Original Paper

Use of Large Language Models for Automated Short-Answer Question Grading in Second-Cycle Medical Education: Performance and Consistency Evaluation Study

Sabrina Sidali[1,2*], MD, PhD; Pierre-Antoine Soret[3,4*], MD, PhD; Mathilde Wagner[5,6], MD, PhD; Thibault Voron[7], MD, PhD; Quentin Vanderbecq[5,6*], MD, PhD

[1]Cordeliers Research Center, Sorbonne University, Inserm, University of Paris, Functional Genomics of Solid Tumors Team, French National Cancer League–Accredited Team, Labex OncoImmunology, F-75006 Paris, Île-de-France, France

[2]University of Paris, Greater Paris University Hospitals (AP-HP), DMU DIGEST, Beaujon Hospital, Reference Center for Vascular Liver Diseases, FILFOIE, ERN RARE-LIVER, Inflammation Research Center, Inserm UMR 1149, Paris, Île-de-France, France

[3]Sorbonne University, Inserm, Saint-Antoine Research Center (CRSA), Paris, Île-de-France, France

[4]Saint-Antoine Hospital, Greater Paris University Hospitals (AP-HP), Paris, Île-de-France, France

[5]Department of Radiology, Saint-Antoine Hospital, Greater Paris University Hospitals (AP-HP), Paris, Île-de-France, France

[6]LIB, UMR 7371, Sorbonne University, CNRS, Inserm U1146, Paris, Île-de-France, France

[7]Department of General and Digestive Surgery, Saint-Antoine Hospital, Greater Paris University Hospitals (AP-HP), Sorbonne University, Paris, Île-de-France, France

[*]these authors contributed equally

Corresponding Author:
Quentin Vanderbecq, MD, PhD
Department of Radiology, Saint-Antoine Hospital, Greater Paris University Hospitals (AP-HP)
184 Rue du Faubourg Saint-Antoine
Paris, Île-de-France, 75012
France
Phone: 33 149282000
Email: q.vanderbecq@gmail.com

Abstract

Background: Short-answer questions (SAQs) effectively assess higher-order reasoning skills in medical education but face challenges in scalability due to time-intensive manual grading and concerns about interrater reliability. Recent advances in large language models (LLMs) may offer a solution through automated assessment.

Objective: This study evaluated the performance and consistency of LLMs—ChatGPT, Claude Sonnet 4, and Mistral Pro—compared with those of human graders for SAQ grading in hepatology and gastroenterology.

Methods: We retrospectively analyzed SAQs from hepatology and gastroenterology examinations at Sorbonne University, Paris, France (2023-2025). Three human evaluators (2 hepatogastroenterologists and 1 radiologist) independently graded all responses twice, with the majority consensus score serving as the reference standard. Six LLMs were evaluated: ChatGPT (GPT-4o, GPT-4.5, GPT-4o3, and GPT-5), Claude Sonnet 4, and Mistral Pro. Each LLM graded all responses twice using standardized prompts, including questions, model answers, grading guidelines, and scoring rubrics. Agreement was assessed using quadratic weighted kappa (QWK), Cohen κ, the intraclass correlation coefficient (ICC), and the mean absolute error (MAE) for interhuman, intrahuman, LLM-versus-human, and intramodel comparisons.

Results: A total of 902 student responses from 12 SAQs were included and analyzed. Interhuman grader agreement was almost perfect (QWK 0.90-0.92; ICC 0.90-0.92; MAE 0.05-0.07). Claude Sonnet 4 and GPT-4o3 demonstrated near-human–level performance when compared with the majority vote of human evaluators (Claude Sonnet 4: QWK 0.88-0.90; ICC 0.88-0.90; MAE 0.06-0.07 and GPT-4o3: QWK 0.82-0.89; ICC 0.82-0.89; MAE 0.07-0.09). GPT-5 showed substantial agreement (QWK 0.77-0.79). Mistral Pro and older ChatGPT versions performed poorly (QWK <0.63). Intramodel consistency was the highest for Claude Sonnet 4 (QWK 0.93), approaching the best human intrarater reliability. Agreement did not correlate with prompt design choices (number of examples or minimum word count).

XSL•FO
RenderX

Conclusions: Advanced LLMs, particularly Claude Sonnet 4 and GPT-4o3, demonstrated near-human–level grading performance, supporting their potential use as assistive tools in SAQ grading. However, performance varied by model, and agreement alone does not ensure educational validity. Until broader multicenter validation is conducted across diverse domains, stakes, and question formats—along with safeguards for bias, privacy, and model drift—LLMs should augment rather than replace human faculty graders.

(JMIR Med Educ 2026;12:e89445) doi: 10.2196/89445

KEYWORDS

large language models; short-answer grading; artificial intelligence; AI; medical education

Introduction

The assessment of learning and competencies is a central pedagogical approach in initial health sciences training programs, whether during the preclinical curriculum, in the context of clinical rotations in the undergraduate program, or in the context of rotations in the postgraduate program [1]. Assessments are used, particularly in the preclinical curriculum, to provide an initial checkpoint for student learning, that is, to verify that students possess adequate knowledge and the necessary skills to enter the clinical curriculum [1]. Among the various assessment formats used in medical education, short-answer questions (SAQs) have emerged as a valuable tool for evaluating student knowledge and clinical reasoning capabilities. Indeed, SAQs constitute a widely recognized assessment format for evaluating the knowledge and cognitive skills of medical students [2-4]. In contrast to multiple-choice questions, which primarily assess the recall of discrete facts and may encourage superficial memorization strategies, SAQs offer the opportunity to evaluate a broader spectrum of competencies, including theoretical understanding, analytic reasoning, and reflective thinking [5]. Furthermore, research in cognitive psychology suggests that the retrieval practice inherent in SAQ formats promotes superior long-term retention compared with recognition-based assessments [6]. This capacity to capture higher-order cognitive processes makes SAQs particularly valuable in the context of competency-based medical education.

