BMC Medical Education

RESEARCH

Open Access

LLM-based automatic short answer grading in undergraduate medical education

Christian Grévisse[1*]

Abstract

Background Multiple choice questions are heavily used in medical education assessments, but rely on recognition instead of knowledge recall. However, grading open questions is a time-intensive task for teachers. Automatic short answer grading (ASAG) has tried to fill this gap, and with the recent advent of Large Language Models (LLM), this branch has seen a new momentum.

Methods We graded 2288 student answers from 12 undergraduate medical education courses in 3 languages using GPT-4 and Gemini 1.0 Pro.

Results GPT-4 proposed significantly lower grades than the human evaluator, but reached low rates of false positives. The grades of Gemini 1.0 Pro were not significantly different from the teachers'. Both LLMs reached a moderate agreement with human grades, and a high precision for GPT-4 among answers considered fully correct. A consistent grading behavior could be determined for high-quality keys. A weak correlation was found wrt. the length or language of student answers. There is a risk of bias if the LLM knows the human grade a priori.

Conclusions LLM-based ASAG applied to medical education still requires human oversight, but time can be spared on the edge cases, allowing teachers to focus on the middle ones. For Bachelor-level medical education questions, the training knowledge of LLMs seems to be sufficient, fine-tuning is thus not necessary.

Keywords Automatic short answer grading, Medical education, Large language models, GPT-4, Gemini

Background

Educational assessment is of utmost importance in teaching and learning. Depending on the target competencies, different forms of assessment are leveraged. For the lower levels of Bloom's taxonomy [3], multiple choice questions (MCQs) are a heavily used question type in exams, e.g., in medical education [26]. Nonetheless, MCQs rely on *recognition* of knowledge, and do not require *recall* or even *creation* as compared to natural language-based questions such as fill-in-the-gap (cloze) questions, short answer questions or longer essays [4]. However, the grading of such questions is time-intensive for teachers, which has a negative impact on scalability [8] and can push teachers toward avoiding them in favor of MCQs [7]. The time-intensive character of manual grading can lead to fatigue, so the order in which exam copies are graded can have an impact on the grade [4]. Graders can also be biased, even more in non-anonymous grading settings. These factors lead to inconsistencies and hamper the *reliability*. In the case of multiple graders, *inter-rater agreement* needs to be reached, which, due to the subjective nature of grading [8], is not always possible. Finally, in-depth personalized feedback is chronophagous and hence infeasible.

Automatic short answer grading (ASAG) focusses on *"assessing short natural language responses to objective questions using computational methods"* [4]. ASAG tries

*Correspondence:
Christian Grévisse
christian.grevisse@uni.lu
[1] Department of Life Sciences and Medicine, University of Luxembourg, 6, avenue de la Fonte, L-4364 Esch-sur-Alzette, Luxembourg

to tackle several of the aforementioned issues of human graders. Generally, literature on ASAG considers a short answer between one phrase and one paragraph long [4], whereas very short answers (between one and five words) are also gaining importance in medical education [2]. Advantages of ASAG are manyfold, for both teachers and students. In formative assessment settings, automatic grading gives an immediate feedback to students, a clear benefit of *intelligent tutoring systems* or other e-learning systems [6]. Teachers are assisted through valuable insights which allow them to take instructional decisions, e.g., when recognizing common misconceptions [14]. The time saved by automatic or - at least - assisted grading could be used for, e.g., personalized guidance and support [1].

The history of ASAG dates back to the 1960s [4]. The beginning was marked by rule-based methods, including concept mapping (i.e., detecting the presence or absence of concepts in student answers) and information extraction (essentially pattern matching, e.g., using regular expressions or parse trees). Later approaches focussed on statistical methods, including corpus-based methods (i.e., statistical properties in large document corpora) and machine learning (Natural Language Processing (NLP) used in classification or regression tasks). Considerable work was conducted regarding evaluation, by comparing different methods through public data sets and competitions [4].

With the recent advent of Large Language Models (LLMs), a new branch of ASAG approaches is coming up in literature [5, 10, 12, 14, 16, 17, 20–23]. LLM-based ASAG methods should enhance fairness, improve efficiency while maintaining equal or higher accuracy compared to human graders [8]. Nonetheless, there is a set of caveats to take into account, such as [5, 8]:

- **Consistency** An LLM should give the same grade to a student answer across multiple independent runs. The underlying model should be able or instructed to reproduce the same output in a deterministic way.
- **Biases** The data with which an LLM was trained or fine-tuned could present inherent biases and be reflected in inference.
- **Knowledge availability** Specialized knowledge from a certain domain might not have been included in the original training data. Depending on the knowledge cut-off date, recent knowledge might also not be present.
- **Hallucinations** In the context of LLMs, a hallucination is an erroneous generation [15], often due to missing knowledge. The LLM could be tempted to fill the gap of missing knowledge by inventing facts and thereby draw wrong conclusions on the grade.
- **Prompt injections** If a student is aware of an LLM grading the exam questions, there is a risk of exploits through tailored indications in the answer. Prompt injections can be defined as the *"action of inserting malicious text with the goal of misaligning an LLM"* [19] and manipulating its output [25] while circumventing content restrictions and filters [11].
- **Transparency** The LLM should be able to consistently justify why a certain grade was determined for a student answer. Given the black-box nature of the underlying neural networks and the risk of hallucinations, this is a non-trivial challenge [12, 14].
- **Privacy** Student answers need to be anonymized, at least if commercial LLMs like GPT or Gemini are used. Open source LLMs such as Vicuna, Alpaca or Llama locally deployed could reduce this issue [10]
- **Readiness and accessibility of technology** As often with new technology, it may not necessarily be available to all educational institutions due to a limited budget or lack of personnel.

Contemporary literature on and a set of tools aiming at LLM-based ASAG is presented in Appendix A. To the best of our knowledge, no study so far focussed on LLM-based ASAG applied to medical education.

In this article, we graded student answers to short open questions from 12 undergraduate medical courses in 3 different languages using GPT-4 and Gemini 1.0 Pro and compared the resulting grades to those given by human evaluators. The research questions we address are the following:

- **RQ1** Is there a significant difference in human grading vs. LLM grading?
- **RQ2** Can the grade proposed by an LLM be influenced by providing it with the human grade a priori?
- **RQ3** Is LLM-based grading consistent?
- **RQ4** Do the language or length of an answer have an impact on the LLM-based grading?

The study contributes to the ASAG field in multiple aspects:

- To the best of our knowledge, this is the first study analysing the capabilities of LLM-based ASAG in the context of medical education. A positive outcome could significantly improve the efficiency of medical teachers in grading tasks.
- To the best of our knowledge, this is the first study using the recent Gemini LLM for ASAG.

Methods

We collected a total of 2288 student answers from 82 questions in 12 undergraduate medical education courses at the University of Luxembourg. Each question was answered by a median of 30 students over a period of 2 years between Summer Term 2021/2022 and Winter Term 2023/2024. For the given period, there would have been a total of 196 questions, but we only included those for which an evaluation rubric or sample solution was provided by the author of the question. A distribution of the number of questions and answers per course is given in Fig. 1. The majority of questions and answers stem from the biopathology and oncology courses.

The questions were asked in summative assessments conducted on Moodle. The answers were evaluated by one human grader (typically the author) and attributed a numerical grade between 0 and max. 10, depending on the weight of the question. Answers were indeed short, with a median length of 190 characters.

