes, achieving up to 89.7% agreement with human graders (Cohen's κ of 0.79)[10] and with an estimated 91% reduction in human effort.

In this study, we add to the growing body of grading literature by assessing whether ChatGPT-4 can accurately and reliably grade medical students' subject, objective, assessment, and plan (SOAP) notes from an OSCE.

METHODS

Data Collection

This study falls under the category of "not human subjects research" and received an exemption from the Institutional Review Board at University of South Florida.

One hundred and twenty-seven SOAP notes were written following third-year medical students' completion of an OSCE as part of the family medicine clerkship. A single family physician proctor previously graded the notes. Students could earn 28 possible points in four categories: history (seven points), physical exam (seven points), differential diagnosis/thought process (nine points), and treatment plan (five points). The students' notes were deidentified, and proctor scores from each of the four categories were collected.

The same grading rubric that the proctor had used previously to grade the notes was uploaded to ChatGPT. The LLM was asked to use the rubric to assign grades to each note, applying the instructions, "Use this rubric to grade the following response." These instructions were given in a "zero-shot" approach, because the LLM was provided with only the rubric and instructions, without training on sample notes or expected responses. The LLM-assigned grades for each of the four categories were then collected.

Statistical Analysis

We used independent sample *t* tests to evaluate differences in mean scores for the assessment categories history, physical exam, differential diagnosis/thought process, treatment plan, and total score. We used contingency table analysis to assess the association between the grader (proctor vs ChatGPT) and grade distribution outcomes (honors and pass).

We calculated effect sizes using Cohen's *d* to evaluate the practical significance of the differences. Statistical significance was set at $P<.05$.

RESULTS

History

The history section of the note reflected the students' documentation of the standardized patient interview. Students could be awarded up to seven points for documenting important historical elements leading to the appropriate diagnoses. The mean score assigned by proctors was 6.30 (SD=0.86), and the mean score assigned by ChatGPT was 6.66 (SD=0.48). While this difference was statistically significant ($P<.001$), the difference in mean scores was practically slight (Cohen's $d=0.52$).

Physical Exam

The physical exam section of the note reflects the students' documentation of the physical exam performed during the standardized patient encounter. Students could earn up to seven points for documenting important exam findings in this category. Students were instructed in the examination to document only exam maneuvers that they actually performed in the encounter. The mean score assigned by the proctor was 5.50 (SD=1.62), and the mean score assigned by ChatGPT was 5.65 (SD=0.69). We found no significant difference between these scores ($P=.34$).

Differential Diagnosis/Thought Process

The differential diagnosis/thought process portion of the notes is where students document potential diagnoses for the patient's complaint, supported by findings from the history and physical exam, which the students must list. Students could earn a maximum of nine points in this section. One point is given for each diagnosis. Up to six points could be earned for the thought process (two points for relevant findings for each diagnosis). The mean score assigned by the proctor was 7.45 (SD=1.29), and the mean score assigned by ChatGPT was 6.86 (SD=0.43). While this difference was statistically significant ($P<.001$), the difference was practically slight (Cohen's $d=0.61$).

Treatment Plan

The treatment plan portion of the note is where the students document the advice they would give, the medicine they would prescribe, and/or the diagnostic testing they would obtain in order to make a definitive diagnosis. In this section, students could earn a maximum of five points for listing clinically reasonable approaches to treatment, as listed in the rubric. The mean score assigned by the proctor was 3.89 (SD=1.16), and the mean score assigned by ChatGPT was 4.94 (SD=0.29). The difference between the mean scores was a statistically significant difference (P<.001), and the difference between the two is notably large (Cohen's d=1.252).