However, despite these advantages, the large-scale implementation of SAQs in medical school examinations remains constrained by several factors. Chief among these is the significant administrative and faculty burden imposed on institutions with large student cohorts. Specifically, the time-intensive nature of manual grading represents a significant practical constraint, particularly for large student cohorts. Moreover, human grading is inherently susceptible to variability and potential bias, thereby raising concerns about interrater reliability and the objectivity of assessment outcomes [7]. These limitations have historically restricted the use of SAQs in high-stakes examinations, despite their pedagogical advantages.

Recent advances in AI, particularly the development of large language models (LLMs), offer a promising avenue for addressing these challenges. LLMs have demonstrated the capacity to process and evaluate natural language responses at scale, potentially enabling automated grading of SAQs with accuracy comparable to that of human evaluators. If validated, such systems could reduce the logistical and temporal barriers associated with SAQ assessment, thereby facilitating their broader adoption in high-stakes medical examinations [7-9].

Therefore, in this study we aimed to evaluate the performance, of 6 models of 3 LLMs—ChatGPT (GPT-4o, GPT-4.5, GPT-4o3, and GPT-5), Claude Sonnet 4, and Mistral Pro—in the automated grading of SAQs from second-cycle examinations in hepatology and gastroenterology.

Methods

Study Design and Data Selection

We analyzed retrospective data from the SAQs of second-cycle hepatology and gastroenterology examinations conducted in the 2023 to 2024 and 2024 to 2025 academic years at Sorbonne University Medical School, Paris, France.

The dataset included the questions administered to students, the students' responses, the grading guidelines provided to evaluators, and the scores awarded to students.

Assessment Procedure

Each student response underwent dual grading: human grading and LLM grading. Human grading was performed twice at 3-month intervals by 3 independent evaluators (2 hepato-gastroenterologist (PAS [evaluator 1] and SS [evaluator 2]), and 1 radiologist (QV [evaluator 3]). The 3 evaluators assigned a score of 0, 0.5, or 1 to each student response, based on the grading guidelines set by the teacher who wrote the question. The final score for human grading corresponded to the majority score assigned by at least 2 of the 3 evaluators from their first evaluation. A median score of 0.5 was assigned to 3 different responses.

LLM grading was performed twice by the 6 selected LLMs: 4 versions of ChatGPT (GPT-4o [July 2025 version], GPT-4.5 [July 2025 version], GPT-4o3 [July 2025 version], and GPT-5 with advanced reasoning ["Thinking"; September 2025 version]; OpenAI), Claude Sonnet 4 (September 2025 version; Anthropic), and Mistral Pro (July 2025 version; Mistral AI). Multiple ChatGPT versions were included to assess performance evolution across model iterations, whereas only the most recent versions of Claude Sonnet 4 and Mistral Pro were tested as these represented the state-of-the-art models available from these providers at the time of the study. All models were queried with default temperature settings unless otherwise specified. For each question, the LLMs received the following information: the question prompt, several examples of correct answers, grading guidelines provided to human evaluators, if available, the list of student responses, and the scoring rubric. Multiple

XSL•FO
RenderX

responses from the same SAQ were provided within the same prompt rather than graded individually in separate prompts. To reduce redundancy, identical student responses were included only once. Each LLM assigned a score of 0, 0.5, or 1 to each student response. An example of a typical prompt used is provided in Multimedia Appendix 1.

Ethical Considerations

This retrospective study was conducted in accordance with institutional guidelines and approved by Sorbonne University Medical School. The study was approved by the institutional review board under approval number CRM-2607-587. All examination data were fully anonymized by the university prior to analysis to ensure student confidentiality, and no individual student identifiers were accessible to the research team. As this study involved only the secondary analysis of existing anonymized educational data, individual informed consent was not required.

Statistical Analysis

Data are presented as median (IQR) or number (proportion) as appropriate. Comparisons of quantitative and qualitative variables were made using the Mann-Whitney *U* test and chi-square or Fisher exact test, respectively.

Comparison of Human Grading

Scores assigned by 3 independent evaluators were compared with each other using quadratic weighted kappa (QWK), Cohen κ, intraclass correlation coefficient (ICC; absolute agreement) and mean absolute error (MAE) to assess consistency among human evaluators. Global interrater agreement among the 3 evaluators was assessed using Fleiss κ. To assess intrarater agreement, each evaluator reannotated all responses after a minimum of a 3-month interval, allowing sufficient time to minimize recall bias. Agreement between the first and second annotations was assessed using QWK, Cohen κ, ICC, and MAE. Interevaluator agreement was scored according to the Landis and Koch criteria [10].

Comparison Between LLMs and Human Grading

The performance of each LLM was compared with reference human grading. For each SAQ, the score assigned by each LLM was compared with the score given by the human grader. Agreement between each LLM and the human grader by majority voting was assessed using QWK, Cohen κ coefficient, ICC, and MAE.

Intramodal Agreement

To assess the stability of LLM grading over time, each model was prompted 2 times on the same set of SAQs. Agreement across repeated inferences from the same model was quantified using QWK, Cohen κ, ICC, and MAE, analogous to intrarater agreement in human evaluators.

Correlation Analysis

A correlation analysis (Spearman rank correlation coefficient) was performed to examine the association between QWK and 2 prompt design variables: the number of examples included in the prompt and the minimum number of words required in the response. The minimum threshold was set with the understanding that a correct response may naturally be longer, depending on the use of synonyms or alternative phrasing.

All tests were 2-sided, and $P<.05$ was considered significant. Data handling and analysis were performed using Python (version 3.11; Python Software Foundation) and scikit-learn (version 1.7) [11].