We developed a small Django web application into which questions and answers were imported and analyzed. Student data was anonymized.

Given the trilingual setting of the Bachelor in Medicine at the University of Luxembourg, 49 questions were asked in French, the remainder (33) in English. Answers were given in French (62%), English (37%) and a minority (1%) in German (Fig. 2). In almost 90%, answers were given in the same language as the question, but sometimes, students feel more comfortable in a different one, which usually is not an issue for teachers. Language was automatically detected using the `lingua` and `langdetect` Python packages. Inconsistencies were resolved manually.

Answers were graded using two LLMs, namely GPT-4 (`GPT4`) and Gemini 1.0 Pro (`Gemini`) through the

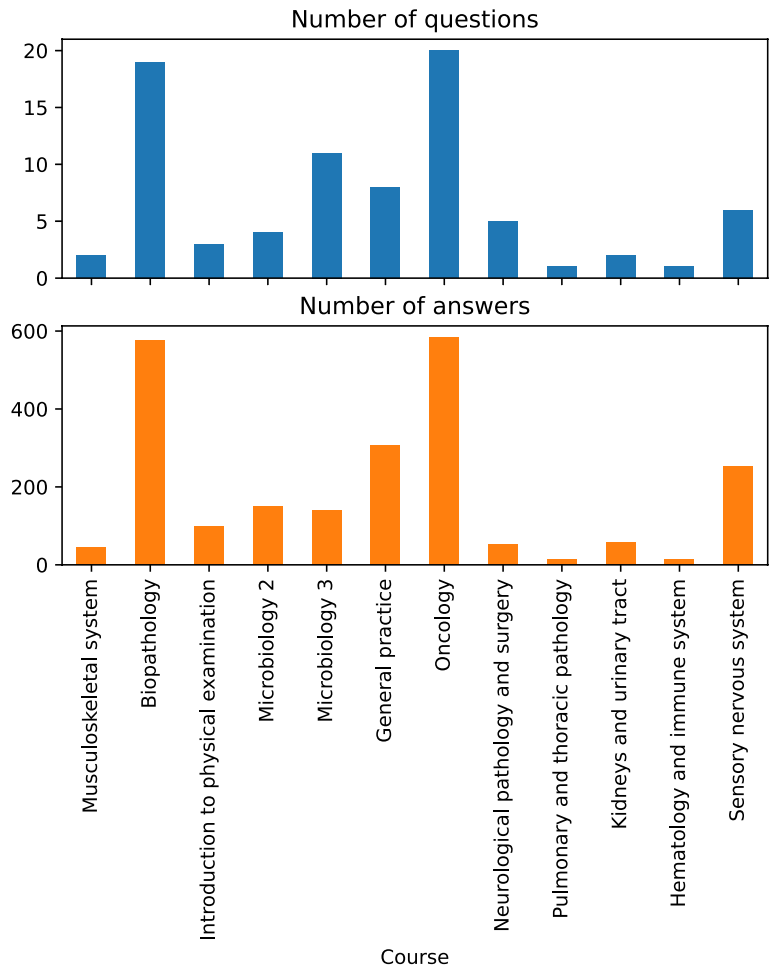


Fig. 1 Number of questions per course

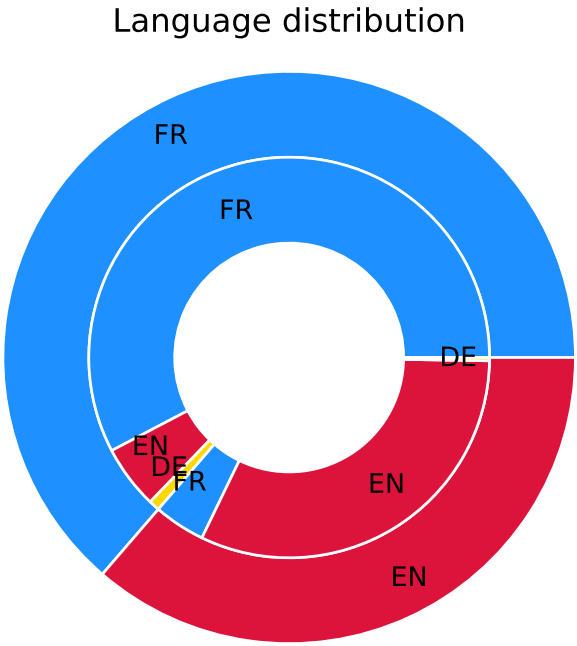


Fig. 2 Distribution of languages. The outer ring shows the language of the question, whereas the inner shows the language of the student answer

`openai` and `vertexai` Python packages respectively. Default *hyperparameters* were used, including the *temperature*[1] which indicates whether responses are more consistent and deterministic or more diverse and creative. Given that grading should appreciate the correctness of an answer compared to a sample solution by a numerical value, the expected output should not depend too much on the temperature. For `gpt-4-0613`[2], temperature ranges from 0.0 too 2.0, where 0.0 is most consistent and 2.0 is most diverse. The default value of 1.0 was used. For `gemini-1.0-pro-001`[3], temperature ranges from 0.0 to 1.0, the default of 0.9 was used. Experiments were conducted in late February 2024. Analysis was done using the Python libraries `pandas`[4], `SciPy`[5] and `scikit-learn`[6]. Visualization was created using the Python libraries `Matplotlib`[7] and `seaborn`[8]. The usual significance level $\alpha = 0.05$ was employed.

[1] https://platform.openai.com/docs/guides/text-generation/how-should-i-set-the-temperature-parameter

[2] https://platform.openai.com/docs/api-reference/chat/create#chat-create-temperature

[3] https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/adjust-parameter-values#temperature

[4] https://pandas.pydata.org

[5] https://scipy.org

[6] https://scikit-learn.org

[7] https://matplotlib.org

[8] https://seaborn.pydata.org

The prompts are shown in Fig. 3. GPT-4 distinguishes between *system* messages, which set the context and the behavior, and *user* messages, which contain the actual requests. At the time of writing, Gemini does not draw this distinction. Both LLMs are provided the question stem, the key (rubric or sample solution), the student answer and the maximum number of points to be attributed.

In addition to `GPT4`, we also included a third automatic grader `GPT4b`, where `b` stands for *biased*. The red text in Fig. 3, is only present in `GPT4b`, and provides the LLM with the grade of the human evaluator. This is done in order to respond to **RQ2**.

Results

Comparison among graders

For comparability reasons, all following calculations used normalized grades by dividing the attributed grade by the maximum grade of a question, resulting in a grade between 0 and 1. Human grades for all 2288 questions had a mean of 0.68($\pm$0.34) and a median of 0.75, indicating generally rather high grades.

We first wanted to check whether there are some significant differences between the human grader (`Human`) and the three LLM-based graders `GPT4`, `GPT4b` and `Gemini`. As shown in Table 1, there were significant differences among all graders except between `Human` and `Gemini`. Indeed, as can be seen in Fig. 4, both `GPT4` and `GPT4b` had globally a lower mean normalized grade (0.65 ± 0.29 respectively 0.64 ± 0.31) than the human grader. `Gemini` was comparable to the human grader, with a mean of 0.68 ± 0.32 and a median of 0.75.

