

Using Large Language Models for Automated Grading of Student Writing about Science

Chris Impey

cimpey@as.arizona.edu

University of Arizona

Matthew Wenger

University of Arizona

Nikhil Garuda

University of Arizona

Shahriar Golchin

University of Arizona

Sarah Stamer


University of Arizona

Research Article

Keywords: student writing, science classes, online education, assessment, machine learning, large language models

Posted Date: February 22nd, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-3962175/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

A challenge in teaching large classes for formal or informal learners is assessing writing. As a result, most large classes, especially in science, use objective assessment tools like multiple choice quizzes. The rapid maturation of AI has created the possibility of using large language models (LLMs) to assess student writing. An experiment was carried out using GPT-3.5 and GPT-4 to see if machine learning methods based on LLMs can rival peer grading for reliability and automation in evaluating short writing assignments on topics in astronomy. The audience was lifelong learners in three massive open online courses (MOOCs) offered through Coursera. However, the results should also be applicable to non-science majors in university settings. The data was answers from 120 students on 12 questions across the three courses. The LLM was fed with total grades, model answers, and rubrics from an instructor for all three questions. In addition to seeing how reliably the LLMs reproduced instructor grades, the LLMs were asked to generate their own rubrics. Overall, the LLMs were more reliable than peer grading, both in the aggregate and by individual student, and they came much closer to the instructor grades for all three of the online courses. GPT-4 generally outperformed GPT-3.5. The implication is that LLMs can be used for automated, reliable, and scalable grading of student science writing.

Introduction

Artificial intelligence (AI) has had a profound effect on diverse fields (Ryan, 2023; Shamshiri et al., 2024; Thirananavukasaru et al., 2023). AI has also impacted education at every level (Zhang and Aslan, 2021). College leaders see both promise and peril in this disruptive technology. In one survey of college leaders' opinions about Generative AI, 78% agreed that the tools offer an opportunity to improve how colleges educate, operate, and conduct research, but 57% also agreed that the same tools pose a threat to how colleges educate, operate, and conduct research (Anft, 2023). Recently, large language models (LLMs) and tools such as ChatGPT (Ouyang et al., 2022) have shown a great potential to help students learn but have also led to concerns about plagiarism and a degradation in the ability of students to write and synthesize information (Grassini, 2023). A complex typology of AI's capabilities affects every aspect of education, from tutoring and assessment to the way institutions admit students and identify those who are at risk (Holmes and Tuomi, 2022). The literature discussing LLMs in the classroom has mostly focused on instructors and students using them to generate educational content (Kasneci et al., 2023). In 2023, Khan Academy and OpenAI announced a partnership using GPT-4 as a learning assistant tool to facilitate student learning (OpenAI 2023; Khan, 2023). Students will be able to ask questions about content as they would an instructor.

We embarked on a project to see if LLMs could be useful in massive open online courses, or MOOCs. MOOCs are typically free and are open to anyone in the world who has access to a computer and the internet. In their first ten years, MOOCs have grown to nearly 20,000 courses, offered by 950 universities, and serving 220 million students worldwide (Shah, 2021). MOOCs are of interest to researchers because they are an informal learning environment where people can learn about science without enrolling in a university class, particularly adult learners (Falk and Needham, 2013). Although MOOCs resemble formal classes, with video lectures, quizzes, and activities, the learning is a self-directed learning environment guided by individual needs and interests (Oakley and Sejnowski, 2019). Unlike in the college setting, learners do not get grades or transferable credit and the classes do not typically contribute to a degree program. MOOCs have an international audience that encompasses many developing countries so they can play an important role in the democratization of education (Impey, 2020). The current study builds on prior work examining peer review grading in MOOCs (Formanek et. al., 2017). This research, as well as work by others (Gamage, Staubitz, and Whiting, 2021) has shown that while participation in peer review is correlated with student engagement and course completion, the grades can be inconsistent and there are problems with reliability and validity. Although there are some opportunities to improve the peer grading process, LLMs may offer a solution that has yet to be explored.

AI techniques are beginning to be used in MOOCs since they can readily be scaled to many thousands of learners. A literature review in 2020 found twenty papers using AI for assessment of students (Sanchez-Prieto et al., 2020). Four analyzed student behaviors, six investigated student feelings or sentiments, and ten assessed student achievement through AI-based methods. Among the ten, the focus was grading multiple choice tests, lab exercises, concept maps, and short-answer questions. None investigated longer writing assignments, as we do in this work. Recently, GPT-3.5 (Ye et al. 2023) has been shown to have an accuracy of 65-95% for grading multiple choice tests across ten different science topics (Alseddiqi et al., 2023). This variability in performance can be attributed to potential data contamination issues in LLMs or their inherent probabilistic behavior during text generation (Golchin and Surdeanu, 2023a,b). One study has used LLMs to validate peer-assigned essay scores in a Coursera MOOC (Morris et al. 2023). With this work, we go further by using LLMs to grade essays and generate rubrics. To make our analysis more robust, we used GPT-4 as well as GPT-3.5 (Bojic, Kovacevic, and Cabarkapa, 2023). We wanted to address the following research questions: can the LLMs (1) generate a grade comparable to that of an instructor, (2) match or exceed the reliability of peer grading of student writing when given an instructor's model answer, and (3) create a grading rubric with similar quality to that of an instructor? This work provides proof of concept, and in the future, we will see if LLMs can reliably scale automated grading to many thousands of students in online classes.

Theoretical Background

The theoretical framework for this research has two components. The first relates to learner engagement. High quality educational experiences depend on pedagogy that will keep learners motivated and involved in their mastery of the material. This is far more difficult online than in a face-to-face class (Martin and Borup, 2022). MOOCs present a particular challenge, since there is little opportunity for direct, real-time interaction with the instructor. The voluntary nature of a MOOC and the lack of summative assessment also affect engagement. One study validated a MOOC engagement scale with the following dimensions: behavioral, cognitive, emotional, and social (Deng, Benckendorff, and Gannaway, 2020). Active learning in a MOOC has been shown to increase engagement and completion rates (Shah et al., 2022). We have used short writing assignments and peer review to increase learner engagement and have demonstrated that participation in writing substantially increases the probability that a learner will complete the course (Formanek et al., 2019).