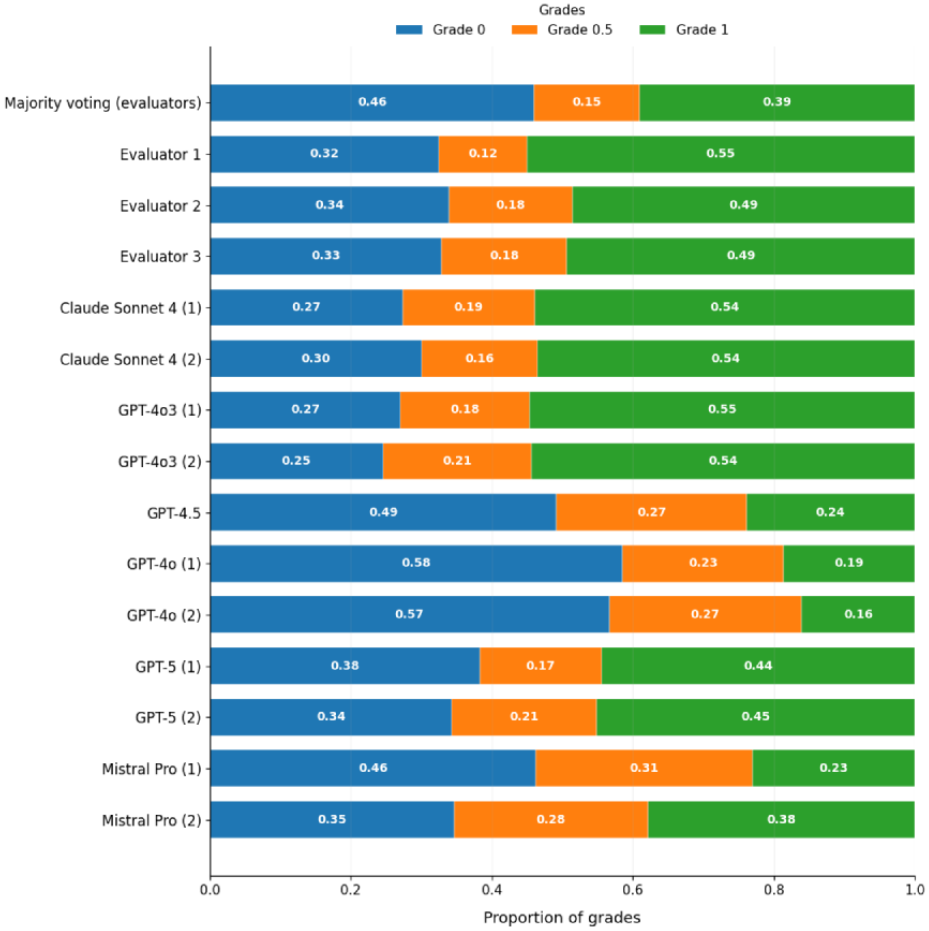
Results

Overview

A total of 902 student responses from 12 SAQs across 6 second-cycle hepatology and gastroenterology examinations were analyzed. The median number of student responses per SAQ was 53 (IQR 26-116; range 16-208). The median number of examples provided to the LLMs was 3.5 (IQR 2.0-4.5; range 1-10). The median prompt length (excluding spaces) was 1213 (IQR 954-1262; range: 672-1518) characters. The distribution of grades across models and human evaluators is represented in Figure 1.

XSL•FO
RenderX

Figure 1. Distribution of grades across models or evaluators.



Comparison of Human Grading

Interrater agreement was high across all evaluator pairs, with QWK values ranging from 0.90 to 0.92, ICCs from 0.90 to 0.92, and MAEs between 0.05 and 0.07. Pairwise Cohen's kappa values ranged from 0.81 to 0.87, confirming strong consistency among graders. Intrarater agreement was also robust: evaluator 2 demonstrated almost perfect reliability (QWK and ICC 0.98), while evaluators 1 and 3 achieved substantial to almost perfect agreement (QWK 0.85 and 0.90, respectively) according to Landis and Koch criteria [10] (Table 1 and Figures 2-3).

XSL•FO
RenderX

Table 1. Interrater and intrarater agreement among human evaluators.

	QWK[a] (95% CI)	Cohen κ (95% CI)	ICC[b] (95% CI)	MAE[c] (95% CI)
Interrater agreement[d]				
Evaluators 1 and 2	0.91 (0.88-0.93)	0.86 (0.83-0.88)	0.91 (0.88-0.93)	0.05 (0.03-0.07)
Evaluators 2 and 3	0.90 (0.87-0.92)	0.81 (0.78-0.84)	0.90 (0.87-0.92)	0.07 (0.05-0.08)
Evaluators 3 and 1	0.92 (0.90-0.94)	0.87 (0.84-0.89)	0.92 (0.90-0.94)	0.05 (0.04-0.06)
Intrarater agreement				
Evaluator 1	0.85 (0.82-0.88)	0.84 (0.81-0.87)	0.85 (0.82-0.88)	0.07 (0.06-0.09)
Evaluator 2	0.98 (0.96-0.99)	0.98 (0.97-0.99)	0.98 (0.96-0.99)	0.009 (0.003-0.017)
Evaluator 3	0.90 (0.88-0.91)	0.76 (0.72-0.79)	0.90 (0.88-0.91)	0.08 (0.07-0.09)
Agreement with majority voting[d]				
Evaluator 1	0.97 (0.96-0.98)	0.95 (0.93-0.97)	0.97 (0.96-0.98)	0.02 (0.01-0.02)
Evaluator 2	0.95 (0.94-0.97)	0.90 (0.88-0.92)	0.95 (0.94-0.97)	0.03 (0.02-0.04)
Evaluator 3	0.95 (0.93-0.97)	0.91 (0.89-0.93)	0.95 (0.93-0.97)	0.03 (0.02-0.04)