On a per-course level, as shown in Fig. 5, this general trend is sometimes broken. In the hematology and immune system course, `GPT4` and `GPT4b` were less severe than the `Human` grader, whereas `Gemini` was giving lower grades than the teacher. In the pulmonary and thoracic pathology course, `GPT4` and `Gemini` were both less severe than `Human`.

In Fig. 6, *Bland-Altman* plots show the agreement between `Human` and LLM graders. For all 3 LLM graders, the mean difference compared to the `Human` grader is very close to 0, indicating a low systematic bias. The *limits of agreement* represent 95% of all differences: Their extension is lowest for `GPT4b` and highest for `Gemini`. The patterns indicate a tendency towards agreeing more for completely wrong or completely correct answers. Disagreement varies among partially correct answers, contributing to the extension of the limits of agreement, specifically for `Gemini`. As overplotting might introduce a distortion, we will later analyze the distribution of the absolute relative difference.

System
You are a medical teacher that needs to grade questions. Compare the answer given by the student to the key and return a grade between 0 and max. <Points> points. Decimal grades are allowed. Only return the grade, nothing else!

User
Question: <Stem>

User
Key: <Key>

User
Student's answer: <Answer>

User
What grade do you give, knowing that a human grader gave <Points>?

Fig. 3 Prompts to grade student answers

Table 1 Paired t-test between the different graders

	GPT4		GPT4b		Gemini	
	p-value	95% CI	*p*-value	95% CI	*p*-value	95% CI
Human	$\mathbf{1.22 \cdot 10^{-12}}$	(0.029, 0.052)	$\mathbf{7.13 \cdot 10^{-43}}$	(0.046, 0.061)	0.32	(–0.007, 0.020)
GPT4			$\mathbf{1.71 \cdot 10^{-4}}$	(0.006, 0.019)	$\mathbf{1.92 \cdot 10^{-10}}$	(–0.044, –0.023)
GPT4b					$\mathbf{3.42 \cdot 10^{-16}}$	(–0.057, –0.035)

Correlation between grade and answer language or length
We wanted to see whether the length or language of an answer had an impact on the grade. We used the Pearson Correlation Coefficient r for calculating the correlation between answer length and grade. For the categorical variable of the answer language, we calculated the correlation ratio η. The results for all 4 graders are shown in Table 2. All values show a weak correlation between the grade and the answer length respectively language.

Accuracy
For teachers, it could be interesting to rely on an LLM-based ASAG assistant if it can reliably grade "extreme" answers, i.e. fully correct or fully wrong, such that teachers could focus on partially correct answers. To determine this reliability, we can calculate the accuracy by categorizing answers based on their grade.

We first defined binary accuracy by categorizing answers into fully correct (1) or not fully correct (0,