The second component involves technology in the service of grading student writing. It is trivial to use algorithms and programs to grade objective assessments like multiple choice quizzes. However, automated grading of student writing is more difficult. Machine learning has advanced rapidly enough that it can be used to grade essays (Borad and Netak, 2021). In a recent study, the BERT language model was effective in evaluating writing based on grammar, semantics, coherence, and prompt relevance (Vanga et al., 2023). Another study using the RoBERTa language model shows that this language model could outperform human raters (Beseiso, Alzubi, and Rashaideh, 2021). We are investigating whether LLMs can surpass peer grading in reliability relative to an instructor. Instructor model answers are one input, but another is an instructor-generated rubric, since predicting rubric scores has been found to be essential to automated essay grading (Kumar and Boulanger, 2020). If LLMs can approach instructor reliability, they can be used in MOOCs with tens of thousands of learners, where grading by peers is a burden and grading by a human instructor is essentially impossible.

Data from the Astronomy MOOCs

Our education research group has been offering MOOCs through Coursera since 2014 (Impey, Wenger, and Austin, 2015). The MOOCs utilized in this study are on the topics of astronomy (Impey et al., 2016), astrobiology (Impey, Wenger, and Riabokin, 2023), and the history and philosophy of astronomy (Impey, 2023). Respectively, they are titled “Astronomy: Exploring Time and Space,” “Astrobiology: Exploring Other Worlds,” and “Knowing the Universe: History and Philosophy of Astronomy.” Together, they have enrolled 410,000 learners in 190 countries. A study of learner engagement and motivation shows that learners who attempt short writing assignments or activities like citizen science as part of the course are much more likely to complete the class than those who do not (Formanek et al., 2019). However, it is impractical for one instructor to grade thousands of writing assignments, so one option is peer grading, where every assignment is graded by three (or four) other students selected randomly, using a very simple rubric. This approach does increase learner motivation, but peer grading has limited reliability and validity (Formanek et al., 2017). We wanted to explore whether LLMs could be effective in grading student writing and in generating grading rubrics. For this purpose, we collected data from each of the Coursera MOOCs. Among the three courses, astronomy and astrobiology are the easiest for a human or a machine to evaluate because they are content based. The course on the history and philosophy of astronomy is more challenging because answers to some of the questions depend on speculation or hypothetical situations, where the judgement can be subjective (Golchin et al. 2024, Impey, 2023).

We gathered answers from 120 learners to each of 12 questions across the three courses. For the introductory astronomy and astrobiology courses, the students were asked to write a 250-to-750-word response, and for the history and philosophy class the instructions asked for 250-to-300 words. For the astronomy and astrobiology courses, the assignments were purposefully sampled to have a spread of peer review grades. For the history and philosophy course, which was launched recently and has a lower enrollment, fewer assignments were available, so the assignments were selected because they covered the selected topics, and a random sample was chosen from those assignments. Ground zero for the comparison was grading by one of the instructors according to rubrics he had created (Stevens and Levi, 2012; Pisano et al., 2021). The grading was made blind by having one of the authors provide writing assignments to the instructor without grade information. The instructor used the same rubrics that are available to the students and peer graders in the course. These same rubrics were provided to the LLMs for this experiment. Example or model answers were written by the instructor to represent a content expert’s version of an acceptable correct answer for each assignment.

Results from Large Language Models

Table 1 shows the average scores for each question in each course. Both LLMs are displayed here with each of the prompts, GPT-3.5 and GPT-4. One-sigma standard deviations for each average were determined from bootstrap resampling of the scores (Efron, 1979). To mimic the situation of a MOOC setting, 10,000 bootstrap samples were taken in each case. For each LLM, Table 1 has four lines. The first line is the average score determined by the LLM when provided with the instructor’s model answer. The second line is the average score from the LLM when provided with the instructor’s model answer and the rubric written by the instructor. The third line is the average score using a rubric created by the LLM based on the instructor’s model answer. The fourth line is the average score using only a rubric created by the LLM. The bootstrap standard deviation is zero for scores obtained via GPT-3.5 for Q2 on the astronomy course using the AI rubric only, and for Q1 on the history and philosophy course using the instructor’s answer and rubric. This is because all 10 students received identical scores for that prompt, resulting in zero standard deviation after the bootstrap resampling process.

Introduction to Astronomy

The class “Astronomy: Exploring Time and Space” comprises five writing questions, with a selection of ten students chosen randomly to represent a diverse scoring range. The first question carries a maximum score of 6 points, while questions 2 to 5 are scored out of 9 points. Across all the prompts, both LLMs consistently graded higher than the instructor, except in cases where GPT-3.5 is provided with the instructor’s model answer and rubric. Notably, both LLMs struggle to align with the instructor’s grades in the first question due to the open-endedness of it. On the second question, there is improved agreement with the instructor, as both LLMs tend to grade higher. On the third question, GPT-4 performs markedly better than GPT-3.5 in matching the instructor grades. Similarly, on the fourth question, GPT-4 performs slightly better than GPT-3.5. The fifth question has the lowest instructor scores and both LLMs grade substantially higher than the instructor, with GPT-4 coming closer than GPT-3.5. It is noteworthy that for the best-performing LLM, GPT-4, the results remain consistent whether the instructor’s answer is paired with an AI-generated rubric or with the instructor’s rubric.

Introduction to Astrobiology

The “Astrobiology: Exploring Other Worlds” course has three writing questions, and once again, we randomly selected ten students to sample the spread of better and worse answers. Each question is scored out of ten points. All the questions here allowing students to give both open-ended, allowing for a thorough exploration of the topics, and well-structured answers, presenting information in a logical and organized fashion. On the first question, GPT-4 consistently scores higher than the instructor for all prompts and GPT-3.5 has slightly better agreement. The best results were obtained when the LLM was provided with an instructor’s model answer and rubric. On the second question, the pattern is similar, with GPT-4 more lenient than GPT-3.5 in its scoring. For both LLMs, the results are within one sigma of the instructor’s scores, except when using the AI-generated rubric alone. On the third question, GPT-3.5 graded more strictly than the instructor while GPT-4 had excellent agreement with the instructor, except when the only prompt was the AI-generated rubric.