Total Score and Grade Distribution

Each note could earn at most 28 points. The mean score assigned by the proctor was 23.13 (SD=2.84), and the mean score assigned by ChatGPT was 24.11 (SD=1.27), which was a statistically significant difference (P<.001). While this was a practically small difference (Cohen's d=0.44), the ChatGPT scores would result in a statistically significant difference in the final grade. The OSCE is one of the elements used to determine a student's final grade, with the cutoff for honors being 80%. Raw grades translated to proctor M=82.62% (SD=10.15%) versus ChatGPT M=86.10% (SD=4.52%). The difference in the final grade assignment (honors vs pass) between human and AI graders led to a statistically significant difference in final grade distribution (χ^2=31.77, P<.001). The human proctor graded the notes as 81 (63.8%) honors and 46 (36.2%) pass, whereas ChatGPT graded 118 (92.9%) as honors and 9 (7.1%) as pass.

In addition, subanalysis revealed that AI-generated scores did not demonstrate a meaningful relationship—either linear or nonlinear—with human-assigned scores across the five evaluated domains. Scatterplots (Appendix A) showed weak or absent patterns, with evident clustering and ceiling effects, particularly in the proctors' scores. Restricted cubic spline modeling, which is capable of capturing flexible nonlinear trends, yielded low R^2 values ranging from 0.01 to 0.098, indicating that less than 10% of the variance in human scores was explained by AI scores.

DISCUSSION

While ChatGPT graded SOAP notes differently than did the human proctor in all categories except for physical exam, on average the difference in total scores between AI and the human grader was only one point (3.5%). These results mirror the findings from Jamieson *et al.*, who found 89.7% agreement between an LLM and human expert graders when using AI in a pass/fail grading system.[10] However, for our tiered grading system, ChatGPT's grading significantly impacted the final grade distribution, even though the absolute numerical differences between graders were small. Grade plays a prominent role in medical education,[11] ultimately determining class rank or percentile, which is arguably more important than the raw scores students receive on assignments. This discrepancy suggests that ChatGPT, in its current version, is not ready to independently grade medical students' summative written assignments in a tiered grading system. Likewise, subanalysis (Appendix A), which revealed lack of a meaningful relationship between AI and human-generated scores, suggests that the AI scores are not aligned with human evaluative judgments and cannot currently be used as a reliable proxy or substitute for human scoring in this OSCE context.

This conclusion presupposes that the human grader is the reference for grading, which is a justified assumption based on review of ChatGPT's comments (Figure 1). The differences in the LLM's grading of the various sections of the SOAP note support using the human grader as the gold standard. The most significant grading discrepancies (as evidenced by Cohen's d) occurred in the categories of differential diagnosis/thought process and treatment plan. One might postulate that these domains' reliance on clinical reasoning, which is too nuanced for ChatGPT, led to this discrepancy. The LLM almost universally gave students full credit for the treatment plan (Table 1). The lower average for differential diagnosis/thought process (and tighter standard deviation) may indicate that the LLM overlooked written evidence of students' clinical reasoning. Alternatively, the findings may indicate leniency bias from the human grader, though the methods do not allow further investigation into this possibility.

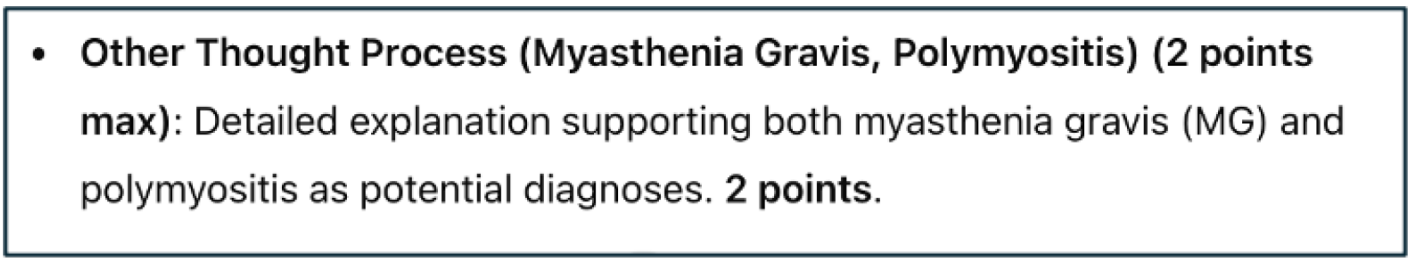
Despite prompting to "use the rubric" with each SOAP note, GPT frequently made its own assumptions about "reasonable" differential diagnoses and treatment plans, which were not explicitly outlined in the grading rubric. Diagnoses such as myasthenia gravis (Figure 1) were awarded points by GPT, despite not being mentioned overtly in the rubric (focal neurological deficits are not present in the patient script). GPT consistently awarded points for ordering lab tests that were not listed in the rubric. It also awarded points for vague tests such as "panel for blood disorders," which a clinical proctor would not consider specific enough.