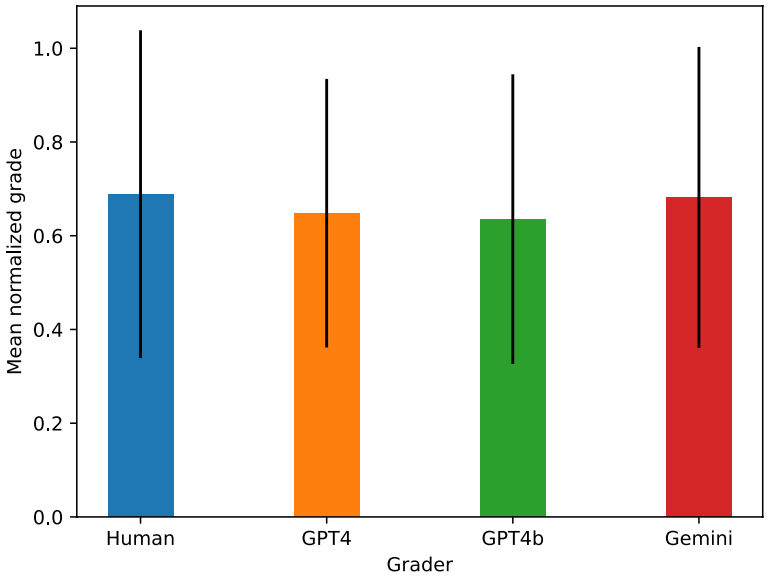


Fig. 4 Global comparison of normalized grades among graders

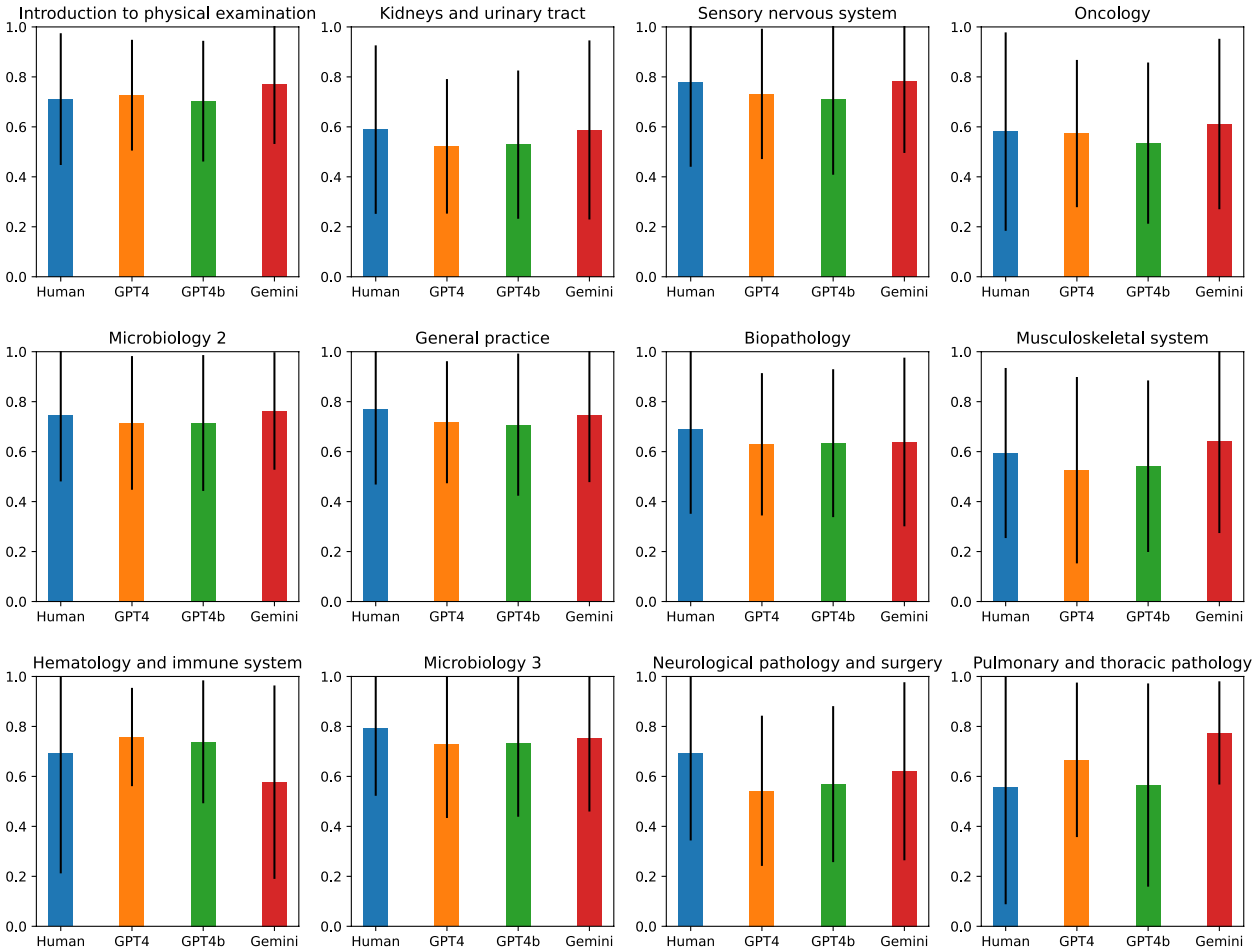


Fig. 5 Per-course comparison of normalized grades among graders

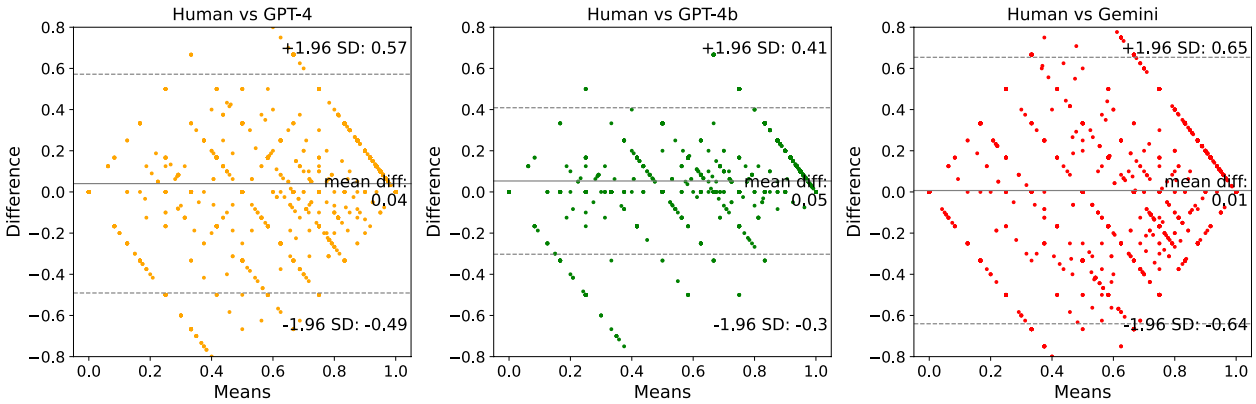
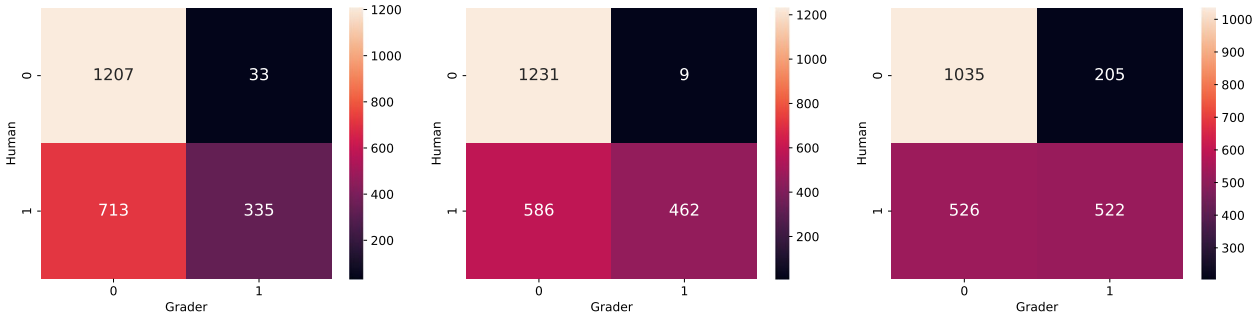


Fig. 6 Bland-Altman plots showing the agreement between `Human` and LLM graders

Table 2 Correlation between grade and answer length/language

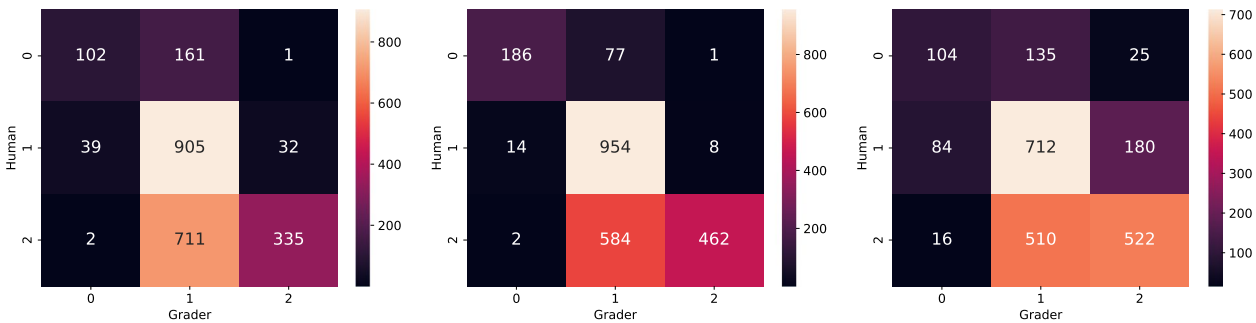
Grader	r_{Length}	$\eta^2_{Language}$
`Human`	0.046	0.017
`GPT4`	0.218	0.023
`GPT4b`	0.134	0.021
`Gemini`	0.112	0.039

including fully wrong and partially correct) and comparing the three LLM graders to the human grader. Heatmaps and accuracy values are shown in Fig. 7. `GPT4b` reaches the highest accuracy, whereas `GPT4` and `Gemini` have a comparable accuracy. `GPT4` and `GPT4b` have few false positives (i.e., the grader wrongly attributing full points), but all three LLM graders suffer from a



(a) `GPT4` ($acc = 0.67$) (b) `GPT4b` ($acc = 0.74$) (c) `Gemini` ($acc = 0.68$)

Fig. 7 Heatmaps showing the binary accuracy of the 3 LLM graders. The fairer the color, the higher the number



(a) `GPT4` ($acc = 0.59$) (b) `GPT4b` ($acc = 0.70$) (c) `Gemini` ($acc = 0.58$)

Fig. 8 Heatmaps showing the ternary accuracy of the 3 LLM graders. The fairer the color, the higher the number

high number of false negatives (i.e., the grader wrongly deducting points).

Ternary accuracy could be defined by categorizing answers into fully wrong (0), partially correct (1) and fully correct (2). Heatmaps and accuracy values are shown in Fig. 8. Again, `GPT4b` has the highest accuracy, the other two LLM graders are comparable. All three give partial grades for answers that were marked as fully wrong or fully correct by the human grader, which would mean that verification by a human grader is still required.

Fully correct or incorrect answers

For those questions that were considered fully correct by the human evaluator, we calculated the *precision*, i.e., the number of answers considered correct by both the LLM and the human grader divided by the number of all answers considered correct by the LLM. Again, `GPT4b` reached the highest precision (0.98), followed by `GPT4` (0.91) and `Gemini` (0.72). The high precision indicates that `GPT4` could be used as a relatively reliable grader for detecting fully correct solutions.

For answers considered fully wrong by the human evaluator, LLM graders were more reluctant in giving no points, as the median relative grade for `GPT4` and `Gemini` was 0.33.