History and Philosophy of Astronomy

The “Knowing the Universe: History and Philosophy of Astronomy” course presents the greatest challenge for either an instructor or for an LLM, because the subject matter is very broad, and it involves conceptually challenging material. An answer might be based as much on plausible speculation as a consensus among experts. Students are therefore challenged to use higher order thinking skills than mere fact retrieval. Also, there were far fewer students and writing assignments to sample from. Instructor’s scores were lower as a percentage than for the other two courses. On the first question, the LLMs scored lower than the instructor by a substantial amount, with the best agreement for a prompt with the AI-generated rubric. On the second question, the agreement is generally better, and in this case the poorest agreement was with the AI-generated rubric. On the third question, the agreement is poor, with three of the prompts several standard deviations away from instructor scores. On the fourth question, the agreement is again poor, with the best result obtained using the instructor model answer along with an AI-generated rubric.

Table 1. Instructor and LLM Grades

Model	Prompt	Courses											
		Astronomy					Astrobiology			History and Philosophy			
		Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q1	Q2	Q3	Q4
	Instructor Grades	3.90 \pm 0.54	8.2 \pm 0.37	7.51 \pm 0.92	7.41 \pm 0.86	5.51 \pm 0.94	6.8 \pm 1.09	6.7 \pm 0.83	7.89 \pm 0.85	3.50 \pm 0.21	2.39 \pm 0.29	2.70 \pm 0.20	2.20 \pm 0.28
GPT-3.5	Instructor Provided Answer	4.50 \pm 0.16	7.60 \pm 0.29	6.81 \pm 0.75	6.81 \pm 0.78	6.41 \pm 0.78	7.41 \pm 0.84	7.01 \pm 0.76	6.70 \pm 0.71	2.70 \pm 0.14	2.90 \pm 0.22	2.69 \pm 0.32	3.00 \pm 0.14
	Instructor Provided Answer + Rubric	2.90 \pm 0.61	8.30 \pm 0.25	7.21 \pm 0.85	7.11 \pm 0.91	7.40 \pm 0.91	6.71 \pm 1.18	7.31 \pm 0.83	6.49 \pm 0.85	2.00 \pm 0.00	1.80 \pm 0.31	1.20 \pm 0.42	1.10 \pm 0.33
	AI Rubric + Instructor Answers	4.10 \pm 0.61	8.40 \pm 0.21	6.61 \pm 0.78	6.81 \pm 0.94	6.41 \pm 1.10	6.60 \pm 0.94	5.53 \pm 1.30	5.02 \pm 1.35	2.85 \pm 0.20	2.56 \pm 0.25	1.70 \pm 0.38	2.20 \pm 0.19
	AI Rubric Only	5.20 \pm 0.42	9.00 \pm 0.00	8.01 \pm 0.85	7.11 \pm 0.10	6.81 \pm 1.11	7.47 \pm 1.25	8.58 \pm 0.92	7.57 \pm 0.97	3.30 \pm 0.31	3.45 \pm 0.15	3.50 \pm 0.14	2.80 \pm 0.36
GPT-4	Instructor Provided Answer	4.75 \pm 0.41	8.65 \pm 0.20	7.61 \pm 0.87	7.51 \pm 0.77	6.21 \pm 0.90	7.50 \pm 0.70	7.91 \pm 0.68	8.10 \pm 0.33	3.50 \pm 0.21	3.25 \pm 0.23	3.65 \pm 0.14	3.2 \pm 0.24
	Instructor Provided Answer + Rubric	4.40 \pm 0.41	8.30 \pm 0.28	7.31 \pm 0.86	6.91 \pm 0.90	5.91 \pm 1.06	7.11 \pm 0.95	7.41 \pm 0.83	7.50 \pm 0.41	3.20 \pm 0.31	3.10 \pm 0.17	3.20 \pm 0.19	2.70 \pm 0.28
	AI Rubric + Instructor Answers	4.40 \pm 0.48	8.50 \pm 0.27	7.61 \pm 0.87	7.06 \pm 0.89	6.36 \pm 1.10	7.11 \pm 0.85	7.11 \pm 0.98	7.50 \pm 0.29	3.20 \pm 0.22	2.95 \pm 0.23	3.20 \pm 0.18	2.95 \pm 0.23
	AI Rubric Only	5.30 \pm 0.38	8.75 \pm 0.15	7.91 \pm 0.86	7.91 \pm 0.60	6.91 \pm 0.95	7.95 \pm 0.82	8.70 \pm 0.49	8.65 \pm 0.25	3.30 \pm 0.17	3.25 \pm 0.19	3.50 \pm 0.14	3.00 \pm 0.20

Performance of Large Language Models

The data in Table 1 can be used to draw inferences about the performance and reliability of the LLMs for grading student writing in an astronomy MOOC. Bootstrap resampling of the LLM scores yields standard deviations comparable to or lower than the standard deviations of scores from the instructor. For the astronomy course, across all the four prompts, this is true for 63% (12/19) of the questions using GPT-3.5 and 75% (15/20) of the questions using GPT-4. Next, considering the astrobiology course, this is true for 58% (7/12) of the questions using GPT-3.5 and 92% (11/12) of the questions using GPT-4. Finally, for the history and philosophy class, this is true for 56% (9/16) of the questions using GPT-3.5 and 88% (14/16) of the questions using GPT-4. Overall, these differences reflect lower internal variance when using GPT-4 than when using GPT-3.5.

Language Models versus Instructor

Next, we can look at the level of agreement between instructor scores and LLM scores. A clear pattern is seen from the data in Table 1, where across all prompts and questions, GPT-3.5 often scores the assignments higher than the instructor, 50% or 24/48 times, as it scores them lower than the instructor. On the other hand, GPT-4 generally scores assignments higher than the instructor, 83% or 40/48 times. In terms of differences between instructor scores and LLM scores relative to the LLM standard deviation from bootstrapping, other patterns are apparent. Table 2 summarizes the differences between instructor and LLM grades. The instructor grades are given in the top row of Table 2, as in Table 1. The subsequent rows treat instructor grades as “perfect” and list the number of standard deviations by which the LLM differs. The two cells with missing data are the cases where bootstrap resampling failed, as described earlier. In terms of root mean square (RMS) agreement of instructor between LLM scores, as anticipated from the comments above, GPT-3.5 performs worse with the history and philosophy class than the astronomy or astrobiology classes, and the same is true for GPT-4. There is little difference in the performance of either LLM between the astronomy class and the astrobiology class. The RMS results after averaging over all questions in each class are shown in Table 3. Two cells with numbers in parentheses reflect missing data for one question in the course. GPT-4 returns answers closer to instructor grades than GPT-3.5 for 3 out of the 4 prompts for the astronomy class and 3 out of the 4 prompts for the astrobiology class, but only 2 out of the 4 prompts for the history and philosophy class.