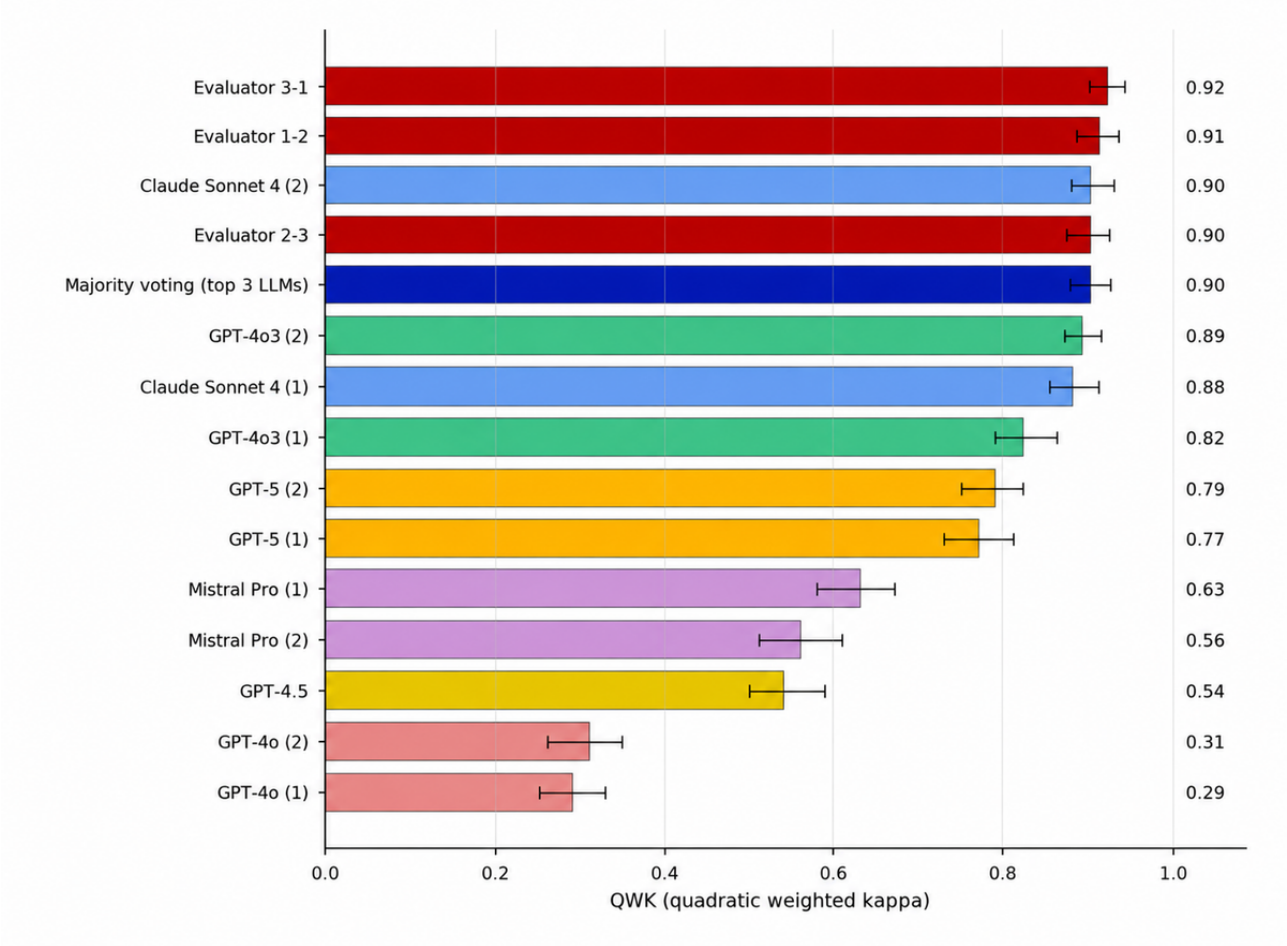
[a]QWK: quadratic weighted kappa.

[b]ICC: intraclass correlation coefficient.

[c]MAE: mean absolute error.

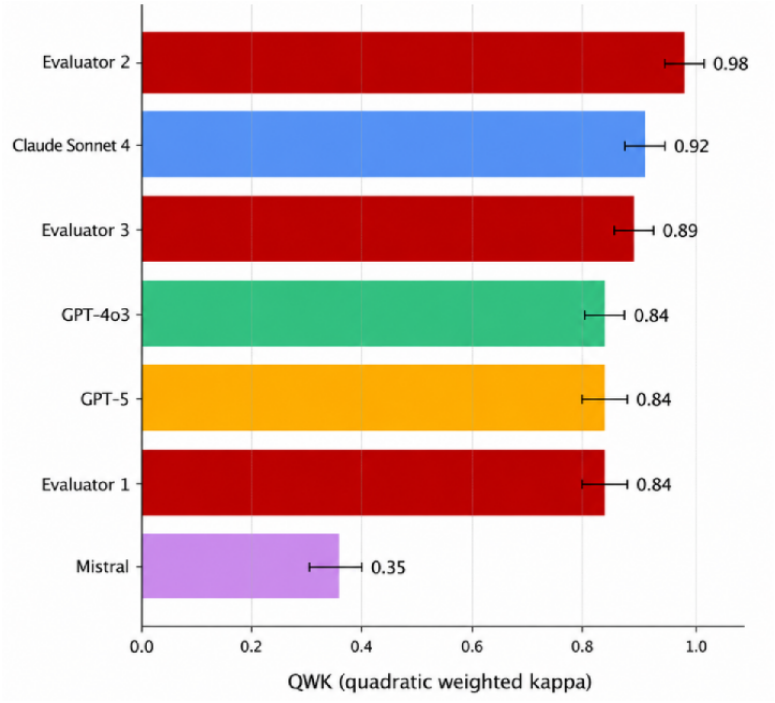
[d]Interrater agreement and agreement against the majority vote were calculated only on the first evaluation of each evaluator.

Figure 2. Quadratic weighted kappa (QWK) score for humans and large language models (LLMs). QWK scores were calculated between pairs of human evaluators based on their initial annotations. For LLMs, QWK scores were calculated against the majority vote of the 3 human evaluators using the initial (1) or repeated (2) LLM assessments. CIs were estimated using bootstrapping with 3000 samples and plotted accordingly. MV: majority voting.



XSL•FO
RenderX

Figure 3. Intrarater and intramodel agreement for evaluators and the best-performing large language models (LLMs). Quadratic weighted kappa (QWK) scores were calculated between pairs of evaluations. CIs were estimated using bootstrapping with 3000 samples and plotted accordingly.