TABLE 1. Comparison of Human Proctor Versus ChatGPT Note Grading

Exam element (points)	Proctor grade*	GPT grade*	*P* value	Cohen's *D*
History (7)	6.30 (0.86)	6.66 (0.48)	<.001	−0.52
Exam (7)	5.50 (1.6)	5.65 (0.69)	.34	–
Differential/ thought (9)	7.45 (1.29)	6.86 (0.43)	<.001	0.61
Treatment plan (5)	3.89 (1.16)	4.94 (0.29)	<.001	−1.25
Total grade (28)	23.13 (2.84)	24.11 (1.27)	<.001	−0.44

*Number expressed as mean (SD).
Abbreviations: GPT, generative pretrained transformer; SD, standard deviation

FIGURE 1. **Example of ChatGPT's Deviation From the Rubric**



Myasthenia gravis was not an acceptable diagnosis listed on the rubric. Despite this, ChatGPT awarded points for this diagnosis that it found to be "reasonable".

While these errors are anecdotal, they support the notion that ChatGPT-4 cannot fully replace a human grader. Deviations from the grading rubric indicate that medical educators bring domain-specific depth and pedagogical insight to grading, which ChatGPT-4 lacks.

However, ChatGPT certainly possesses potential to augment medical educators' efforts through grading and feedback. The LLM was able to reliably grade the history and physical exam sections of the SOAP notes. It gave detailed explanations for its grading (Figure 2) and, when prompted, summarized feedback for the student (Figure 3). Given the aforementioned limitations, the present technology seems best suited for formative educational assessments as opposed to graded, summative assessments, particularly in a tiered grading system. Alternatively, ChatGPT could be used to grade more straightforward aspects of documentation, such as the history and physical exam, while human proctors still evaluate the assessment and plan. This approach would nevertheless improve human proctors' grading efficiency.

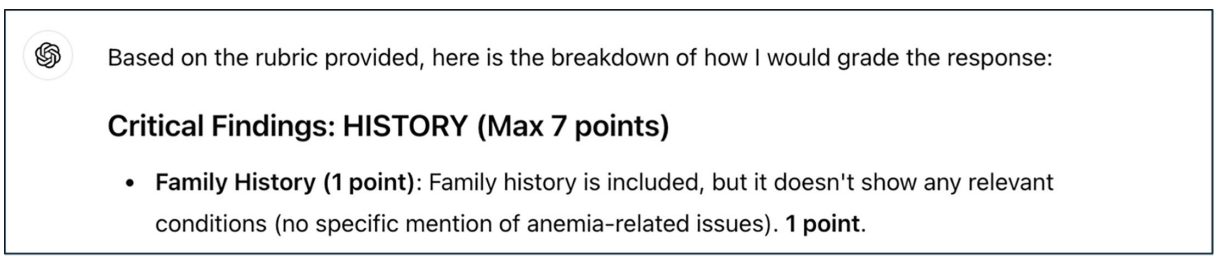
This study was limited in several ways. While it utilized a single LLM, studies comparing multiple LLMs show differences between their performance of the same task.[2,5] Just as there are multiple LLMs, there are also multiple ways to ask an LLM to perform the same task, including few-shot learning and complex training through fine-tuning. For this study, we elected to provide ChatGPT with a simple prompt and the established rubric to simulate the way a busy academic physician might use an LLM in the real world. This technique is known as "zero-shot" prompting, in which an LLM is simply asked to perform a new task for which it was not trained. To more definitively evaluate AI's ability to grade students' notes, one could consider fine-tuning an LLM. However, fine-tuned models are not yet widely available for this purpose. Future studies could compare the performance of different AI programs and training methods.