Table 2. Differences between Instructor and LLM Grades

Model	Prompt	Courses											
		Astronomy					Astrobiology			History and Philosophy			
		Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q1	Q2	Q3	Q4
	Instructor Grades	3.9	8.2	7.5	7.4	5.5	6.8	6.7	7.9	3.5	2.4	2.7	2.2
GPT-3.5	Instructor Answer	+4.0	-2.0	-1.0	-0.8	+1.2	+0.7	+0.4	-1.9	-5.5	+2.5	+0.0	+5.8
	Instructor Answer + Rubric	-1.6	+0.4	-0.3	-0.8	+1.2	+0.7	+0.4	-1.9	...	+2.5	+0.0	+5.8
	AI Rubric + Instructor Answer	+0.3	+1.0	-1.2	-0.7	+0.8	-0.2	-1.4	-2.2	-3.3	+0.3	-2.7	+0.0
	AI Rubric Only	+3.0	...	+0.6	-3.0	+1.2	+0.6	+1.0	-0.3	-0.7	+8.5	+5.5	+1.7
GPT-4	Instructor Answer	+2.1	+2.3	+0.1	+0.2	+0.8	+1.0	+1.7	+0.6	+0.0	+3.5	+6.1	+4.1
	Instructor Answer + Rubric	+1.2	+0.3	-0.3	-0.5	+0.4	+0.3	+0.8	-0.9	-1.0	+3.3	+2.5	+2.7
	AI Rubric + Instructor Answer	+1.0	+1.0	+0.1	-0.3	+0.9	+0.3	+0.4	-1.3	-1.4	+2.3	+2.7	+3.2
	AI Rubric Only	+3.4	+3.6	+0.5	+0.9	+1.4	+1.3	+2.0	+3.0	-1.1	+4.1	+5.4	+4.0

The best performance is observed for GPT-4 when the prompt includes both the instructor's model answer and their rubric, as well as when the prompt comprises the instructor's model answer plus an AI-generated rubric. In these

instances, scores for 7 out of the 8 questions in the astronomy and astrobiology courses fall within one standard deviation. Notably, the worst performance for GPT-4 is seen when the prompt is solely an AI-generated rubric, followed by the situation where the prompt includes only the instructor's model answer. However, across all four prompts with GPT-4 and spanning 12 assignments, the dispersion in the Language Model (LLM) grades is consistently less than or comparable to the dispersion in instructor grades. The agreement is expected to be closest to an instructor when using the 'Instructor + Rubric' prompt since it aligns closely with what the instructor will be grading. Incorporation of an AI-generated rubric has the potential to create more robust rubrics by leveraging correlations from instructor answers. It's important to note, however, that relying solely on rubrics may not be sufficient, and the presence of a master answer is crucial for ensuring greater reliability in the grading process.

Table 3. RMS Differences between Instructor and LLM Grades

Model	Prompt	Astronomy	Astrobiology	History and Philosophy
GPT-3.5	Instructor Grades	4.8	2.1	6.3
GPT-3.5	Instructor Answer	2.7	1.7	(4.7)
GPT-3.5	Instructor Answer + Rubric	1.9	2.5	6.3
GPT-3.5	AI Rubric + Instructor Answer	(4.6)	1.2	10.3
GPT-4	AI Rubric Only	3.2	2.1	8.1
GPT-4	Instructor Answer	1.4	1.2	5.0
GPT-4	Instructor Answer + Rubric	1.7	1.4	4.7
GPT-4	AI Rubric + Instructor Answer	5.2	3.8	7.9

Language Models versus Peer Grading

Peer assessment is often used in MOOCs and a major challenge is improving the reliability and consistency of this type of evaluation (Gamage, Staubitz, and Whiting, 2021). We analyzed peer grading for the astronomy course used here (Formanek et al., 2017). In that study, using data from 2015, peer review for 300 assignments out of 4 points had a mean score of 3.39, slightly lower than the instructor mean score, with a standard deviation of 0.78. In terms of reliability, unsurprisingly, instructor grades were the most reliable, followed by trained undergraduate graders, followed by peer graders. The correlation between the median instructor grades and the median peer grades was moderate ($r = 0.49$).

In this work, we compare instructor grades for a subset of the students in three courses with peer review grades for the entire course. Sample sizes are 19,661 for the astronomy course, 1705 for the astrobiology course, and 113 for the history and philosophy course. The numbers reflect the fact that the astronomy course started in 2015, the astrobiology course started in 2019, and the history and philosophy course started in 2022. Table 4 compares mean grades from the instructor and peer graders for the same questions as shown in Table 1. Uncertainties are one sigma standard deviations from the sample mean. Peer graders give higher scores and are more lenient than the instructor in almost every case, and the dispersion in their grades is higher than the instructor dispersion, by a factor of two for the astronomy and astrobiology courses, and by a factor of four for the history and philosophy course. We encountered some examples of student plagiarism in analyzing this data. Plagiarism and other forms of academic dishonesty are known to be an issue with some MOOCs (Surahman and Wang, 2021), but it does not affect the statistical results we report here.

Table 4. Instructor and Peer Grades

Source of Grades	Courses											
	Astronomy					Astrobiology			History and Philosophy			
	Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q1	Q2	Q3	Q4
Instructor Grades	3.90 \pm 0.54	8.2 \pm 0.37	7.51 \pm 0.92	7.41 \pm 0.86	5.51 \pm 0.94	6.8 \pm 1.09	6.7 \pm 0.83	7.89 \pm 0.85	3.50 \pm 0.21	2.39 \pm 0.29	2.70 \pm 0.20	2.20 \pm 0.28
Peer Grades	5.13 \pm 1.28	7.92 \pm 1.66	7.97 \pm 1.66	7.83 \pm 1.81	7.98 \pm 1.65	8.83 \pm 2.19	8.77 \pm 2.27	8.88 \pm 2.08	3.50 \pm 0.91	3.69 \pm 0.70	3.65 \pm 0.92	3.49 \pm 1.10

Summarizing the three-way comparison between instructor grades, peer review grades, and the grades assigned by the LLMs, both GPT-3.5 and GPT-4 generally produce grades that are closer to the instructor grades and have lower dispersion than peer grades. This is true for all three courses. Given the different scoring schemes for different assignments we use a score normalized to one. For the introductory course, the mean scatter in normalized scores is 15% between the instructor and GPT-4, as opposed to 27% between the instructor and peer graders. For the astrobiology course, the mean scatter in normalized scores is 17% between instructor and GPT-4, as opposed to 22% between instructor and peer graders. Last, for the history and philosophy course, the mean scatter in normalized scores is 26% between instructor and GPT-4, as opposed to 38% between instructor and peer graders. LLMs clearly have the potential to act as proxies for the instructor, avoiding some of the pitfalls and limitations of using novices (other students in the class) to grade student assignments.