Differences and regrading

We were interested in the absolute relative difference Δ between the LLM graders and the human evaluator. Figure 9 shows that the distribution for all 3 LLM graders is left-skewed, indicating generally low differences, with a median of 0.17 for `GPT4` and `Gemini` and a median of 0.0 for `GPT4b`.

For answers where $\Delta > 0.5$, we asked the corresponding LLM to regrade the answer. As shown in Fig. 10, the prompt indicated the previous grade of the LLM and the human grade, and asked the LLM to explain its divergence compared to the human grade or if it changed opinion (and then eventually indicate a new grade).

Tables 3 and 4 show the outcome of this regrading exercise. Over 200 answers were regraded where `Gemini` initially differed over 50% with the human grade, and over 100 for `GPT4`. Note that there was no significant difference between the human and original LLM grade for the subset of answers where $\Delta > 0.5$ and a new grade was determined. While the new grades also did not significantly differ from the original LLM grade, in the case of `Gemini`, they were significantly higher than the original grades.

This can also be appreciated in Fig. 11. Both `GPT4` and `GPT4b` gave a higher grade after regrading to half of the answers, while lowering it for a third of them. `Gemini` changed to a higher or lower grade for a quarter of the answers respectively.

Explanations

We used the LLMs to summarize their explanations why they changed or maintained their grade compared to the human one (Fig. 12). For `GPT4b`, we additionally indicated that it knew the human grade and asked whether it considers to have been biased.

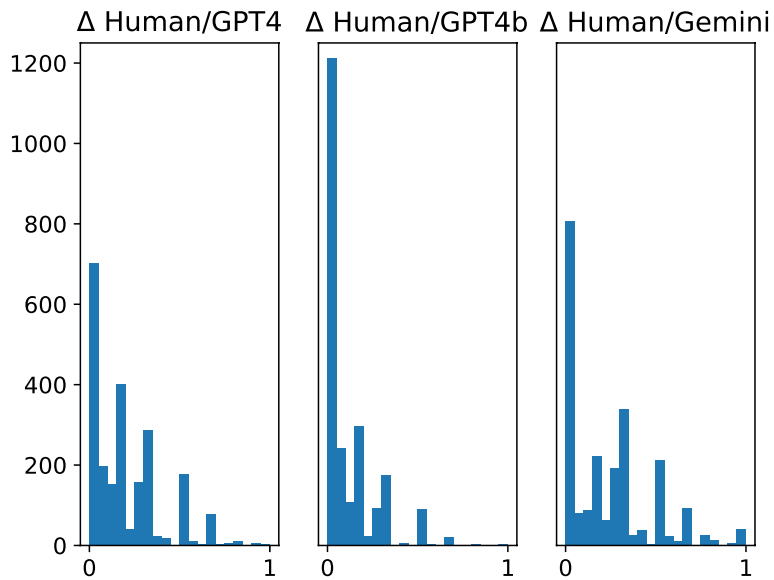


Fig. 9 Distribution of absolute relative difference Δ between human and LLM-based grades

System
You are a medical teacher who previously graded the following exam question, based on the given key and student's answer.

User
Question: `<Stem>`

User
Key: `<Key>`

User
Student's answer: `<Answer>`

User
You graded this answer `<Points>` out of `<Points>` points, whereas a human teacher gave `<Points>`.

User
Return a JSON object where you give an explanation (key: 'explanation') for the difference in grade (justify why you think your grade is correct or the human grader's is incorrect, or indicate whether you changed your opinion). Optionally, in a separate key 'new_grade', indicate a new grade if you changed your opinion. If so, only put the new grade in the value, nothing else! If you stick with your previous grade, omit this key!"

Fig. 10 Prompt asking the LLM to explain the high divergence from the human grade

Table 3 Number of regraded answers

	GPT4	GPT4b	Gemini
Answers regraded	122	36	215
Answers with new different grade	100	29	109

All 3 LLM graders were more lenient in their grading when the essence of the key was captured in the student's answer, this to *"encourage foundational understanding over rote memorization"* (GPT4) and to *"acknowledge the effort"* (GPT4b). GPT4b mentioned that it *"tended to*

Table 4 Paired t-tests between human, original LLM and new LLM grades

	GPT4		GPT4b		Gemini	
	p-value	95% CI	*p*-value	95% CI	*p*-value	95% CI
Human vs. Original LLM grade	0.985	(–0.143, 0.140)	0.293	(–0.135, 0.433)	0.964	(–0.151, 0.144)
Original LLM grade vs. new LLM grade	0.068	(–0.162, 0.006)	0.202	(–0.223, 0.049)	0.307	(–0.187, 0.059)
Human vs. new LLM grade	0.111	(–0.177, 0.019)	0.605	(–0.181, 0.305)	**0.020**	(–0.124, –0.011)

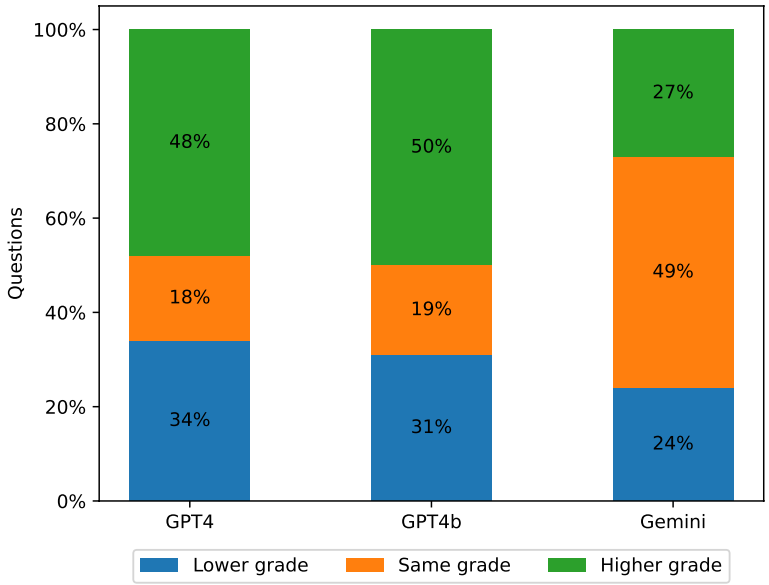


Fig. 11 Evolution of grade after regrading

System

You have previously graded exam questions and compared your grades to those of a human medical teacher. You knew the human teacher's grade at the time of grading. You have given explanations why you have given a higher, lower or the same grade. You will be given these explanations, separated for each answer by the delimiter ///.

User

The explanations are: `<Explanations>`

User

Based on these explanations, summarize why you gave more, less or the same grade than the human grader. Also, indicate whether you think to have been biased in your decision by knowing the grade of the human teacher a priori.