Comparison Between LLMs and Human Grading

A comprehensive comparison of all metrics across models and human evaluators is presented in Table 2. When compared with the majority vote of human evaluators (Table 2 and Figure 2), the best-performing models were Claude Sonnet 4 (QWK 0.88-0.90; ICC 0.88-0.90; MAE 0.06-0.07) and GPT-4o3 (QWK 0.82-0.89; ICC 0.82-0.89; MAE 0.07-0.09), both achieving almost perfect agreement according to the Landis and Koch criteria. These values approached those observed between human evaluator pairs (QWK 0.90-0.92; ICC 0.90-0.92; MAE 0.05-0.07), indicating near-human–level performance. GPT-5 demonstrated substantial agreement (QWK 0.77-0.79; ICC 0.77-0.79; MAE 0.12-0.13), though lower than any human evaluator pair. Mistral Pro showed moderate alignment (QWK 0.56-0.63; ICC 0.56-0.63; MAE 0.22-0.25), while older versions of ChatGPT (GPT-4o and GPT-4.5) exhibited fair to moderate agreement (QWK <0.55 and MAE >0.25) clearly below human consistency levels.

In terms of reproducibility (Table 3 and Figure 3), intramodel consistency was the highest for Claude Sonnet 4 (QWK 0.93; ICC 0.93; MAE 0.04), approaching but not exceeding the best human intrarater agreement (evaluator 2: QWK 0.98; ICC 0.98; MAE 0.009) and surpassing evaluators 1 and 3 (QWK 0.85-0.90; ICC 0.85-0.90; MAE 0.07-0.08). GPT-4o3 and GPT-5 also demonstrated almost perfect stability (QWK 0.85; ICC 0.85; MAE≈0.08), comparable to human evaluators 1 and 3. By contrast, Mistral Pro showed fair intramodel reliability (QWK 0.36; ICC 0.36; MAE 0.28), substantially below any human evaluator's test-retest performance.

Question-level analysis revealed a strong correlation between interhuman agreement and LLM performance. Among the 5 SAQs with the lowest interhuman agreement, 4 also yielded the poorest performance for GPT-4o3 and Claude Sonnet 4. These questions were neither those with the greatest diversity of student responses nor those requiring the longest average response length (number of words).

XSL•FO
RenderX

Table 2. Agreement between large language models (LLMs) versus majority voting of evaluators[a].

LLM vs majority voting	QWK[b] (95% CI)	Cohen κ (95% CI)	ICC[c] (95% CI)	MAE[d] (95% CI)
Claude Sonnet 4 (2)	0.90 (0.88-0.93)	0.83 (0.80-0.86)	0.90 (0.88-0.93)	0.06 (0.05-0.07)
Claude Sonnet 4 (1)	0.88 (0.85-0.91)	0.81 (0.77-0.84)	0.88 (0.85-0.91)	0.07 (0.06-0.08)
GPT-4o3 (2)	0.89 (0.87-0.91)	0.79 (0.76-0.82)	0.89 (0.87-0.91)	0.07 (0.05-0.08)
GPT-4o3 (1)	0.82 (0.79-0.86)	0.78 (0.74-0.81)	0.82 (0.79-0.86)	0.09 (0.07-0.11)
GPT-5 (2)	0.79 (0.75-0.82)	0.69 (0.65-0.72)	0.79 (0.75-0.82)	0.12 (0.11-0.14)
GPT-5 (1)	0.77 (0.73-0.81)	0.67 (0.63-0.71)	0.77 (0.73-0.81)	0.13 (0.11-0.15)
GPT-4.5	0.54 (0.50-0.59)	0.40 (0.36-0.44)	0.54 (0.50-0.59)	0.26 (0.24-0.29)
GPT-4o (2)	0.31 (0.26-0.35)	0.31 (0.27-0.35)	0.31 (0.26-0.35)	0.37 (0.34-0.40)
GPT-4o (1)	0.29 (0.25-0.33)	0.23 (0.19-0.26)	0.29 (0.25-0.33)	0.39 (0.36-0.42)
Mistral Pro (1)	0.63 (0.58-0.67)	0.43 (0.38-0.47)	0.63 (0.58-0.67)	0.22 (0.20-0.24)
Mistral Pro (2)	0.56 (0.51-0.61)	0.38 (0.33-0.43)	0.56 (0.51-0.61)	0.25 (0.23-0.28)

[a]The numbers (1) and (2) following model names indicate the initial and repeated grading assessments, respectively.

[b]QWK: quadratic weighted kappa.

[c]ICC: intraclass correlation coefficient.

[d]MAE: mean absolute error.

Table 3. Intramodel agreement among large language models.

	QWK[a] (95% CI)	Cohen κ (95% CI)	ICC[b] (95% CI)	MAE[c] (95% CI)
GPT-4o3	0.85 (0.82-0.88)	0.77 (0.74-0.80)	0.85 (0.82-0.88)	0.08 (0.07-0.09)
GPT-5	0.85 (0.83-0.88)	0.79 (0.76-0.82)	0.85 (0.83-0.88)	0.08 (0.07-0.10)
Mistral Pro	0.36 (0.30-0.42)	0.38 (0.34-0.43)	0.36 (0.30-0.42)	0.28 (0.25-0.30)
Claude Sonnet 4	0.93 (0.91-0.95)	0.90 (0.87-0.92)	0.93 (0.92-0.95)	0.04 (0.03-0.05)

[a]QWK: quadratic weighted kappa.

[b]ICC: intraclass correlation coefficient.

[c]MAE: mean absolute error.