Comparisons for Individual Students

The comparisons just described are averages across all ten students sampled. However, a student cares about their own grade more than the class average, so we also made a direct comparison of instructor, GPT-4, and peer graded scores on all twelve assignments for the ten students individually. In addition, we investigated the dispersion among the peer grades, where there were four for each assignment in the introductory course, and three for each assignment in the astrobiology and history and philosophy courses. Figure 1 is a scatterplot of the difference in grades between instructor and median for the peer grades on the x-axis and the difference in grades between the instructor and GPT-4 on the y-axis. Histograms are also shown.

For the introductory course, the mean instructor minus GPT-4 score is -0.01 ± 0.15 , compared to the mean instructor minus peer grading score of -0.07 ± 0.26 . The LLM grade is very close to the instructor grade, with a smaller dispersion than the peer grade. For the astrobiology course, the mean instructor minus GPT-4 score is -0.02 ± 0.17 , compared to the mean instructor minus peer grading score of -0.09 ± 0.20 . The LLM grade is very close to the instructor grade, with a slightly smaller dispersion than the peer grade. Last, for the history and philosophy course, the mean instructor minus GPT-4 score is -0.09 ± 0.24 , compared to the mean instructor minus peer grading score of -0.23 ± 0.30 . The LLM grade is much closer to the instructor grade, and has a slightly smaller dispersion, than the peer grade. LLM grades and peer grades are always more lenient than instructor grades. Overall, GPT-4 performs better than peer grading.

To further understand why there was a higher discrepancy amongst peer grades, we retrieved from Coursera the individual peer grades for each of the students to see the dispersion of peer grades for each question. We calculate a representative score by taking the average of the 10 students and converting them into percentages. Then we calculate the mean absolute deviation for each of the peer grades and then choose the highest and lowest out of the sample of three or four to see the spread in the percentages for each of the question. This is shown in Figure 2. The dispersion

ranges from 4% to 33% for the introductory course, from 0% to 26% for the astrobiology course, and from 0% to 20% for the history and philosophy course. The dispersion among peer grades for a particular assignment is substantial and it contributes to the dispersion in peer grades for the entire sample. This analysis of grades for individual students affirms that GPT-4 is superior to Coursera's peer grading mechanism, and it comes very close to matching the grades assigned by the instructor.

Discussion

The promise and peril of AI for education cannot be fully elucidated in a simple pilot study like this. However, the result of using LLMs to grade student writing assignments in these MOOCs are promising. Performance of the LLMs is excellent for the astronomy and astrobiology courses, and markedly worse in the history and philosophy class, where questions are more open-ended, and it is challenging even for an instructor to create a concrete rubric. To address the research questions posed earlier, it is possible for GPT-4 to score within one standard deviation of an instructor when the prompt includes the instructor's model answer and a rubric. In an astronomy MOOC where a direct comparison can be made, GPT-3.5 and GPT-4 both do as well as peer grading in matching the instructor grades with a small dispersion. When the LLM is used to create a rubric, that prompt alone does not give good results, but it does in combination with an instructor's model answer. Lastly, these AI methods can easily be scaled to evaluate the science writing of thousands or tens of thousands of online learners in real time. In this analysis, we have treated instructor grading as the "gold standard" and shown that GPT-4 comes close to the instructor score in three different MOOCs. Equipped with a rubric and a detailed explanation of the instructional goals for the assignment, it is likely that a future LLM could match the performance and reliability of a human instructor. In fact, since we have assumed instructors to be perfect, when in fact they are fallible, it's plausible that an LLM could one day eclipse the reliability of an instructor.

Results from this initial study have provided significant insights on the strengths and weaknesses of an LLM grading system. The writing assignments analyzed for this project were gathered from an existing course that was developed without any plan to grade assignments using an LLM. Our future work will explore the opportunities and challenges of applying these computer aided systems to assessing student writing assignments. Progress could be made in using LLMs to generate model answers and rubrics for open-ended questions and questions asking for speculation. Another approach involves the development of writing assignments and grading systems specifically designed to play to the strengths of the LLM. From an educational perspective, the kinds of student assignments that are best suited to LLM grading occupy the first five levels of Bloom's taxonomy (Krathwohl, 2002). They are: Remember, Understand, Apply, Analyze, and Evaluate. These levels all require factual knowledge, conceptual knowledge, and procedural knowledge, which can be drawn from an existing base of information on which the LLM can be or may have already been trained. Any assignments that fall in the sixth level (Create) will be more challenging for the LLM to assess, as was seen in the results from the Astrobiology class assignments. Current LLMs fall short of being able to evaluate creative assignments and any writing that requires metacognitive thinking.

This study dealt with science writing by lifelong, adult learners in MOOCs. A natural next step for this research is to apply it in the college classroom. College instructors agree that writing is an important tool for helping undergraduates learn science and apply the principles of scientific thinking (Moon, Gear, and Schultz, 2018). However, in the large introductory classes where most students get their only experience of science, the grading burden of evaluating student writing is severe. Peer grading can be used, but as in the MOOCs described here, undergraduates are not always reliable graders (Biango-Daniels and Sarvary, 2020). We plan to use LLMs to help instructors grade student writing in large General Education classes that satisfy the science requirement for graduation. Initially, it would be for formative assessment, where the LLM delivers a grade based on the instructor's model answer and rubric, as in this study. Beyond that, feedback could be provided using the claim, evidence, reasoning framework that is widely used in middle and high

school science classrooms, and recently at the college level (Eden, 2023). LLMs have recently been used for fact-checking and for identifying claim-evidence pairs in scientific content (Koneru, Wu, and Rajtmajer, 2023; Wang et al., 2023; Zeng and Zubiaga, 2024). The hope is that LLMs could provide instructors and their students with assessment of the scientific validity of student writing, aiming for the “gold standard” of conceptual learning (Gere et al., 2019).

Declarations

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

This work was supported in part by a grant from the National Science Foundation, award DUE-2020784.