Fig. 12 Prompt to GPT-4 summarizing the explanations

give the benefit of the doubt". They did critize the human graders:

> *"The variance in grades between myself and the human teacher often stemmed from* **differing interpretations of the breadth and depth of knowledge** *demonstrated by the students and the educational objectives of emphasizing either broad understanding or detailed specificity."* (`GPT4b`)

> *"In several instances, I identified a partial understanding or effort from the student that* **the human grader might have overlooked** *, leading to a more lenient grading on my part."* (GPT4b)

> *"In cases where the student's answer matched the key or demonstrated understanding consistent with the key but was graded* **harshly** *by the human grader."* (Gemini)

More strict grading was explained by significant misunderstanding or inaccuracies *"where precision is crucial"* (e.g., treatment) or when the answer was too broad (GPT4). Gemini further stated to have given lower grades if these inaccuracies *"detracted from the overall correctness or completeness as per the criteria laid out in the key"*. In some cases, Gemini also acknowledged its own error, in an almost human way:

> *"Occurred in situations where,* **after further reflection or additional scrutiny** *of the student's response in comparison to the key, it was evident that* **the human grader's assessment more precisely reflected** *the accuracy or completeness of the student's answer according to the provided grading criteria."* (Gemini)

Finally, regarding a possible bias, GPT4b gave a twofold answer:

> *"Reflecting on my decision-making process, there might have been a* **bias towards ensuring students received recognition for their efforts and partial understanding** *. This inclination towards acknowledging any correct elements in an answer, even if not fully detailed or precise, might have led to* **slightly more generous grading in some cases** *compared to the human teacher. Furthermore, my grading may also reflect a bias toward facilitating learning through* **positive reinforcement** *, particularly where foundational knowledge was demonstrated but lacked depth."* (GPT4b)

> *"* **Knowing the grade of the human teacher a priori might have influenced my grading decisions to some extent** *. This prior knowledge could have set anchoring points for reconsideration, potentially biasing my evaluations towards justifying or questioning the human teacher's judgments.* **My reflections often considered the human teacher's perspective, indicating a level of bias in attempting to align with or understand their grading rationale.** *While efforts were made to objectively assess each student's understanding based on the explanations provided, the insight into the human teacher's grading likely influenced the degree of leniency or strictness applied in my reevaluations. This influence is particularly apparent in cases where adjustments were made upon reflecting on the human teacher's perspective, suggesting a* **predisposition to seek justification for their grades before finalizing my own.** *"* (GPT4b)

Variability

Given that LLMs were up to changing their initial grade when asked to regrade, we wanted to analyze the general variability. For this, we repeated the prompt from Fig. 3 on all 373 answers where $\Delta > 0.5$ 9 further times to have a total of 10 independent grades.

For each answer, we calculated the standard deviation of these 10 normalized grades, called henceforth σ_{10}. The highest average standard deviation was reached by Gemini ($\sigma_{10} = 0.15$), the lowest by GPT4 ($\sigma_{10} = 0.08$). There was no significant difference between the original grade and the average of the 9 additional ones. The third quartile $Q3$ of σ_{10} is 0.17, which indicates a low overall variability. The maximum $\sigma_{10} = 0.44$ was attained by Gemini, which was responsible for 90% of the answers with σ_{10} in the upper quartile range.

Two answers where Gemini reached $\sigma_{10} > 0.4$ are particularly interesting here. The first question asked for alarm symptoms and the key contained a list of possible symptoms to mention. The student's answer was quite long, with many symptoms but not all tailored to alarm symptoms. The second question asked for three causes of fibrinous pericarditis. The student answered with an empty enumeration (i.e., only writing "1. ... 2. ... 3. ..."). For both answers, Gemini fluctuated between 0 and full points. GPT4 and GPT4b agreed with the human grader.

There were 6 questions having at least 4 answers with σ_{10} in the upper quartile range, resulting in a set of 27 answers. In 23 cases, again Gemini was the variable grader. Most often, the related questions were asking for a number of items (e.g., *"Name 3 ..."*) and the key indicated a larger number of possible correct answers. Gemini sometimes considered that the student needed to give all the answers. The LLM was also quite variable in its grading when the key was relatively short, leaving room for interpretation. Language differences between question and answers were only present in one case and hence did not play a role.

Grading of sample solutions

Finally, we were interested in seeing how GPT4 would grade the sample solution. The prompt used for this is shown in Fig. 13. To not introduce any potential bias

System
You are a medical teacher that needs to grade questions. Return a grade between 0 and max. `<Points>` points. Decimal grades are allowed. Only return the grade, nothing else!

User
Question: `<Stem>`

User
Student's answer: `<Answer>`

User
What grade do you give?

Fig. 13 Prompt asking the LLM to grade the sample solution

in favor of the solution, the LLM was told that it was a student answer. Lacking a key to compare against in this case, the LLM needed to rely on its training knowledge.

We calculated the relative difference between the attributed grade and the maximum number of points. The third quartile of this relative difference was 0.0, indicating a high agreement between the grade and the number of points. On a per course-level, the highest mean difference could be noted in the neuropathology course (0.15). Here again, the issue was caused by questions whose key was quite concise. Indeed, there is a weak negative correlation ($r = -0.11$) between the relative difference and the length of the key. This means that a shorter key may lead to a lower grade. Keys had a median length of 192 characters, the longest key had 920 characters.

Discussion

Overall, grades by GPT-4 were significantly lower than the grades from the human evaluator. This is different from the results in [5], where GPT would assign higher scores than human evaluators, and only become more severe in one-shot settings. Tobler also noted a high agreement with manual grading and sometimes a stricter evaluation by GPT-4 [23]. Gemini 1.0 Pro did not present a significant difference compared to human grades.

Binary accuracy was higher than ternary, which was also observed in [10]. As opposed to [10], though, we found low rates of false positives with GPT-4, and GPT-4 achieved a high precision for fully correct answers. Nevertheless, there remained questions where LLMs were reluctant to give 0 respectively full points, so human oversight is still necessary [21]. But given that recognition of fully correct answers is reliable, more time will be available to focus on the middle cases. This aligns well with Kortemeyer [13], who observed that GPT *"performs best at the extreme ends of the grading spectrum"*, and with Schneider et al., who stated that GPT has a tendency to the middle [21].

Regarding **RQ1**, we thus acknowledge a significant difference for GPT-4, but with a low rate of false positives, and a non-significant difference for Gemini.

When asking **GPT4b** about a potential bias directly, it could not exclude it. This bias could also have led it to reach the highest accuracy in both binary and ternary settings. Of course, the usage of **GPT4b** is only realistic if the LLM can be provided a human grade and would issue a second opinion. Regarding **RQ2**, we cannot exclude a potential bias. Hence, prompt design should take this into account.

Hackl et al. stated that *"inconsistent ratings could lead to unfair outcomes"* [12]. In our case, variability was not significant, but the highest was observed with Gemini,

which sometimes fluctuated between 0 and full points. This behavior could be explained by certain characteristics of the key, if it was either too short or enumerated many possible answers in a question asking for a number of items. This issue could be solved by providing more explicit keys respectively adapting the prompt in case of enumeration questions. In [12], the authors even executed multiple iterations over different time frames, and still reached a high inter-rater reliability with GPT-4. Whether the temperature hyperparameter has an influence on the variability of the grade needs to be further analysed, although the diversity in the output should not have an impact on judging factual knowledge, as LLMs typically do not provide false information on purpose unless otherwise instructed. Furthermore, as the variability was not noted among the whole corpus of answers, the temperature would at least not be the only influencing factor. With respect to **RQ3**, overall we can say that LLM-based grading is consistent for high quality keys.

We detected a weak correlation between LLM-based grades and the length or language of an answer. Interestingly, Chang and Ginter found a negative association between answer length and model performance [5], while Schneider et al. found that grading criteria are language sensitive [21]. Regarding **RQ4**, we conclude that - in our setting - language or length of an answer did not have an impact on the LLM-based grading.

Assessing the instructor's answers allows to see whether they are too tightly coupled to the learning material [21]. In our case, the keys reached full points in 3 out of 4 cases, which also shows that, at least for Bachelor-level medical education questions, the training knowledge is sufficient and does not require model fine-tuning.

The explanations given by the LLM graders extended from giving the *"benefit of the doubt"* to sanctioning inaccuracies *"where precision is crucial"*, which is indeed the case for medical education. Condor proposes to use ASAG as an assistant to rather than a replacement of the human grader, even more in high-stakes scenarios [7]. Failing students merely based on a grade provided by an LLM raises ethical questions [21, 23]. Indeed, if a low agreement among the automated model and the human rater would be found, a third human opinion should be asked [7]. Employing the LLM itself as a second opinion to a human grader can be interesting, e.g., if anonymous grading cannot be applied [23], but it could also give a false sense of security [21].

Matelsky et al. observed that it was also *"important to consider how students perceive AI graders"* [16]. Schultze et al. recently studied *algorithm aversion*, i.e., whether students appreciate LLM-augmented feedback [22]. They conducted an experiment with undergraduate psychology students in the UK. The results showed no algorithm aversion, and the quality of feedback by GPT-3.5 was rated higher than the feedback from a human teacher, as it used more positive phrasing, was more detailed and showed better readability.

Coming back to the caveats mentioned at the beginning, the LLM graders were quite consistent, even though some variability could be noted depending on the quality of the key. The knowledge availability can be determined based on the knowledge cutoff date. For the employed LLMs, `GPT-4` includes training data up to September 2021, whereas `Gemini`'s cutoff is November 2023. With respect to the knowledge assessed in an undergraduate medical education program, these cutoff dates seem fairly reasonable. Biases in the underlying training data should be monitored for, e.g., by using LLM evaluation frameworks like *DeepEval*[9]. Transparency can be enabled by requesting explanations on the grade, ideally at the same time as the grade is provided to avoid a misalignment. It cannot be excluded that LLMs could have hallucinated while providing the explanations of the grades, albeit their summary seems overall rather plausible. Prompt injections would most likely just be a risk if students would assume that answers were only graded by an LLM, as such a kind of manipulation in the answer would seem strange to a human grader. We do stress the need of human supervision. Nonetheless, there exist technical solutions to try to detect prompt injection attacks, e.g., *Rebuff AI*[10]. Privacy can be ensured by not providing personal data to the LLM and/or using LLMs deployed on-premise. Finally, accessibility to technology being often a question of budget or personnel depends on the very institution.

What are the practical implications of our findings? For **medical educators**, while LLMs were reluctant to give 0 points, GPT-4 achieved a high precision of fully correct answers, which means that educators can save time respectively can dedicate more time to non-edge cases. The trade-off, though, is to focus on creating high quality keys or grading rubrics which can be used by the LLM. At least for questions and answers in French and English, and for Bachelor-level knowledge, our approach is promising, which means that open questions could be more frequently used in medical education exams. As students tend to perform better in multiple-choice questions due to recognition of the correct answer [2], a higher usage of short answer questions facilitated by LLM-based automatic grading support could better reflect real student performance. For **medical students**,

[9] https://github.com/confident-ai/deepeval

[10] https://github.com/protectai/rebuff

LLM-based feedback could be valuable in formative assessment, where the stakes of incorrect grading is not so high. However, feedback should be monitored, as hallucinations could foster misconceptions. On the other hand, misconceptions could also be identified through students' answers [18]. It was reported that students did not present algorithm aversion and rather appreciated the feedback from an LLM [22]. However, they could develop algorithm aversion in case the LLM consistently attributed lower grades than a human grader in summative assessments. Finally, for **medical schools**, applying the described approach would incur costs, either for infrastructure or AI APIs. Also, quality should be closely monitored, as students could complain about unfair or incorrect grading.

The main limitation of this study is that LLM grades were only compared to a single human grader. Even if the graders were all domain experts and mostly authors of the questions they graded, another grader could have been more lenient or severe. Also, answers were graded in an isolated way, i.e., knowledge across questions was not taken into account by the LLMs, but might have been appreciated by a human grader. The temporal evolution of LLMs can also have an impact on the consistency of grading, and should thus be verified periodically [12].

Conclusions

In this article, we analyzed the capabilities of LLM-based ASAG in the context of medical education in a multilingual setting. GPT-4 showing low rates of false positives and a high precision among fully correct answers can spare time to teachers who may focus on the middle cases. In our setting, as the grades were overall rather high, this could lead to a substantial time saving among medical teachers, which in our case are mostly active practitioners. To enable consistent gradings, teachers would need to invest a bit of time in writing high quality sample solutions or rubrics, but the return on this investment is time spared during grading. Overall, this could mean a non-negligible boost in efficiency to medical teachers.

The training knowledge seems to be sufficient for Bachelor-level medical education questions, so no fine-tuning would be needed. Human oversight is still necessary, even more for high-stakes exams, e.g., as a second or third opinion in addition to human graders. For formative settings, automated feedback to medical students would be feasible as well.

For future work, we support the idea of Schneider et al. to use LLMs for assessing open questions before an exam and check whether their rubric or sample solution is too tightly coupled with the learning material [21]. One LLM could respond to the question, while another instance would grade this answer based on the key. Different answer styles could be produced by adjusting the temperature hyperparameter. Such an approach was also used in [6] to augment datasets and reduce overfitting during training. Finally, while the results of Gemini 1.0 Pro in this study are already promising, we would like to further investigate the behavior of Gemini 1.5. The recently released Llama 3.1 and GPT4o will also be included.

Appendix A: Related work and tools

Contemporary literature on LLM-based ASAG approaches has covered different fields such as mechanical engineering [10] or data/computer science [21]. The first transformer-based ASAG works we found stem from 2020 [7, 9]. Condor trained Google's BERT (Bidirectional Encoder Representations from Transformers) on expert ratings [7]. Gaddipati et al. compared ASAG performance of ELMo, BERT, GPT-1 and GPT-2 on the Mohler dataset, which includes questions and answers from the computer science domain [9]. In that time, according to their findings, ELMo outperformed the other transformers. Chang and Ginter used ChatGPT to grade 2000 student answers in Finnish from ten undergraduate courses [5]. No grading criteria, such as a rubric or the instructor's solution, were provided, but sample student answers were used with the corresponding grade. GPT-4 outperformed GPT-3.5, a common finding in contemporary LLM research. Gao et al. used Vicuna for ASAG in mechanical engineering courses [10]. Hackl et al. used GPT-4 to evaluate macroeconomics essays on both content and style [12]. Regarding the content, a sample solution was provided. The LLM tended to give higher scores for content than for style, which can be explained by the high linguistic quality reached and hence expected by GPT-4. Okgetheng and Takeuchi graded 300 Japanese essays using the Open-Calm LLM family [17]. Pinto et al. used GPT-4 to grade open-ended questions in job training at a software development company [20]. Schneider et al. assessed the instructor's answer, a student's answer in general as well as a student's answer with respect to the instructor's answer through GPT 3.5. The answers were both in German and English, situated in the data science and computer science domains [21]. Xiao et al. used GPT-4 and a fine-tuned GPT-3.5 for Automated Essay Scoring (AES) in second-language learners courses [24]. The aim was to assist novice graders. The result was that novice graders with LLM feedback reach an accuracy comparable to expert graders, while the performance and consistency of expert graders got boosted.

There is also a set of tools aiming at LLM-based ASAG. Matelsky et al. developed *FreeText*, a web application (alternatively also a widget for Jupyter notebooks) that provides LLM-based feedback for open-ended responses

[16]. Under the hood, it uses *Guidance*[11], a library by Microsoft that enables handling agnostic of the underlying LLM. Tobler developed *SmartGrading*, a GPT-4-based ASAG web application [23]. The *AI Text* quiz question type plugin[12] for the Moodle Learning Management System[13] by Marcus Green extends the default essay question type by giving automatic feedback and marking based on GPT.

Authors' contributions
The sole author realized the study conception, literature search, data analysis and manuscript writing.

Funding
No funding was received.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 19 May 2024 Accepted: 13 September 2024
Published online: 27 September 2024

References

1. Adıgüzel T, Kaya MH, Cansu FK. Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. Contemp Educ Technol. 2023;15(3). https://doi.org/10.30935/cedtech/13152.
2. Bala L, Westacott RJ, Brown C, Sam AH. Twelve tips for introducing very short answer questions (VSAQs) into your medical curriculum. Med Teach. 2023;45(4):360–7. https://doi.org/10.1080/0142159X.2022.2093706.
3. Bloom BS. Taxonomy of educational objectives: The classification of educational goals. Boston: Allyn and Bacon; 1956.
4. Burrows S, Gurevych I, Stein B. The Eras and Trends of Automatic Short Answer Grading. Int J Artif Intell Educ. 2015;25(1):60–117. https://doi.org/10.1007/s40593-014-0026-8.
5. Chang LH, Ginter F. Automatic Short Answer Grading for Finnish with ChatGPT. Proc AAAI Conf Artif Intell. 2024;38(21):23173–81. https://doi.org/10.1609/aaai.v38i21.30363.
6. Cochran K, Cohn C, Rouet JF, Hastings P. Improving Automated Evaluation of Student Text Responses Using GPT-3.5 for Text Data Augmentation. In: Wang N, Rebolledo-Mendez G, Matsuda N, Santos OC, Dimitrova V, editors. Artificial Intelligence in Education. Cham: Springer Nature Switzerland; 2023. pp. 217–28. https://doi.org/10.1007/978-3-031-36272-9_18.
7. Condor A. Exploring Automatic Short Answer Grading as a Tool to Assist in Human Rating. In: Bittencourt II, Cukurova M, Muldner K, Luckin R, Millán E, editors. Artificial Intelligence in Education. Cham: Springer International Publishing; 2020. pp. 74–9. https://doi.org/10.1007/978-3-030-52240-7_14.
8. Fagbohun O, Iduwe N, Abdullahi M, Ifaturoti A, Nwanna O. Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices. J Artif Intell Mach Learn Data Sci. 2024;2(1):1–8. https://doi.org/10.51219/JAIMLD/oluwole-fagbohun/19.
9. Gaddipati SK, Nair D, Plöger PG. Comparative Evaluation of Pretrained Transfer Learning Models on Automatic Short Answer Grading. 2020. https://doi.org/10.48550/arXiv.2009.01303.
10. Gao R, Thomas N, Srinivasa A. Work in Progress: Large Language Model Based Automatic Grading Study. In: 2023 IEEE Frontiers in Education Conference (FIE). 2023. https://doi.org/10.1109/FIE58773.2023.10343006.
11. Greshake K, Abdelnabi S, Mishra S, Endres C, Holz T, Fritz M. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. AISec '23. New York: Association for Computing Machinery; 2023. pp. 79–90. https://doi.org/10.1145/3605764.3623985.
12. Hackl V, Müller AE, Granitzer M, Sailer M. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. Front Educ. 2023;8. https://doi.org/10.3389/feduc.2023.1272229.
13. Kortemeyer G. Toward AI grading of student problem solutions in introductory physics: A feasibility study. Phys Rev Phys Educ Res. 2023;19(2). https://doi.org/10.1103/physrevphyseducres.19.020163.
14. Latif E, Zhai X. Fine-tuning ChatGPT for automatic scoring. Comput Educ Artif Intell. 2024;6. https://doi.org/10.1016/j.caeai.2024.100210.
15. Masters K. Medical Teacher's first ChatGPT's referencing hallucinations: Lessons for editors, reviewers, and teachers. Med Teach. 2023;45(7):673–5. https://doi.org/10.1080/0142159X.2023.2208731.
16. Matelsky JK, Parodi F, Liu T, Lange RD, Kording KP. A large language model-assisted education tool to provide feedback on open-ended responses. 2023. https://doi.org/10.48550/arXiv.2308.02439.
17. Okgetheng B, Takeuchi K. Estimating Japanese Essay Grading Scores with Large Language Models. In: 30th Annual Conference of the Language Processing Society (NLP2024). Japan: The Association for Natural Language Processing; 2024. https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/B3-2.pdf
18. Olde Bekkink M, Donders ARTR, Kooloos JG, de Waal RMW, Ruiter DJ. Uncovering students' misconceptions by assessment of their written questions. BMC Med Educ. 2016;16(1):221. https://doi.org/10.1186/s12909-016-0739-5.
19. Perez F, Ribeiro I. Ignore Previous Prompt: Attack Techniques For Language Models. In: NeurIPS ML Safety Workshop. 2022. https://doi.org/10.48550/arXiv.2211.09527.
20. Pinto G, Cardoso-Pereira I, Monteiro D, Lucena D, Souza A, Gama K. Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT. In: Proceedings of the XXXVII Brazilian Symposium on Software Engineering. SBES '23. New York: Association for Computing Machinery; 2023. pp. 293–302. https://doi.org/10.1145/3613372.3614197.
21. Schneider J, Schenk B, Niklaus C, Vlachos M. Towards LLM-based Autograding for Short Textual Answers. 2023. https://doi.org/10.48550/arXiv.2309.11508.
22. Schultze T, Kumar VS, McKeown GJ, O'Connor PA, Rychlowska M, Sparemblek K. Using Large Language Models to Augment (Rather Than Replace) Human Feedback in Higher Education Improves Perceived Feedback Quality. 2024. https://doi.org/10.31234/osf.io/tvcag.
23. Tobler S. Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments. Methods X. 2024;12. https://doi.org/10.1016/j.mex.2023.102531.
24. Xiao C, Ma W, Xu SX, Zhang K, Wang Y, Fu Q. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. 2024. https://doi.org/10.48550/arXiv.2401.06431.

[11] https://github.com/guidance-ai/guidance

[12] https://github.com/marcusgreen/moodle-qtype_aitext

[13] https://moodle.org

25. Yip DW, Esmradi A, Chan CF. A Novel Evaluation Framework for Assessing Resilience Against Prompt Injection Attacks in Large Language Models. In: 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). 2023. https://doi.org/10.1109/CSDE59766.2023.10487667.
26. Zuckerman M, Flood R, Tan RJB, Kelp N, Ecker DJ, Menke J, et al. ChatGPT for assessment writing. Med Teach. 2023;45(11):1224–7. https://doi.org/10.1080/0142159X.2023.2249239.

Publisher's Note