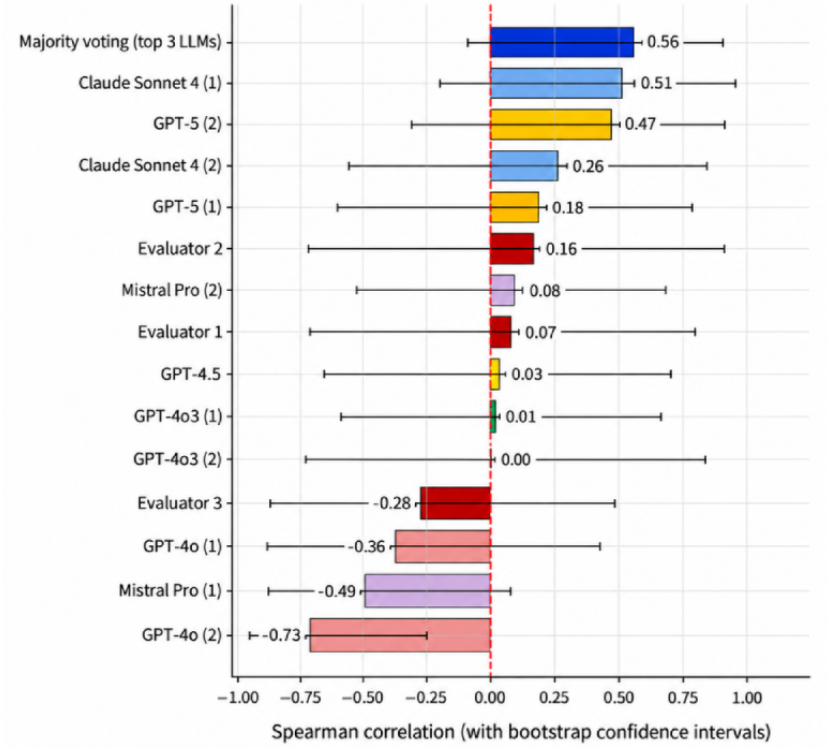
Correlation Analysis: Association Between Grading Agreement (QWK) and Prompt Design

Correlations between QWK and the number of examples provided were generally weak and not statistically significant across all models and human evaluators (Figure 4 and Table 4). Most coefficients were close to 0, with wide CIs spanning both negative and positive values. The strongest positive trends were observed for the consensus score of the 3 best-performing LLMs ("LLM (top3)" (ρ [rho]=0.56, 95% CI −0.09 to 0.92; *P*=.09) and Claude Sonnet 4 at first annotation ("Claude Sonnet 4 (1)" (ρ [rho]=0.51, 95% CI −0.18 to 0.90; *P*=.13), but neither reached statistical significance.

Correlations between QWK and the minimum number of words required for a full score were also low overall (Table 4). No consistent or significant association was observed across models or evaluators. The strongest negative trend appeared for Mistral Pro at first annotation ("Mistral Pro (1)" ρ [rho]=−0.59, 95% CI −0.89 to 0.02; *P*=.07), suggesting a possible inverse relationship, though this did not achieve statistical significance. Human evaluators showed comparable weak correlations, with values ranging from −0.16 to 0.50 and wide CIs.

XSL•FO
RenderX

Figure 4. Correlation between quadratic weighted kappa scores for student responses and the number of examples provided. The numbers (1) and (2) indicate the initial and repeated assessments by the large language models (LLMs), respectively.



XSL•FO
RenderX

Table 4. Correlation between response length and number of examples provided[a].

Correlation characteristic and model or evaluator	ρ (rho) (95% CI)	*P* value[b]
Number of examples provided		
GPT-4.5	0.03 (−0.67 to 0.71)	.94
LLM[c] MV[d] (top3)	0.56 (−0.09 to 0.92)	.09
GPT-5 (1)	0.20 (−0.60 to 0.75)	.58
GPT-5 (2)	0.49 (−0.30 to 0.87)	.15
Mistral pro (1)	−0.49 (−0.86 to 0.06)	.15
Mistral pro (2)	0.08 (−0.53 to 0.68)	.84
GPT-4o3 (1)	0.01 (−0.59 to 0.67)	.97
GPT-4o3 (2)	0.00 (−0.74 to 0.81)	>.99
Claude Sonnet 4 (1)	0.51 (−0.18 to 0.90)	.13
Claude Sonnet 4 (2)	0.26 (−0.53 to 0.82)	.46
Evaluator 1	0.07 (−0.71 to 0.76)	.85
Evaluator 2	0.18 (−0.71 to 0.89)	.63
Evaluator 3	−0.34 (−0.87 to 0.51)	.34
Minimum number of words required for a full score		
GPT-4.5	−0.34 (−0.97 to 0.51)	.34
LLM MV (top3)	0.21 (−0.58 to 0.95)	.56
GPT-5 (1)	−0.23 (−0.86 to 0.57)	.51
GPT-5 (2)	−0.16 (−0.82 to 0.61)	.66
Mistral Pro (1)	−0.59 (−0.89 to 0.02)	.07
Mistral Pro (2)	−0.02 (−0.74 to 0.75)	.96
GPT-4o3 (1)	0.08 (−0.59 to 0.91)	.83
GPT-4o3 (2)	−0.28 (−0.82 to 0.59)	.44
Claude Sonnet 4 (1)	0.36 (−0.42 to 0.95)	.30
Claude Sonnet 4 (2)	0.31 (−0.39 to 0.83)	.38
Evaluator 1	0.36 (−0.40 to 0.86)	.30
Evaluator 2	0.50 (−0.17 to 0.90)	.14
Evaluator 3	−0.16 (−0.72 to 0.78)	.65

[a]Spearman correlations were computed between the number of examples provided or the minimum words required for a full score and QWK scores of models or evaluators against the majority vote of the 3 human evaluators. The numbers (1) and (2) indicate the initial and repeated assessments by the large language models, respectively. 95% CIs were estimated by bootstrapping.

[b]*P* values were obtained from Spearman rank test.

[c]LLM: large language model.

[d]MV: majority voting.

Discussion

In our study, Claude Sonnet 4 and GPT-4o3 showed the closest alignment with human grading, with stable reproducibility across repeated inferences. None of the models surpassed human evaluators in either agreement or reproducibility; the best systems reached human‐level ranges of agreement and reproducibility. Correlation analyses did not reveal clear associations between the prompt design features examined and model performance. These findings suggest that LLM-based grading systems hold promise for enhancing the scalability, efficiency, and potential objectivity of medical education assessment. However, it should be noted that human and LLM biases differ fundamentally: while human graders may be affected by fatigue, familiarity with students, or inconsistent rubric interpretation, LLMs may exhibit systematic biases inherited from pretraining data, sensitivity to prompt framing, or tendencies toward certain response patterns—none of which are yet fully characterized in medical education assessment contexts.