Author Contribution

All authors contributed to the design and conception of the study. Evaluation material for the MOOCs was created by M.W., who also acted as the model instructor for this project. Implementation of the LLMs for grading student writing was done by N.G. and S.G. Gathering of the peer grading information was carried out by N.G. and S.S. The first draft of the manuscript was written entirely by C.I. All authors commented on successive versions of the manuscript and all authors read and approved the final manuscript.

Acknowledgements

We acknowledge fruitful conversations with Sanlyn Bunxer on the educational implications of this research, and with Alexander Danehy on the computational basis for the analysis and on the strengths and weaknesses of the current generation of large language models. This work was supported in part by a grant from the National Science Foundation, award DUE-2020784.

References

1. Alseddiqi, M., Al-Mofleh, A., Albalooshi, L, and Najam, O. (2023). Revolutionizing Online Learning: The Potential of ChatGPT in Massive Open Online Courses. *European Journal of Education and Pedagogy*, 4(4), 1-5. <https://doi.org/10.24018/ejedu.2023.4.4.686>
2. Anft, M. (2023). Perspectives on Generative AI: College Leaders Assess the Promise and the Threat of a Game-Changing Tool. *The Chronicle of Higher Education*, Washington, DC.
3. Beseiso, M., Alzubi, O.A., and Rashaideh, H. (2021). A Novel Automated Essay Scoring Approach for Reliable Higher Education Assessments. *Journal of Computing in Higher Education*, 33, 727-746.
4. Biango-Daniels, M., and Sarvary, M. (2020). A Challenge in Teaching Scientific Communication: Academic Experience Does Not Improve Undergraduates' Ability to Assess Their or Their Peers' Writing. *Assessment and Evaluation in Higher Education*, 46(5), 809-820.
5. Bojic, L., Kovacevic, P., and Cabarkapa, M. (2023). GPT-4 Surpassing Human Performance in Linguistic Pragmatics.

6. Borad, J.G., and Netak, L.D. (2021). Automated Grading of Essays: A Review. In: Singh, M., Kang, DK., Lee, JH., Tiwary, U.S., Singh, D., Chung, WY. (eds) *Intelligent Human Computer Interaction. IHCI 2020. Lecture Notes in Computer Science*, Vol. 12615. Springer, Cham. https://doi.org/10.1007/978-3-030-68449-5_25
7. Deng, R., Benckendorff, P., and Gannaway, B. (2020). Learner Engagement in MOOCs: Scale Development and Validation. *British Journal of Educational Technology*, 51(1), 245-262.
8. Eden, A. (2023). A Modified Claim, Evidence, Reasoning Organizer to Support Writing in the Science Classroom. *The American Biology Teacher*, 85(5), 289-291.
9. Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1-26
10. Falk, J. H., and Needham, M. D. (2013). Factors Contributing to Adult Knowledge of Science and Technology. *Journal of Research in Science Teaching*, 50(4), 431-452.
11. Formanek, M., Wenger, M., Buxner, S., Impey, C.D., and Sonam, T. (2017). Insights About Large-Scale Online Peer Assessment from an Analysis of an Astronomy MOOC. *Computers and Education*, 113, 243.
12. Formanek, M., Buxner, S., Impey, C., and Wenger, M. (2019). Relationship between Learners' Motivation and Course Engagement in an Astronomy Massive Open Online Course. *Physical Review Physics Education Research*, 15, 020140.
13. Gamage, D., Staubitz, T., and Whiting, M. (2021). Peer Assessment in MOOCs; Systematic Literature Review. *Distance Education*, 42(2), 268-289. <https://doi.org/10.1080/01587919.2021.1911626>
14. Gere, A.R., Limlamai, N., Wilson, E., Saylor, K.M., and Pugh, R. (2019). Writing and Conceptual Learning in Science: An Analysis of Assignments. *Written Communication*, 36(1), 99-135.
15. Golchin, S., Garuda, N., Impey, C., and Wenger, M. (2024). Large Language Models as MOOCs Graders. <https://arxiv.org/abs/2402.03776>
16. Golchin, S., and Surdeanu, M. (2023a). Time Travel in LLMs: Tracing Data Contamination in Large Language Models. <https://arxiv.org/abs/2308.08493>
17. Golchin, S., and Surdeanu, M. (2023b). Data Contamination Quiz: A Tool to Detect and Estimate Contamination in Large Language Models. <https://arxiv.org/abs/2311.06233>
18. Grassini, S. (2023). Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences*, 13(7), 692.
19. Holmes, W., and Tuomi, I. (2022). State of the Art and Practice in AI in Education. *European Journal of Education*, 57(4), 542-570.
20. Impey, C.D., Wenger, M.C., and Austin, C.L. (2015). Astronomy for Astronomical Numbers: A Worldwide Massive Open Online Class. *The International Review of Research in Open and Distributed Learning*, 16(1), 57-79.
21. Impey, C.D., Wenger, M., Formanek, M., and Buxner, S. (2016). Bringing the Universe to the World: Lessons Learned from a Massive Open Online Class on Astronomy. *Communicating Astronomy with the Public Journal*, 21, 20-30.
22. Impey, C.D. (2020). Higher Education Online and the Developing World. *Journal of Education and Human Development*, 9(2), 17-24.
23. Impey, C.D. (2023). Knowing the Universe: Teaching the History and Philosophy of Astronomy. *Astronomy Education Journal*. <https://doi.org/10.32374/AEJ.2023.3.1.058aep>
24. Impey, C.D., Wenger, M., and Riabokin, X. (2023). The Design and Delivery of an Astrobiology Massive Open Online Course. *Astrobiology*, 23(4), 460-468.
25. Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023). ChatGPT for good? On Opportunities and Challenges of