In addition, human grading of written assignments contains inherent subjectivity. All 127 SOAP notes were graded by the same proctor, thereby limiting interrater variability; but other factors, such as students' writing style, note organization, human error, and the order in which notes were reviewed, could have affected the proctor-assigned grades. Multiple human graders would provide further insight to an LLM's grading; but in everyday experience, one grader is a common practice, even for high-stakes written exams. If human error is present, it would partially explain the differences between proctor- and ChatGPT-assigned grades. Lastly, the cutoff for honors (80% or higher) was based on historical consensus, but analysis of the LLM's grading could provide a different cutoff that yields the desired grade distribution. AI's uses change and grow on a seemingly daily basis. Medical educators can use the results of this study to continue exploring LLMs' ability to grade, particularly as chain-of-thought models become more widely available for multistep tasks.

CONCLUSIONS

In conclusion, when compared to a human proctor, ChatGPT-4 assigned statistically different grades to medical students'

FIGURE 2. **Sample From ChatGPT's Grading Response**



 https://doi.org/10.22454/FamMed.2025.954255

FIGURE 3. Example of Summative Feedback Provided by ChatGPT

- **Areas for Improvement**:
  - **History**: Include GI symptoms, such as blood in stool or constipation, to complete the anemia assessment.
  - **Physical Exam**: Add capillary refill or mucous membrane findings and additional extremity details for completeness.
  - **Differential Diagnosis**: Expand on supporting evidence for hypothyroidism and depression to strengthen diagnostic reasoning.

The note demonstrates good clinical reasoning and thorough documentation, with minor adjustments needed to enhance completeness and depth in differential support.

SOAP notes based on an OSCE. While the difference in the mean grades assigned by proctor versus ChatGPT was small, the LLM's grading significantly impacted the final, tiered grade distribution. Medical educators should thoroughly investigate an LLM's performance characteristics in their local grading framework before using AI to augment grading of summative, written assessments.

REFERENCES

1. McCoy LG, Ci Ng FY, Sauer CM, et al. Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: a narrative review. *BMC Med Educ*. 2024;24(1):1096.
2. Ilgaz HB, Çelik Z. The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and Google Bard. *Cureus*. 2023;15(9):e45301.
3. Stretton B, Kovoor J, Arnold M, Bacchi S. ChatGPT-based learning: generative artificial intelligence in medical education. *Med Sci Educ*. 2024;34(1):215–217.
4. 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.
5. Hanna RE, Smith LR, Mhaskar R, Hanna K. Performance of language models on the family medicine in-training exam. *Fam Med*. 2024;56(9):555–560.
6. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141(8):798–800.
7. Hopkins BS, Nguyen VN, Dallas J, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *J Neurosurg*. 2023;139(3):904–911.
8. Wang L, Wan Z, Ni C, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. *J Med Internet Res*. 2024;26:e22769.
9. Gonzalez VH, Mattingly S, Wilhelm J, Hemingson D. Using artificial intelligence to grade practical laboratory examinations: sacrificing students' learning experiences for saving time? *Anat Sci Educ*. 2024;17(5):932–936.
10. Jamieson AR, Holcomb MJ, Dalton TO, et al. Rubrics to prompts: assessing medical student post-encounter notes with AI. *NEJM AI*. 2024;1(12).
11. Saguil A, Balog EK, Goldenberg MN, et al. The association between specialty match and third-year clerkship performance. *Mil Med*. 2012;177(9 Suppl):47–52.