XSL•FO
RenderX

Our results are consistent with existing data. Several studies suggest that LLMs can approximate, but not consistently exceed, human grading in medical education. Seneviratne and Manathunga [12] reported a strong correlation between an AI-assisted SAQ scoring tool and human examiner markings, supporting the feasibility of automated SAQ scoring when aligned with examiner rubrics. Similarly, Rajan et al [13] reported that GPT-4 scoring could be statistically equivalent to faculty grading at the aggregate level, while discrepancies persisted on items requiring more complex reasoning, supporting the feasibility of LLM-assisted scoring without establishing validity across all unseen assessment contexts. Beyond SAQs, Burke et al [14] showed that ChatGPT could score structured free-text clinical notes with fewer scoring errors than human graders, suggesting that LLMs may perform particularly well when evaluation criteria are explicit and fact-based.

At the same time, published findings remain heterogeneous regarding the LLM grading performance in medical education. For example, Grevisse [7] observed moderate agreement between GPT-4 or Gemini and SAQ instructors and noted that GPT-4 tended to be conservative in awarding full credit, highlighting the need for human oversight. Similarly, Bolgova et al [15] showed that expert-expert agreement remained higher than LLM-expert agreement, with performance varying across questions and depending on rubric structure, highlighting the value of expert-expert agreement for identifying settings in which human oversight remains important. In a medical education setting on mechanical ventilation, ChatGPT demonstrated weak agreement with human graders, and the authors cautioned against using LLMs as sole graders in high-stakes contexts [16].

Emerging evidence also suggests that agreement between LLMs and faculty may not be uniform across question types: in a case-based short-answer exam, LLM grading was comparable to faculty overall, but discrepancies increased for harder questions and for higher Bloom's levels (eg, understanding and evaluating), implying that LLM performance may depend on the cognitive processes targeted by each SAQ rather than reflecting a stable "global" grading ability [13]. In other domains, GPT-4 achieved near-human consistency for lower-stakes short-answer scoring in K-12 education, reinforcing the idea that task difficulty, stakes, and rubric clarity likely influence observed agreement [17].

In our dataset, some questions yielded lower interrater consistency among human graders than among certain LLMs. Rather than indicating that humans "underperformed," this pattern more plausibly suggests that a subset of SAQs may have been insufficiently precise or supported by incomplete grading guidance, increasing interpretive latitude. This interpretation aligns with works showing that question framing and prompt design can strongly influence LLM outputs [18,19]. Clearer and more specific prompts consistently yield more accurate and relevant outputs because prompt engineering acts as a "steering mechanism" that guides the model's reasoning. A notable example is chain-of-thought prompting: simply adding a phrase like "Let's think step by step" can encourage systematic reasoning and dramatically improve accuracy on complex tasks; for instance, boosting GPT-3's performance on math word problems from about 17% to 78% [20]. Numerous studies likewise show that prompts providing context, explicit instructions, or stepwise guidance enhance LLM outputs across domains, including education, where well-designed prompts have been found to significantly improve the usefulness of responses [21]. Together, these findings emphasize that observed LLM-human agreement is not solely a property of the model but also of assessment and implementation design, including question construction, rubric specificity, and prompting strategy (instruction clarity, context provided, and scoring constraints).

This study has several limitations. First, the limited sample size (12 SAQs from a single institution and specialty) constrained statistical power, generalizability, and the robustness of question-level exploratory analyses. Moreover, strong performance on this specific dataset does not guarantee correct assessment in other medical domains, as LLM outputs depend not only on pretraining data but also on rubric quality, question framing, and contextual factors specific to each assessment setting. Second, grading guidelines were not uniformly available, which may have introduced variability in both human and model scoring. Third, multiple student responses were presented within a single prompt for each SAQ rather than being graded in separate prompts. Although identical responses were included only once to reduce redundancy, potential effects of context or order cannot be excluded. Fourth, we evaluated specific LLM versions (July-September 2025) using default parameters without optimization; rapidly evolving models may demonstrate different capabilities. Specifically, LLMs were not trained using previous questions to improve scoring performance. Fifth, practical implementation challenges, including cost, latency, privacy, and student acceptance, were not examined. Finally, our analysis focused on binary and partial-credit scoring for SAQs; performance may differ for other assessment formats (eg, extended responses, oral exams, and objective structured clinical examination [OSCE] narratives) and more granular scales. In addition, although the majority vote of human graders served as a practical reference standard, high agreement with this standard and strong reproducibility should not be interpreted as equivalent to educational validity, particularly in high-stakes assessment contexts. Multicenter studies across specialties, languages, and SAQ formats are needed to better characterize reliability, reproducibility, and external validity in real-world settings.

In conclusion, Claude Sonnet 4 and GPT-4o3 demonstrated the strongest alignment with human grading in our second-cycle medical education SAQ context, potentially enabling more scalable use of SAQs while preserving near-human consistency. However, models have not yet surpassed human evaluators in reproducibility or accuracy across all question types, and performance appears sensitive to rubric quality and cognitive complexity. Approaches such as retrieval-augmented generation and parameter-efficient fine-tuning may further improve task-specific reliability. At present, LLMs should be considered supportive tools integrated into grading workflows with transparent validation and ongoing monitoring, rather than autonomous replacements for faculty graders, particularly for high-stakes assessments where interpretability, fairness, and educational validity are essential.

XSL•FO
RenderX

Acknowledgments

The authors would like to acknowledge the contributions of Professor Fleur Cohen and the AI Sorbonne University committee for providing the authors with the medical school short-answer examination and for supporting this project.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: SS, PAS, QV, MW
Data curation: SS, PAS, QV, TV
Formal analysis: SS, PAS, QV
Investigation: SS, PAS, QV
Writing—original draft: SS, PAS, QV, MW, TV
Writing—review and editing: SS, PAS, QV, MW, TV

Conflicts of Interest

SS received funding from Roche and AstraZeneca for invited talks, and from Elsevier for writing educational content. PAS received funding from AbbVie, Gilead, Ipsen, and Mirum for invited talks and participation on advisory boards. MW received travel support from Guerbet. The other authors declare no conflicts of interest.