- Large Language Models for Education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>.
26. Khan, S. (2023). Harnessing GPT-4 so that all students benefit. A nonprofit approach for equal access. <https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/>. Accessed 7 Feb. 2024.
 27. Koneru, S., Wu, J, and Rajtmajer, S. (2023). Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences. <https://arxiv.org/abs/2309.06578>
 28. Krathwohl, D.R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice*, 41(4), 212-218.
 29. Kumar, V.S., and Boulanger, D. (2020). Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined? *International Journal of Artificial Intelligence in Education*, 31, 538-584.
 30. Martin, F., and Borup, J. (2022). Online Learner Engagement: Conceptual Definitions, Research Themes, and Supportive Practices. *Educational Psychologist*, 57(3), 162-177.
 31. Moon, A., Gear, A.R., and Schultz, G.V. (2018). Writing in the STEM Classroom: Faculty Conceptions of Writing and its Role in the Undergraduate Classroom. *Science Education*, 102(5), 1007-1028.
 32. Morris, W., Crossley, S. A., Holmes, L., and Trumbore, A. (2023). Using Transformer Language Models to Validate Peer-Assigned Essay Scores in Massive Open Online Courses (MOOCs). *LAK23: 13th International Learning Analytics and Knowledge Conference*, 315-323. <https://doi.org/10.1145/3576050.3576098>
 33. Oakley, B.A., and Sejnowski, T.J. (2019). What We Learned from Creating One of the World's Most Popular MOOCs. *MJP Science Learning*. <https://doi.org/10.1038/s41539-019-0046-0>
 34. OpenAI (2023). OpenAI Customer Stories: Khan Academy. <https://openai.com/customer-stories/khan-academy>. Accessed 7 Feb. 2024.
 35. Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. <https://arxiv.org/abs/2203.02155>
 36. Pisano, A., Crawford, A., Huffman, H., Graham, B., and Kelp, N. (2021). Development and Validation of a Universal Science Writing Rubric that is Applicable to Diverse Genres of Science Writing. *Journal of Microbiology and Biology Education*, 22:e00189-21. <https://doi.org/10.1128/jmbe.00189-21>
 37. Ryan, M. (2023). The Societal and Ethical Impacts of Artificial Intelligence in Agriculture: Mapping Agricultural AI Literature. *AI and Society*, 38, 2473-2485.
 38. Sánchez-Prieto, J. C., Gamazo, A., Cruz-Benito, J., Therón, R., and García-Peñalvo, F. J. (2020). AI-Driven Assessment of Students: Current Uses and Research Trends. In P. Zaphiris and A. Ioannou (Eds.), *Learning and Collaboration Technologies. Design, Experiences. 7th International Conference, LCT 2020, Copenhagen, Denmark*, Springer Nature, 1, 292-302. https://doi.org/10.1007/978-3-030-50513-4_22
 39. Shah, D. (2021). By the Numbers: MOOCs in 2021. Class Central. <https://www.classcentral.com/report/mooc-stats-2021/>
 40. Shah, V., Murthy, S., Warriem, J., Saharashbudhe, S., Banergee, G., and Iyer, S. (2022). Learner-centric MOOC Model: A Pedagogical Design Model Towards Active Learner Participation and Higher Completion Rates. *Educational Technology Research and Development*, 70, 263-288.
 41. Shamshiri, A., Ryu, K.R., and Park, J.Y. (2024). Text Mining and Natural Language Processing in Construction. *Automation in Construction*, 158. <https://doi.org/10.1016/j.autcon.2023.105200>
 42. Stevens, D.D., and Levi, A.J. (2012). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning* (2nd ed.). Routledge. <https://doi.org/10.4324/9781003445432>

43. Surahman, E., and Wang, T.-H. (2021). Academic Dishonesty and Trustworthy Assessment in Online Learning: A Systematic Literature Review. *Journal of Computer Assisted Learning*, 38, 1535-1553.
44. Thirunavukasar, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large Language Models in Medicine. *Nature Medicine*, 29, 1930-1940.
45. Vanga, R.R., Sindhu, C., Bharath, M.S., Reddy, T.C., and Kanneganti, M. (2023). Autograder: A Feature-Based Quantitative Essay Grading System Using BERT. In: Tuba, M., Akashe, S., Joshi, A. (eds) *ICT Infrastructure and Computing. ICT4SD 2023. Lecture Notes in Networks and Systems*, Vol. 754. Springer, Singapore.
https://doi.org/10.1007/978-981-99-4932-8_8
46. Wang, Y., Reddy, R.G., Mujahid, Z.M., Arora, A., Rubashevskii, A., Geng, J., Afzal, O.M., Pan, L., Borenstein, N., Pillai, A., Augenstein, I., Gurevych, Y., and Nakov, P. (2023). Factcheck-GPT: End-to-End Fine-Grained Document-Level Fact-Checking and Correction of LLM Output. <https://arxiv.org/abs/2311.09000>
47. Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhao, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., and Huang, X. (2023). A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models.
<https://arxiv.org/abs/2303.10420>
48. Zeng, X., and Zubiaga, A. (2024). MAPLE: Micro Analysis of Pairwise Language Evolution for Few-Shot Claim Verification. <https://arxiv.org/abs/2401.16282>
49. Zhang, K., and Aslan, A.B. (2021). AI Technologies for Education: Recent Research and Future Directions. *Computers and Education: Artificial Intelligence*, 2, 100025.

Figures

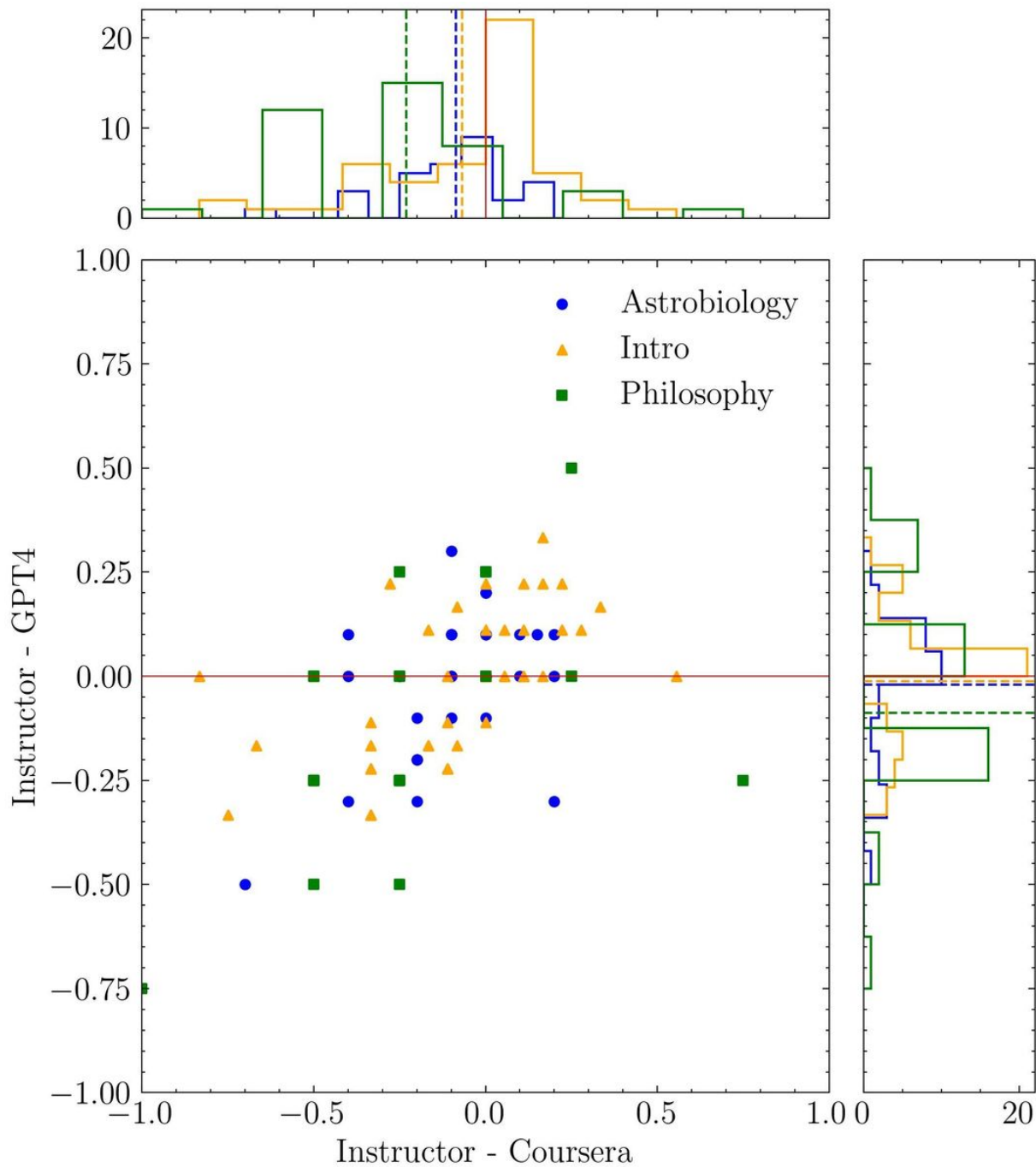


Figure 1

Scatter plot showing the difference of the grades by the instructor and the median of the peer scores received by Coursera for each student on the x-axis and the difference of the grades by the instructor and the LLM Model (GPT-4) on the y-axis. Symbols for the three courses are yellow triangles (Introductory Astronomy), blue circles (Astrobiology), and green squares (History and Philosophy of Astronomy). The two histograms show the distribution of the data points of these differences for each of the axes. Dashed lines in the histograms show the means of the three classes for the two measures of grade difference.

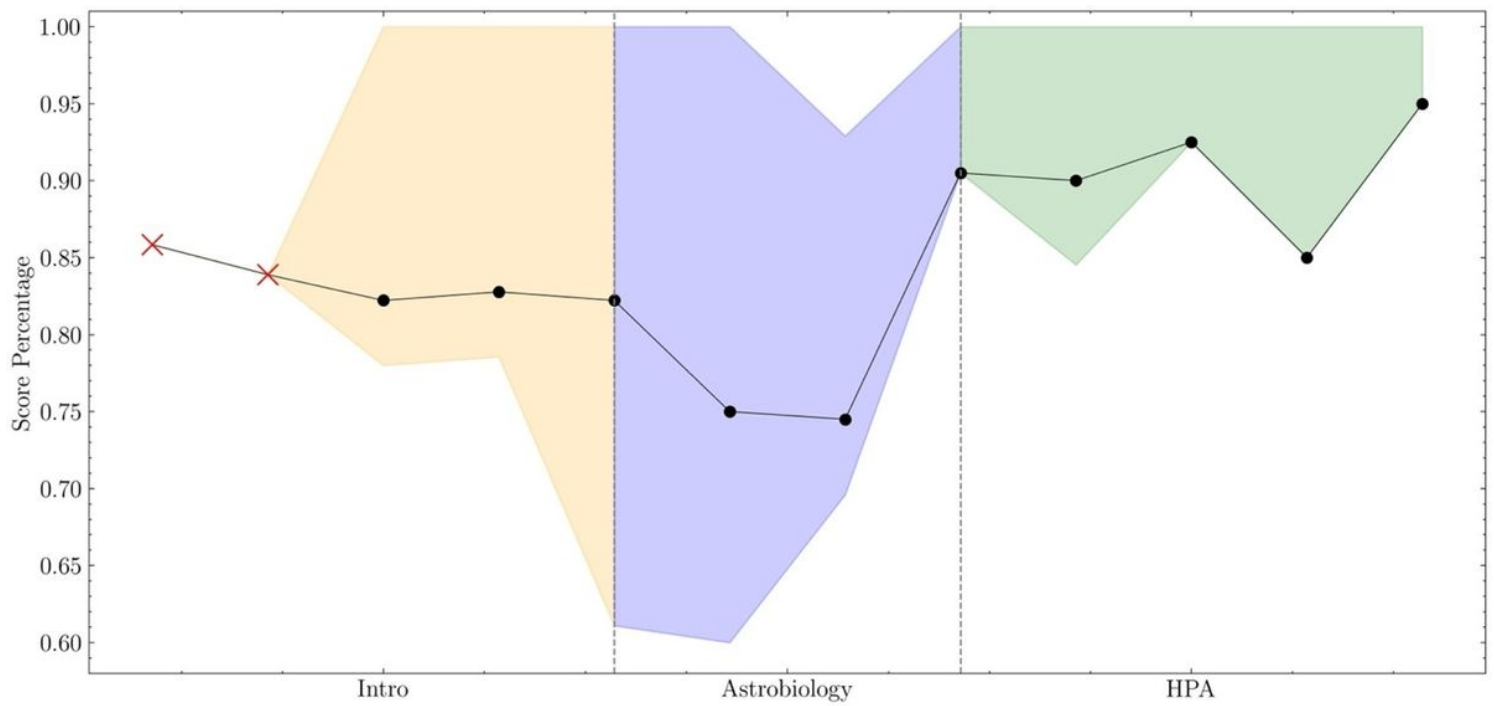


Figure 2

Dispersion of peer grades as a function of the questions in each of the three courses. The spread of the peer grades is shown as the colored shading (using the same color scheme as Fig. 1) and the points are the mean scores of the selected sample of students. Crosses in the plot show the mean scores but Coursera could not retrieve peer grade data for the first two assignments in the Introductory Astronomy course.