Multimedia Appendix 1

Example of a standardized prompt used to grade short-answer questions with large language models.
[DOCX File , 13 KB-Multimedia Appendix 1]

References

1. van der Vleuten C. Validity of final examinations in undergraduate medical training. BMJ. Nov 11, 2000;321(7270):1217-1219. [FREE Full text] [doi: 10.1136/bmj.321.7270.1217] [Medline: 11073517]
2. Mee J, Pandian R, Wolczynski J, Morales A, Paniagua M, Harik P, et al. An experimental comparison of multiple-choice and short-answer questions on a high-stakes test for medical students. Adv Health Sci Educ Theory Pract. Jul 2024;29(3):783-801. [FREE Full text] [doi: 10.1007/s10459-023-10266-3] [Medline: 37665413]
3. Sam AH, Field SM, Collares CF, van der Vleuten CP, Wass VJ, Melville C, et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ. Apr 2018;52(4):447-455. [FREE Full text] [doi: 10.1111/medu.13504] [Medline: 29388317]
4. Puthiaparampil T, Rahman MM. Very short answer questions: a viable alternative to multiple choice questions. BMC Med Educ. May 06, 2020;20(1):141. [FREE Full text] [doi: 10.1186/s12909-020-02057-w] [Medline: 32375739]
5. Burrows S, Gurevych I, Stein B. The eras and trends of automatic short answer grading. Int J Artif Intell Educ. Mar 23, 2015;25(1):60-117. [doi: 10.1007/s40593-014-0026-8]
6. Greving S, Richter T. Practicing retrieval in university teaching: short-answer questions are beneficial, whereas multiple-choice questions are not. J Cogn Psychol. Jun 08, 2022;34(5):657-674. [FREE Full text] [doi: 10.1080/20445911.2022.2085281]
7. Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ. Sep 27, 2024;24(1):1060. [FREE Full text] [doi: 10.1186/s12909-024-06026-5] [Medline: 39334087]
8. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. Jun 01, 2023;9:e48291. [FREE Full text] [doi: 10.2196/48291] [Medline: 37261894]
9. Impey C, Wenger M, Garuda N, Golchin S, Stamer S. Using large language models for automated grading of student writing about science. Int J Artif Intell Educ. Dec 2025;35(4):1825-1859. [doi: 10.1007/s40593-024-00453-7]
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159-174. [Medline: 843571]
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. arXiv. Preprint posted online on January 2, 2012. [doi: 10.48550/arXiv.1201.0490]

XSL•FO
RenderX

12. Seneviratne HM, Manathunga SS. Artificial intelligence assisted automated short answer question scoring tool shows high correlation with human examiner markings. BMC Med Educ. Aug 05, 2025;25(1):1146. [FREE Full text] [doi: 10.1186/s12909-025-07718-2] [Medline: 40764994]
13. Rajan A, Alexander SM, Shenvi CL. Can AI grade like a professor? Comparing artificial intelligence and faculty scoring of medical student short-answer clinical reasoning exams. Adv Health Sci Educ Theory Pract. Apr 2026;31(2):607-617. [doi: 10.1007/s10459-025-10462-3] [Medline: 40767985]
14. Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, et al. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ. Jul 25, 2024;10:e56342. [FREE Full text] [doi: 10.2196/56342] [Medline: 39118469]
15. Bolgova O, Ganguly P, Ikram MF, Mavrych V. Evaluating large language models as graders of medical short answer questions: a comparative analysis with expert human graders. Med Educ Online. Dec 2025;30(1):2550751. [FREE Full text] [doi: 10.1080/10872981.2025.2550751] [Medline: 40849930]
16. Jade T, Yartsev A. ChatGPT for automated grading of short answer questions in mechanical ventilation. arXiv. Preprint posted online on May 5, 2025. [FREE Full text] [doi: 10.48550/arXiv.2505.04645]
17. Henkel O, Boxer A, Hills L, Roberts B, Levonian Z. Can large language models make the grade? An empirical study evaluating LLMs ability to mark short answer questions in K-12 education. In: L@S '24: Proceedings of the Eleventh ACM Conference on Learning @ Scale. New York, NY. Association for Computing Machinery; 2024:300-304.
18. Wu JH, Nishida T, Moghimi S, Weinreb RN. Effects of prompt engineering on large language model performance in response to questions on common ophthalmic conditions. Taiwan J Ophthalmol. Aug 27, 2024;14(3):454-457. [doi: 10.4103/tjo.TJO-D-23-00193] [Medline: 39430350]
19. He J, Rungta M, Koleczek D, Sekhon A, Wang FX, Hasan S. Does prompt formatting have any impact on LLM performance? arXiv. Preprint posted online on November 15, 2024. [doi: 10.48550/arXiv.2411.10541]
20. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv. Preprint posted online on May 24, 2022. [doi: 10.48550/arXiv.2205.11916]
21. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med. Feb 20, 2024;7(1):41. [FREE Full text] [doi: 10.1038/s41746-024-01029-4] [Medline: 38378899]

Abbreviations

ICC: intraclass correlation coefficient
LLM: large language model
MAE: mean absolute error
OSCE: objective structured clinical examination
QWK: quadratic weighted kappa
SAQ: short-answer question

Edited by P Kanzow; submitted 15.Dec.2025; peer-reviewed by ABL Narayanan, L Peng; comments to author 11.Apr.2026; revised version received 07.Jul.2026; accepted 11.Jul.2026; published 02.Nov.2026

Please cite as:
Sidali S, Soret P-A, Wagner M, Voron T, Vanderbecq Q
Use of Large Language Models for Automated Short-Answer Question Grading in Second-Cycle Medical Education: Performance and Consistency Evaluation Study
JMIR Med Educ 2026;12:e89445
URL: https://mededu.jmir.org/2026/1/e89445
doi: 10.2196/89445
PMID:

©Sabrina Sidali, Pierre-Antoine Soret, Mathilde Wagner, Thibault Voron, Quentin Vanderbecq. Originally published in JMIR Medical Education (https://mededu.jmir.org), 02.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX