Grading Massive Open Online Courses Using Large Language Models

Shahriar Golchin[†*]**, Nikhil Garuda**[◇]**, Christopher Impey**[◇]**, Matthew Wenger**[◇]

[†]Department of Computer Science, University of Arizona
[◇]Department of Astronomy, University of Arizona
golchin@arizona.edu

Abstract

Massive open online courses (MOOCs) offer free education globally. Despite this democratization of learning, the massive enrollment in these courses makes it impractical for an instructor to assess every student's writing assignment. As a result, peer grading, often guided by a straightforward rubric, is the method of choice. While convenient, peer grading often falls short in terms of reliability and validity. In this study, we explore the feasibility of using large language models (LLMs) to *replace peer grading in MOOCs.* To this end, we adapt the zero-shot chain-of-thought (ZCoT) prompting technique to *automate the feedback process* once the LLM assigns a score to an assignment. Specifically, to instruct LLMs for grading, we use three distinct prompts based on ZCoT: (1) ZCoT with instructor-provided correct answers, (2) ZCoT with both instructor-provided correct answers and rubrics, and (3) ZCoT with instructor-provided correct answers and LLM-generated rubrics. We tested these prompts in 18 different scenarios using two LLMs—GPT-4 and GPT-3.5—across three MOOCs: Introductory Astronomy, Astrobiology, and the History and Philosophy of Astronomy. Our results show that ZCoT, when augmented with instructor-provided correct answers and rubrics, produces grades that are more aligned with those assigned by instructors compared to peer grading. Finally, our findings indicate a promising potential for automated grading systems in MOOCs, especially in subjects with well-defined rubrics, to improve the learning experience for millions of online learners worldwide.

1 Introduction

Massive open online courses (MOOCs)[1] have revolutionized access to free education by offering a wide array of free courses to anyone with an internet connection. This platform has democratized education and made it globally accessible, accommodating an unlimited number of participants (Impey et al., 2016, 2015). However, the resulting increase in participant numbers, though beneficial to a broader audience, introduces challenges in delivering personalized and constructive feedback on student performance, including the task of grading their assignments. To address this issue, peer grading has become a common practice. While peer grading enhances students' engagement and motivation, its reliability and validity are often questioned (Formanek et al., 2017).

In this study, we investigate the potential of LLMs in handling the task of grading assignments in MOOCs and the viability of replacing the current peer grading system with LLMs on this platform, aiming to establish more personalized and automated grading feedback to enhance the learning experience for online learners. Specifically, our methodology is grounded on employing the zero-shot chain-of-thought (ZCoT) prompting technique by asking the LLMs to grade assignments "step by step" (Kojima et al., 2022). Although this technique was originally designed to enhance the reasoning abilities of LLMs, we adapt it to generate feedback to students by having the LLMs verbalize the reasons behind score deductions. We also further refine it by providing more in-context information about the given task. In particular, we augment this technique in three ways: (1) ZCoT with instructor-provided correct answers, (2) ZCoT integrated with both instructor-provided correct answers and rubrics, and (3) ZCoT coupled with instructor-provided correct answers and LLM-generated rubrics based on the correct answers. We use these prompting strategies to evaluate the grading performance of two performant LLMs—GPT-4 (OpenAI, 2023) and GPT-3.5 (Ouyang et al., 2022)—in three MOOC subjects: Astrobiology (Impey, 2024a), Introductory Astronomy (Impey,

---

[*]Corresponding author.
[1]https://www.mooc.org/

2024b), and the History and Philosophy of Astronomy (Impey, 2024c). Our findings suggest that GPT-4 using ZCoT with instructor-provided answers and rubrics outperforms peer grading in terms of score alignment with those given by instructors. However, grading courses that require imaginative or speculative thinking, such as History and Philosophy of Astronomy, is found to be a challenge for both LLMs and peer grading. Despite this, GPT-4 still excels over peer grading by producing scores more closely aligned with those assigned by the instructors in such scenarios.

The key contributions of this paper are:

(1) We study the potential of replacing the peer grading system with the use of LLMs and adapting the ZCoT prompting technique in MOOCs, aiming to provide automated grading feedback that is more accurate and reliable than peer grading.

(2) Using a rigorous evaluation across 18 settings, our results reveal that when integrating ZCoT with instructor-provided correct answers and rubrics, along with GPT-4, it outperforms all other settings, including peer grading. This approach results in grades that are more closely aligned with those provided by instructors and offers more useful feedback than peer grading. However, the score alignment is stronger in courses that require less creative and imaginative thinking.

2 Related Work

LLMs in Education. The use of LLMs in education has primarily focused on how instructors and students leverage them to generate and respond to educational content (Kasneci et al., 2023; Alseddiqi et al., 2023). A key example is the implementation of GPT-4 at Khan Academy, where it acts as a learning assistant, answering subject-specific questions (OpenAI, 2024). Recent studies expanded on this, exploring LLMs' capabilities in providing feedback on open-ended questions and essays (Matelsky et al., 2023; Pinto et al., 2023; Meyer et al., 2024; Stahl et al., 2024), grading short-answer questions (Chang and Ginter, 2024), generating programming exercises (Del Carpio Gutierrez et al., 2024), and helping students with coding questions (Tripaldelli et al., 2024) or offering feedback on their code (Pankiewicz and Baker, 2023; Kiesler et al., 2023; Zhang et al., 2024; Gabbay and Cohen, 2024; Heickal and Lan, 2024; Jacobs and Jaschke, 2024). Before LLMs became widely used, research in this area relied on the encoder-only component of Transformer networks (Devlin et al., 2018; Vaswani et al., 2017) to evaluate assignments in short question-answering contexts (Morris et al., 2023; Funayama et al., 2022; Riordan et al., 2017).

LLMs as Judges. The most common use of LLMs is to approximate human judgment. Beyond grading, LLMs are used in various tasks requiring evaluative decisions. Among many examples, they were used to assess the quality of generated outputs in generation tasks such as summarization (Liu et al., 2023), evaluate translation quality (Kocmi and Federmann, 2023), and detect factual inconsistencies (Luo et al., 2023). LLMs were also applied to assess how well other LLMs follow instructions (Zeng et al., 2023), gauge the similarity between reference and LLM-generated texts (Golchin and Surdeanu, 2023b), and evaluate code (Zhuo, 2023). Building on this, fine-tuned LLMs as judges were proposed to enhance the performance of general-purpose LLMs (Zhu et al., 2023).

This research builds on earlier studies of peer grading in MOOCs, which revealed challenges in achieving consistency, reliability, and validity in peer-assigned grades (Formanek et al., 2017; Gamage et al., 2021). Our study explores the potential of using LLMs to replace peer grading in MOOCs, aiming to improve and automate the feedback and grading process.

3 Approach

3.1 Prompts

To guide LLMs in grading assignments, we use zero-shot chain-of-thought (ZCoT) (Kojima et al., 2022) as our base prompting technique. There are two main reasons behind this choice. First, we performed a pilot study and found that ZCoT yields scores more closely aligned with those given by instructors compared to vanilla zero-shot prompting. Second, ZCoT enables us to observe the reasoning behind each generated score, as the LLM is required to verbalize its reasoning process "step by step." This allows us to analyze the reliability and validity of the grading steps provided as feedback to students and proactively mitigate potential hallucinations during the grading process. Note that our use of ZCoT differs from its conventional application. Typically, ZCoT involves reasoning before making predictions. However, in our approach, the score (prediction) is generated first, followed by reasoning as feedback on any score deductions. Thus, we use ZCoT not to enhance performance,

but to automatically generate feedback.

We develop three different prompts, which incorporate ZCoT in combination with: (1) instructor-provided correct answers, (2) instructor-provided correct answers and rubrics, and (3) instructor-provided correct answers with LLM-generated rubrics. Each prompt is detailed in the following.

ZCoT with Instructor-Provided Correct Answers. Initially, we implement ZCoT by incorporating the correct answers, supplied by the course instructor, as in-context information to generate grades for assignments. Further, each student's assignment is embedded into the input prompt, and the LLM is tasked with grading the assignment based on the instructor-provided answer for each question. Note that, in this prompt, there is no guidance on how the grading should be carried out, except for the total points for each question. Finally, the grading is conducted for each student and question separately, ensuring that there is no influence from other questions or the grades of other students on the downstream grading task. Figure 1 shows the ZCoT with correct answers integrated.

ZCoT with Instructor-Provided Correct Answers and Rubrics. Our second prompting strategy enhances the previous approach by integrating the instructor-provided rubric for each question in addition to the correct answer. This method aims to align the LLM more closely with the instructor's grading criteria, ensuring the grades and score deductions match the instructor's standards. Retaining all components from the prior prompting technique, we follow the same grading procedure as in the previous stage. Figure 2 illustrates ZCoT with the inclusion of the correct answers and rubrics.

ZCoT with Instructor-Provided Correct Answers and LLM-Generated Rubrics. In this approach, we diverge from the two previous methods where the reference information in the prompt is centered around the human (instructor), and instead, include a portion of information generated by the LLM. This idea is inspired by the fact that LLMs can generate better rubrics compared to those provided by an instructor. The reasoning behind this is that LLMs are trained on an extensive amount of data, and thus, possess comprehensive interdisciplinary knowledge, such as knowledge from both education and astronomy domains in our case, that allows them to produce improved rubrics. To implement this, we prompt GPT-4 with the template prompt shown in Figure 3 to generate rubrics

based on the provided correct answer, total grade, and the question itself. This LLM-generated rubric is then replaced with the instructor-provided rubric used in the previous stage (Figure 2), and the same process is repeated to grade all assignments.

3.2 Evaluation of LLM-Assigned Grades

As noted, MOOCs cater to a large number of participants for each course, and our study aims to evaluate the effectiveness of LLMs in grading assignments for a substantial student population. However, to control the budget when using proprietary LLMs, and carefully analyze the grades assigned by LLMs to individual questions, we conduct the grading process on a manageable subset of the entire collection of students' assignments (see Section 4 under Data for more details). On the other hand, it is crucial to analyze the performance of LLMs in grading a significant volume of assignments, representative of the regular situations in MOOCs. To address this, we use the bootstrap resampling (Efron, 1979; Efron and Tibshirani, 1993).

Qualitatively, we compare LLM-assigned grades with those given by instructors (ground truth). Specifically, we examine the *absolute difference* in average grades per question for each course, considering both lenient and strict grading. The hypothesis we investigate here is that the difference between LLM-assigned and instructor-assigned grades should *not be statistically significant*, as determined through bootstrap resampling. In fact, this implies that the average grades given by both LLM and the instructor could be similar enough to be considered interchangeable. In a different analysis, we also use bootstrap resampling to evaluate average scores for potential bias in LLM grading. This helps us understand how the *overall average scores* from LLMs compare to those from instructors, providing insight into the reliability and consistency of LLM grading. In addition, we compute the mean absolute error (MAE) to measure the discrepancies between LLM- and instructor-assigned grades.

Quantitatively, we assess each question by comparing the LLM-assigned grades with the instructor-assigned grades using evaluations conducted by course instructors. A summary of this evaluation is then reported as our qualitative analysis.

3.3 Baseline

We analyze average scores and MAEs produced by the LLMs for each question across all courses, comparing them to the peer grading as the baseline

Figure 1: An illustration of the ZCoT prompt along with answers provided by the course instructor. Each question is assessed individually for every student. We repeat this process for all questions and students, incorporating their answers into the prompt, and instructing the LLM to grade the assignments. In this example, the instructor-assigned grade is 6/9, with GPT-4 serving as the underlying LLM.

and the instructor-assigned grades as the ground truth. By evaluating the average scores and MAEs with respect to the ground truth, our goal is to identify the most effective type of prompt and the corresponding LLM. In this context, a prompt, along with the respective LLM, is deemed superior if it demonstrates greater alignment with the instructor-assigned grades compared to those obtained through peer grading.

4 Experimental Setup

Data. This study employs data from three MOOCs: Astrobiology (Impey, 2024a), Introductory Astronomy (Impey, 2024b), and the History and Philosophy of Astronomy (Impey, 2024c). The process of obtaining the data followed a strict ethical roadmap, guided by an Institutional Review Board from the university. Participants voluntarily completed a survey and provided formal consent for their data to be used in the research. Also, all identifying information was anonymized. Since the original dataset had most assignments clustered at either very high or low grades, we sampled data to cover a wider range of grades for a more comprehensive evaluation of LLM grading performance.

Most students in these MOOCs are between the ages of 25 and 44. A slight majority of students are male across all courses, with the smallest gender gap in the Astrobiology course (45% female vs. 54% male) and the largest in the History and Philosophy course (34% female vs. 65% male). Geographically, most students are from the United States and India.

Regarding the attributes of the assignments, stu-

Figure 2: An illustration of ZCoT prompt that incorporates both instructor-provided correct answers and rubrics for grading assignments. The grading process used in ZCoT with correct answers only (Figure 1) is also applied here. Similar to Figure 1, the instructor-assigned grade for this question is 6/9, and GPT-4 is the base model. As shown, including rubrics in the prompt helps the LLM generate a grade consistent with the grade assigned by the instructor.

dents are advised to limit their answers to 500 words, although this restriction is not strictly enforced. Specifically, in the Introductory Astronomy course, there are five assignments, each worth 9 points, except for the first question, which is valued at 6 points. The Astrobiology course comprises three questions, each worth 10 points, while the Philosophy and History of Astronomy course includes four questions, each valued at 4 points.[2]

Following previous studies (Pankiewicz and Baker, 2023; Kiesler et al., 2023; Gabbay and Cohen, 2024, inter alia), we use a subset of data for our experiments. Specifically, we analyze 12 assignments from a group of 10 students in each of the three courses, resulting in analyzing 120 questions per prompt. Note that the data in this study is sourced from a proprietary repository, proactively accounting for the potential issue of data contami-

nation in LLMs (Golchin and Surdeanu, 2023b,a).

Setting. For all experiments with GPT-4 and GPT-3.5, we employ the `gpt-4-0613` and `gpt-3.5-turbo-0613` snapshots, accessed via the OpenAI API. We set the temperature to zero and the maximum token limit to 2048 to prompt LLMs. These hyperparameters remain consistent across all settings throughout our evaluations.

Instructor Grading. In our study, we regard the grades given by instructors (university professors) as our ground truth. These professors are experts in the relevant fields of each course and carefully assess the assignments in question. The grading process is double-blind, i.e., instructors are unaware of peer grading and students' information.

Peer Grading. Each writing assignment is evaluated by three to four randomly selected classmates using grading rubrics provided by the instructor,

---

[2] See Appendix A for comprehensive details on questions.

Instruction: Your task is to design a rubric that addresses the following questions/answers. The followings are the homework assignments for the Astrobiology course. This course is designed for undergraduate students majoring in Astronomy. The rubric must be unbiased and adaptable, capable of fairly evaluating any kind of student writing assignment.

The scoring breakdown for each question should be as follows: the score for Question 1 is 10, the score for Question 2 is 10, and the score for Question 3 is 10. Your thoughtful consideration in creating this rubric will ensure all students' work is evaluated equitably and consistently. Make sure that the rubric will provide points that are whole numbered.

Each rubric should be dedicated to each question separately.

– – –

Questions/Answers:

Question 1:

{As Figure 1.}

Answer 1:

{As Figure 1.}

Full Score: 10/10

Question 2:

Discuss how habitable zone range and spectral type are related. [...]

{The question is partially cut due to space constraints.}

Answer 2:

The habitable zone is defined as the range of distances from a star within which liquid water can exist on a planet. [...]

{The answer is partially cut due to space constraints.}

Full Score: 10/10

Question 3:

Discuss why/whether it is plausible that exobiology exists at all, in terms of exoplanet type, spectral type and orbital distance. [...]

{The question is partially cut due to space constraints.}

Answer 3:

The star around HD-127688 has a K spectral type. [...]

{The answer is partially cut due to space constraints.}

Full Score: 10/10

– – –

Rubric for Question 1:

- Identification of detection methods (2 points): The student correctly identifies the detection methods used for each exoplanet.

- Explanation of detection methods (2 points): The student provides a clear and accurate explanation of how each detection method works.

- Identification of physical characteristics (2 points): The student correctly identifies the physical characteristics that can be learned from each set of data.

- Explanation of physical characteristics (2 points): The student provides a clear and accurate explanation of why these physical characteristics can be learned from the data.

- Identification of Earth-like exoplanet (2 points): The student correctly identifies one exoplanet as Earth-like. [...]

{The answer is partially cut due to space constraints.}

Figure 3: An illustration of the prompt utilized to generate rubrics using GPT-4 for the Astrobiology course. This procedure is repeated for all courses under study, substituting the course name, correct answers, total grades, and questions accordingly. The generated rubrics are then integrated into the ZCoT prompt along with the correct answers for assignment grading. Specifically, the prompt showcased in Figure 2 is employed for grading, where the instructor-provided rubrics are replaced with rubrics generated by LLM.

identical to the one used for instructor grading. Despite anonymous grading to maintain impartiality, the final grade for each assignment is determined by calculating the median of scores from peer graders.

Bootstrap Resampling. In all settings involving bootstrap resampling, we conduct our experiments with 10,000 iterations using replacement. Additionally, in these settings, the significance of the results is reported based on a p-value threshold of 0.05 where relevant.

5 Results and Discussion

Tables 1 and 2 present the outcomes of our assessment of writing assignments using GPT-4 and GPT-3.5 across three MOOC subjects. These assessments are grounded on three different prompts devised in our study. Table 1 shows the average grades derived directly from grading the assignments of 10 students per question for each course, whereas Table 2 displays the results after conducting bootstrap resampling with 10,000 iterations on the data from Table 1. In the following, we provide both our qualitative and quantitative analyses.

5.1 Qualitative Analysis: Average Grades Evaluation

Qualitatively, our results unfold several insights:

(1) First and foremost, Table 1 reports no statistically significant differences in average grades produced by LLMs versus instructor-assigned grades. This suggests that LLM grading, akin to peer grading, is capable of generating grades consistent with

Table 1: Average grades from grading 10 students' writing assignments using the three proposed prompting techniques: ZCoT with instructor-provided correct answers, ZCoT combined with instructor-provided correct answers and rubrics, and ZCoT with instructor-provided correct answers and LLM-generated rubrics. Results encompass three MOOCs: Introductory Astronomy, Astrobiology, and the History and Philosophy of Astronomy. Grading is conducted using GPT-3.5 and GPT-4 as the underlying LLMs. Each average grade includes a subscripted standard deviation. As the differences in average grades for both peer and LLM grades relative to instructor grades are *not statistically significant*, no grades are marked with an asterisk (see Appendix C for details on p-values).

Courses	Questions	Instructor Grades	Peer Grades	GPT-3.5 ZCoT W/ Answers	GPT-3.5 ZCoT W/ Answers & Rubrics	GPT-3.5 ZCoT W/ Answers & LLM Rubrics	GPT-4 ZCoT W/ Answers	GPT-4 ZCoT W/ Answers & Rubrics	GPT-4 ZCoT W/ Answers & LLM Rubrics
Introductory Astronomy	Q1	$3.90_{\pm1.70}$	$5.15_{\pm0.87}$	$4.50_{\pm0.50}$	$2.90_{\pm1.92}$	$4.10_{\pm1.92}$	$4.75_{\pm1.29}$	$4.40_{\pm1.28}$	$4.40_{\pm1.50}$
	Q2	$8.20_{\pm1.17}$	$7.55_{\pm2.36}$	$7.60_{\pm0.92}$	$8.30_{\pm0.78}$	$8.40_{\pm0.66}$	$8.65_{\pm0.63}$	$8.30_{\pm0.90}$	$8.50_{\pm0.87}$
	Q3	$7.50_{\pm2.91}$	$7.40_{\pm2.59}$	$6.80_{\pm2.36}$	$7.20_{\pm2.68}$	$6.60_{\pm2.46}$	$7.60_{\pm2.73}$	$7.30_{\pm2.72}$	$7.60_{\pm2.75}$
	Q4	$7.40_{\pm2.73}$	$7.45_{\pm1.44}$	$6.80_{\pm2.48}$	$7.10_{\pm2.88}$	$6.80_{\pm2.96}$	$7.50_{\pm2.42}$	$6.90_{\pm2.84}$	$7.05_{\pm2.81}$
	Q5	$5.50_{\pm2.97}$	$7.40_{\pm2.62}$	$6.40_{\pm2.46}$	$7.40_{\pm2.91}$	$6.40_{\pm3.47}$	$6.20_{\pm2.86}$	$5.90_{\pm3.36}$	$6.35_{\pm3.49}$
Astrobiology	Q1	$6.80_{\pm3.49}$	$7.50_{\pm2.54}$	$7.40_{\pm2.65}$	$6.70_{\pm3.74}$	$6.58_{\pm2.98}$	$7.50_{\pm2.20}$	$7.10_{\pm3.05}$	$7.10_{\pm2.70}$
	Q2	$6.70_{\pm2.65}$	$7.45_{\pm3.24}$	$7.00_{\pm2.41}$	$7.30_{\pm2.65}$	$5.52_{\pm4.11}$	$7.90_{\pm2.17}$	$7.40_{\pm2.65}$	$7.10_{\pm3.11}$
	Q3	$7.90_{\pm2.70}$	$9.05_{\pm1.65}$	$6.70_{\pm2.28}$	$6.50_{\pm2.69}$	$5.03_{\pm4.22}$	$8.10_{\pm1.04}$	$7.50_{\pm1.28}$	$7.50_{\pm0.92}$
History & Philosophy of Astronomy	Q1	$3.50_{\pm0.67}$	$3.60_{\pm0.66}$	$2.70_{\pm0.46}$	$2.00_{\pm0.00}$	$2.85_{\pm0.63}$	$3.50_{\pm0.67}$	$3.20_{\pm0.98}$	$3.20_{\pm0.71}$
	Q2	$2.40_{\pm0.92}$	$3.70_{\pm0.64}$	$2.90_{\pm0.70}$	$1.80_{\pm0.98}$	$2.56_{\pm0.78}$	$3.25_{\pm0.72}$	$3.10_{\pm0.54}$	$2.95_{\pm0.72}$
	Q3	$2.70_{\pm0.64}$	$3.40_{\pm1.20}$	$2.70_{\pm1.00}$	$1.20_{\pm1.33}$	$1.70_{\pm1.19}$	$3.65_{\pm0.45}$	$3.20_{\pm0.60}$	$3.20_{\pm0.56}$
	Q4	$2.20_{\pm0.87}$	$3.80_{\pm0.60}$	$3.00_{\pm0.45}$	$1.10_{\pm1.04}$	$2.20_{\pm0.60}$	$3.25_{\pm0.75}$	$2.70_{\pm0.90}$	$2.95_{\pm0.72}$

Table 2: Average grades from grading 10 students' writing assignments using bootstrap resampled results in Table 1 with 10,000 iterations. Each average grade includes a subscripted standard deviation.

Courses	Questions	Instructor Grades	Peer Grades	GPT-3.5 ZCoT W/ Answers	GPT-3.5 ZCoT W/ Answers & Rubrics	GPT-3.5 ZCoT W/ Answers & LLM Rubrics	GPT-4 ZCoT W/ Answers	GPT-4 ZCoT W/ Answers & Rubrics	GPT-4 ZCoT W/ Answers & LLM Rubrics
Introductory Astronomy	Q1	$3.90_{\pm0.54}$	$5.15_{\pm0.27}$	$4.50_{\pm0.16}$	$2.90_{\pm0.61}$	$4.10_{\pm0.61}$	$4.75_{\pm0.41}$	$4.40_{\pm0.41}$	$4.40_{\pm0.48}$
	Q2	$8.20_{\pm0.37}$	$7.55_{\pm0.75}$	$7.60_{\pm0.29}$	$8.30_{\pm0.25}$	$8.40_{\pm0.21}$	$8.65_{\pm0.20}$	$8.30_{\pm0.28}$	$8.50_{\pm0.27}$
	Q3	$7.51_{\pm0.92}$	$7.41_{\pm0.82}$	$6.81_{\pm0.75}$	$7.21_{\pm0.85}$	$6.61_{\pm0.78}$	$7.61_{\pm0.87}$	$7.31_{\pm0.86}$	$7.61_{\pm0.87}$
	Q4	$7.41_{\pm0.86}$	$7.45_{\pm0.46}$	$6.81_{\pm0.78}$	$7.11_{\pm0.91}$	$6.81_{\pm0.94}$	$7.51_{\pm0.77}$	$6.91_{\pm0.90}$	$7.06_{\pm0.89}$
	Q5	$5.51_{\pm0.94}$	$7.40_{\pm0.83}$	$6.41_{\pm0.78}$	$7.40_{\pm0.91}$	$6.41_{\pm1.10}$	$6.21_{\pm0.90}$	$5.91_{\pm1.06}$	$6.36_{\pm1.10}$
Astrobiology	Q1	$6.81_{\pm1.09}$	$7.50_{\pm0.79}$	$7.41_{\pm0.84}$	$6.71_{\pm1.18}$	$6.60_{\pm0.94}$	$7.50_{\pm0.70}$	$7.11_{\pm0.95}$	$7.11_{\pm0.85}$
	Q2	$6.71_{\pm0.83}$	$7.46_{\pm1.02}$	$7.01_{\pm0.76}$	$7.31_{\pm0.83}$	$5.53_{\pm1.30}$	$7.91_{\pm0.68}$	$7.41_{\pm0.83}$	$7.11_{\pm0.98}$
	Q3	$7.89_{\pm0.85}$	$9.04_{\pm0.52}$	$6.70_{\pm0.71}$	$6.49_{\pm0.85}$	$5.02_{\pm1.35}$	$8.10_{\pm0.33}$	$7.50_{\pm0.41}$	$7.50_{\pm0.29}$
History & Philosophy of Astronomy	Q1	$3.50_{\pm0.21}$	$3.60_{\pm0.21}$	$2.70_{\pm0.14}$	$2.00_{\pm0.00}$	$2.85_{\pm0.20}$	$3.50_{\pm0.21}$	$3.20_{\pm0.31}$	$3.20_{\pm0.22}$
	Q2	$2.39_{\pm0.29}$	$3.69_{\pm0.20}$	$2.90_{\pm0.22}$	$1.80_{\pm0.31}$	$2.56_{\pm0.25}$	$3.25_{\pm0.23}$	$3.10_{\pm0.17}$	$2.95_{\pm0.23}$
	Q3	$2.70_{\pm0.20}$	$3.40_{\pm0.38}$	$2.69_{\pm0.32}$	$1.20_{\pm0.42}$	$1.70_{\pm0.38}$	$3.65_{\pm0.14}$	$3.20_{\pm0.19}$	$3.20_{\pm0.18}$
	Q4	$2.20_{\pm0.28}$	$3.80_{\pm0.19}$	$3.00_{\pm0.14}$	$1.10_{\pm0.33}$	$2.20_{\pm0.19}$	$3.25_{\pm0.24}$	$2.70_{\pm0.28}$	$2.95_{\pm0.23}$

those assigned by instructors.

(2) Tables 1 and 2 reveal that GPT-4 overall outperforms GPT-3.5 by generating grades that are more closely aligned with the instructor-assigned grades, especially in the context of the History and Philosophy of Astronomy.

(3) Focusing on the prompts in Tables 1 and 2, ones that integrate both instructor-provided rubrics and LLM-generated rubrics give rise to grades that are better aligned with those from the instructors. This is true for both GPT-4 and GPT-3.5.

(4) Looking at Tables 1 and 2, in almost all scenarios, GPT-4, when prompted with instructor-provided correct answers and rubrics, generates grades that are either superior to or on par with

peer grades. This even includes the History and Philosophy of Astronomy, which needs speculative or imaginative thinking abilities.

(5) When GPT-4 serves as the underlying base model, LLM-generated rubrics demonstrate comparable performance to instructor-provided rubrics, as illustrated in Tables 1 and 2. This finding is interesting, indicating the potential use of LLMs for rubric generation in educational contexts, enabling the automation of grading processes with only the input of correct answers.

(6) Tables 3 and 4 (in Appendix B) show that LLMs more often produce the lowest MAEs across all questions compared to peer grading, indicating that LLM-assigned grades are more accurate, with fewer errors or deviations from the instructor-assigned grades. Additionally, GPT-4 consistently assigns grades with lower MAEs than GPT-3.5.

(7) From the results listed in Tables 1 and 2, it is clear that the most difficult course to grade is the History and Philosophy of Astronomy. Both LLMs and peer graders struggle to match instructor-assigned grades closely. Nevertheless, even in these cases, GPT-4 outperforms peer grading, generating grades more aligned with instructor-assigned ones.

5.2 Quantitative Analysis: Question-by-Question Evaluation

Quantitatively, after assessing grades and feedback produced by LLMs on a question-by-question basis, we summarize our main observations as follows:

(1) Generally, LLM-assigned grades are higher and more generous than instructor-assigned grades. Although these scores are higher, they align more closely with instructor-assigned grades than peer grades do, particularly in the History and Philosophy of Astronomy. Despite this alignment, some instability was observed. For example, in the Astrobiology course, both models vary dramatically regardless of the prompt strategy employed, with GPT-3.5 scoring low and GPT-4 scoring high, differing from the instructor-assigned scores. We also observed that LLMs encounter challenges based on the length of students' answers, particularly with extremely short or excessively long responses. In such cases, while GPT-4 handled short answers well, it struggled with long ones.

(2) There is a bias towards middle scores and against extremes, especially when ZCoT combines with correct answers and, to a lesser extent, when ZCoT couples with correct answers and LLM-

generated rubrics. Both prompts lead GPT-3.5 and GPT-4 to be less likely to assign full marks compared to the instructor. Also, it is rare for LLMs to give a zero score on the same questions as the instructor. This is particularly evident in the History and Philosophy of Astronomy, where GPT-3.5 rarely gave full credit and assigned more zeros than the instructor for the ZCoT with correct answers and LLM-generated rubrics prompt, which performed well with other assignments and with GPT-4 on the same assignments.

(3) One of the most interesting results of this study is the feedback from LLMs. Both GPT-3.5 and GPT-4 produced subject-appropriate explanations specific enough to justify point deductions, referring to rubric criteria when provided. Most importantly, LLM feedback was more consistent and comprehensive than peer graders, who often failed to leave feedback or provided insufficient detail for students to improve. In contrast, LLMs, regardless of the prompt strategy, provided accurate and constructive feedback. Namely, when GPT-4 employs ZCoT with correct answers and instructor-provided rubrics, it often matches instructor feedback, including details leading to point deductions.

(4) In our experiments with GPT-3.5 and GPT-4, we did not observe any instances of hallucinations when generating grades and feedback for assignments. This can be attributed to the inclusion of detailed rubrics and correct answers in the input prompt, which guide the output into a more structured and factual format. These findings suggest the potential of LLMs to serve as reliable automated grading systems when provided with sufficient guidance in the input prompt.

6 Conclusion

We studied the feasibility and validity of using large language models (LLMs) to replace peer grading in massive open online courses (MOOCs). To this end, we developed three distinct prompts based on the zero-shot chain-of-thought (ZCoT) prompting technique, each with varying in-context information. Specifically, our methods include ZCoT with instructor-provided correct answers, ZCoT with instructor-provided correct answers and rubrics, as well as ZCoT with instructor-provided correct answers and LLM-generated rubrics.

We evaluated the effectiveness of each prompt using two LLMs, GPT-4 and GPT-3.5, by grading assignments from three MOOCs: Introductory

Astronomy, Astrobiology, and the History and Philosophy of Astronomy. Our findings indicated that ZCoT with instructor-provided correct answers and rubrics, particularly with GPT-4, consistently provides grades that are more aligned with instructor-assigned grades compared to peer grading. Further, we found that grading courses requiring reasoning/imaginative skills, e.g., the History and Philosophy of Astronomy, is more challenging. In such cases, GPT-4 also outperforms peer grading.

Overall, our results show the potential of using LLMs in MOOCs to automate grading systems and replace peer grading, thus enhancing the learning experience for millions of online students globally.

7 Limitations

In this research, we studied the promise of LLM grading as a substitute for peer grading in MOOCs to automate the grading process. Yet, opportunities exist for refining the alignment between LLM-assigned and instructor-assigned grades, notably in advanced disciplines such as Philosophy and Mathematics, which demand robust reasoning abilities. Hence, further research is warranted to refine grading methodologies for improved congruence with instructor evaluations.

8 Ethical Considerations

The use of LLMs for automated grading introduces several ethical concerns that warrant careful examination. One major concern is fairness in grading. In fact, biases present in the training data of LLMs could lead to systematic disparities in scores across different demographic groups. For example, if an LLM has been trained on data that reflects societal biases, it might inadvertently favor certain linguistic styles, cultural norms, argumentation styles, or perspectives, thereby disadvantaging students who do not conform to these patterns.

Students' perceptions of machine-generated feedback also present ethical concerns. Feedback from LLMs might be perceived as less personal or less tailored to individual learning paths compared to human-generated feedback. It is essential to communicate clearly with students about how automated grading systems using LLMs operate, the rationale behind their feedback, and the ways in which human oversight is incorporated into the grading process.

In conclusion, while automated grading with LLMs offers significant benefits in efficiency and scalability, it is imperative to address the aforesaid ethical considerations proactively. Ensuring fairness, minimizing biases, and fostering positive perceptions among students are crucial steps toward leveraging the advantages of LLMs while upholding the integrity of the educational experience.

References

Mohamed Alseddiqi, Anwar AL-Mofleh, Leena Albalooshi, and Osama Najam. 2023. Revolutionizing online learning: The potential of chatgpt in massive open online courses. *European Journal of Education and Pedagogy*.

Li-Hsin Chang and Filip Ginter. 2024. Automatic short answer grading for finnish with chatgpt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23173–23181.

Andre Del Carpio Gutierrez, Paul Denny, and Andrew Luxton-Reilly. 2024. Evaluating automatically generated contextualised programming exercises. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 289–295, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

B. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Martin Formanek, Matthew C. Wenger, Sanlyn R. Buxner, Chris D. Impey, and Tenzin Sonam. 2017. Insights about large-scale online peer assessment from an analysis of an astronomy mooc. *Computers & Education*, 113:243–262.

Hiroaki Funayama, Tasuku Sato, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Balancing cost and quality: An exploration of human-in-the-loop frameworks for automated short answer scoring. *Cornell University - arXiv*.

Hagit Gabbay and Anat Cohen. 2024. Combining llm-generated and test-based feedback in a mooc for programming. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 177–187.

Dilrukshi Gamage, Dilrukshi Gamage, Thomas Staubitz, Thomas Staubitz, Mark E. Whiting, and Mark E. Whiting. 2021. Peer assessment in moocs: Systematic literature review. *Distance Education*.

Shahriar Golchin and Mihai Surdeanu. 2023a. Data contamination quiz: A tool to detect and estimate contamination in large language models. *CoRR*, abs/2311.06233.

Shahriar Golchin and Mihai Surdeanu. 2023b. Time travel in llms: Tracing data contamination in large language models. *CoRR*, abs/2308.08493.

Hasnain Heickal and Andrew Lan. 2024. Generating feedback-ladders for logical errors in programming using large language models. *arXiv preprint arXiv:2405.00302*.

Chris Impey. 2024a. Astrobiology: Exploring other worlds. Accessed: 2024-06-08.

Chris Impey. 2024b. Astronomy: Exploring time and space. Accessed: 2024-06-08.

Chris Impey. 2024c. Knowing the universe: An introduction to astronomy. Accessed: 2024-06-08.

Chris Impey, Matthew Wenger, Martin Formanek, and Sanlyn Buxner. 2016. Bringing the universe to the world: lessons learned from a massive open online class on astronomy. *Communicating Astronomy with the Public*, 21:20–30.

Chris D. Impey, Matthew C. Wenger, and Carmen L. Austin. 2015. Astronomy for astronomical numbers: A worldwide massive open online class. *The International Review of Research in Open and Distributed Learning*, 16(1).

Sven Jacobs and Steffen Jaschke. 2024. Evaluating the application of large language models to generate feedback in programming education. *arXiv preprint arXiv:2403.09744*.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, J. Weller, Jochen Kühn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*.

Natalie Kiesler, Dominic Lohr, and Hieke Keuning. 2023. Exploring the potential of large language models to generate formative programming feedback. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–5.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv:2303.15621*.

Jordan K Matelsky, Felipe Parodi, Tony Liu, Richard D Lange, and Konrad P Kording. 2023. A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv:2308.02439*.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Wesley Morris, Scott A. Crossley, Langdon Holmes, and Anne Trumbore. 2023. Using transformer language models to validate peer-assigned essay scores in massive open online courses (moocs). *International Conference on Learning Analytics and Knowledge*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. Openai partners with khan academy to offer math exercises. Accessed: 2024-06-08.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Maciej Pankiewicz and Ryan S Baker. 2023. Large language models (gpt) for automating feedback on programming assignments. *arXiv preprint arXiv:2307.00150*.

Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Braga Ribeiro, Danilo Lucena, Alberto Souza, and Kiev Gama. 2023. Large language models for education: Grading open-ended questions using chatgpt. *Brazilian Symposium on Software Engineering*.

Brian Riordan, Brian Riordan, Andrea Horbach, Andrea Horbach, Aoife Cahill, Aoife Cahill, Torsten Zesch, Torsten Zesch, Chong Min Lee, and Chong Min Lee. 2017. Investigating neural architectures for short answer scoring. *BEA@EMNLP*.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

Alessia Tripaldelli, George Pozek, and Brian Butka. 2024. Leveraging prompt engineering on chatgpt for enhanced learning in cec330: Digital system design in aerospace. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Zhengdong Zhang, Zihan Dong, Yang Shi, Thomas Price, Noboru Matsuda, and Dongkuan Xu. 2024. Students' perceptions and preferences of generative artificial intelligence feedback for programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23250–23258.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

Terry Yue Zhuo. 2023. Ice-score: Instructing large language models to evaluate code. *arXiv preprint arXiv:2304.14317*.

A   Details on Assignment Questions

To offer comprehensive insight into the question structure of each course, we list all questions per course below. Note that we cannot publicly release all correct answers and rubrics to these questions due to the active status of the respective courses on Coursera. The questions are as follows:

- **Course:** Introductory Astronomy

  1. In terms of the scientific method, how does astronomy differ from a lab science like chemistry or biology? How can astronomers be confident in their understanding of objects that are remote from the Earth? Ancient cultures built some impressive structures that incorporated astronomical functions and information (Stonehenge, Chichen Itza, the Great Pyramid). A friend or acquaintance of yours tries to argue that some of these structures and artifacts are evidence of "ancient astronauts" or visits by intelligent aliens. How would you rebut or argue against this idea?

  2. What are the advantages of large telescopes? Provide at least one. Why do astronomers want telescopes in space when putting them there is expensive? What are some examples of wavelength regions beyond the spectrum of visible light where astronomers can learn about the universe? Provide at least two.

  3. What are the two main, indirect methods for finding exoplanets? Why is it so difficult to see exoplanets directly in an image? What are some similarities or differences between our Solar System and new, distant planet systems? Provide at least one similarity and/or difference.

  4. What is the source or cause of the Sun's light, and how do all the elements in the periodic table get produced? What is the general process by which a large diffuse cloud of gas turns into a star and surrounding planets? Name the two end states of stars much more massive than the Sun and describe their physical properties?

  5. Why do astronomers often say that large telescopes are like time machines, or equivalently, why is distant light old light? What is the evidence that the universe began in a hot, dense state of 13.8 billion years ago? The atoms in our bodies and in all the stars in all 100 galaxies form a small percentage of the contents of the universe. What are the two dominant ingredients of the universe and why are astronomers so unsure of their physical nature?

- **Course:** Astrobiology

  1. Clearly identify the detection methods used to gather data for each exoplanet. Briefly explain how each detection method works. Correctly identify

both detection methods. Clearly explain how each detection method works. Correctly identify which physical characteristics can be learned from each set of data, and explain why. Clearly identify the physical characteristics of both exoplanets. Clearly identify one exoplanet as Earth-like.

2. Discuss how habitable zone range and spectral type are related.
Yousef says that all three Earth-like planets likely have liquid surface water because they all orbit at 1 astronomical unit (AU). Since the Earth orbits at 1 AU from the sun, and we know Earth has liquid surface water, then these exoplanets should as well. Address Yousef's statement. Clearly state whether you agree or disagree with the conclusion. Explain your answer with evidence and use data to support your answer.
Lora says that both exoplanets 2 and 3 will have liquid water, but not exoplanet 1. The star for exoplanet 1 is spectral type A, which is too big and hot and would evaporate water on exoplanets. But exoplanet 2 and exoplanet 3 orbit around favorable spectral types G and M, therefore they likely have liquid surface water. Clearly state whether you agree or disagree with Lora's conclusion. Explain your answer with evidence and use data to support your answer.

3. Discuss why/whether it is plausible that exobiology exists at all, in terms of exoplanet type, spectral type, and orbital distance. Clearly state the geologic eon and make a strong argument for the state of exobiology, given their choice of geologic eon. Discusses whether the exobiology is unicellular, multicellular, intelligent, etc. Use geologic eon or age, exoplanet type, and examples from class to support the argument. Present arguments in a clear, logical way that demonstrates understanding of concepts. Correctly use scientific terms or language. Connections between concepts should be well developed.

- **Course:** History & Philosophy of Astronomy
  1. Describe the practical and philosophical

importance of astronomy for humans living a nomadic lifestyle in 20,000 B.C. This assignment will be graded on the clarity and completeness of your answer as well as your use of course topics to support your claims. Your answer should be between 200 and 300 words (about one paragraph).

2. According to the current understanding of the universe, the cosmos has a definite beginning but an infinite future. What are the philosophical problems and implications of this? This assignment will be graded on the clarity and completeness of your answer as well as your use of course topics to support your claims. Your answer should be between 200 and 300 words (about one paragraph).

3. According to Enlightenment philosophy, why might liberty and personal rights be connected to the pursuit of science? This assignment will be graded on the clarity and completeness of your answer as well as your use of course topics to support your claims. Your answer should be between 200 and 300 words (about one paragraph).

4. Imagine if the universe was eternal and unchanging as proposed by the steady state model in the 1950's. What are the philosophical and scientific implications of this? This assignment will be graded on the clarity and completeness of your answer as well as your use of course topics to support your claims. Your answer should be between 200 and 300 words (about one paragraph).

B Mean Absolute Error Results

Tables 3 and 4 present the MAE values for LLM-assigned grades and peer grades relative to instructor-assigned grades. The results show that LLMs generally achieve lower MAE values than peer grades, with GPT-4 most frequently obtaining the lowest MAEs.

C Details on p-values

As detailed in Section 4 under Bootstrap Resampling, we use 10,000 iterations with replacement in all settings with bootstrap resampling. Table

Table 3: MAE values for 10 students' writing assignments, comparing LLM-assigned and peer grades to instructor-assigned grades. **Bold** values indicate the lowest MAE for each question.

Courses	Questions	Peer Grades	GPT-3.5			GPT-4		
			ZCoT W/ Answers	ZCoT W/ Answers & Rubrics	ZCoT W/ Answers & LLM Rubrics	ZCoT W/ Answers	ZCoT W/ Answers & Rubrics	ZCoT W/ Answers & LLM Rubrics
Introductory Astronomy	Q1	1.85	1.20	1.40	1.60	**1.05**	1.30	1.30
	Q2	1.25	0.80	0.90	1.00	**0.45**	0.50	0.50
	Q3	0.70	1.10	0.90	1.30	0.50	**0.40**	**0.40**
	Q4	1.15	1.00	0.70	0.80	**0.50**	0.70	0.55
	Q5	2.10	1.30	2.30	**1.10**	1.30	1.40	1.25
Astrobiology	Q1	1.30	1.60	1.30	1.68	1.70	**0.70**	1.30
	Q2	1.35	1.90	**1.00**	2.38	1.40	1.10	1.20
	Q3	1.85	2.20	2.00	3.02	1.60	1.80	2.00
History & Philosophy of Astronomy	Q1	0.70	0.80	1.50	0.95	**0.40**	0.50	**0.40**
	Q2	1.30	0.90	1.20	1.04	1.05	1.10	**0.75**
	Q3	1.30	0.80	1.50	1.00	0.95	**0.70**	**0.70**
	Q4	1.40	**0.60**	1.30	**0.60**	1.25	0.90	0.95
Overall Average	All	1.35	1.18	1.33	1.37	1.01	**0.92**	0.94

Table 4: MAE values for 10 students' writing assignments using bootstrap resampled results with 10,000 iterations, comparing LLM-assigned and peer grades to instructor-assigned grades. **Bold** values indicate the lowest MAE for each question.

Courses	Questions	Peer Grades	GPT-3.5			GPT-4		
			ZCoT W/ Answers	ZCoT W/ Answers & Rubrics	ZCoT W/ Answers & LLM Rubrics	ZCoT W/ Answers	ZCoT W/ Answers & Rubrics	ZCoT W/ Answers & LLM Rubrics
Introductory Astronomy	Q1	1.67	**1.30**	2.32	2.06	1.72	1.68	1.77
	Q2	1.68	1.32	1.01	0.96	**0.91**	1.02	0.98
	Q3	2.43	2.69	2.58	2.85	2.37	2.51	**2.35**
	Q4	**2.14**	2.46	2.62	2.74	2.34	2.68	2.62
	Q5	3.30	**2.96**	3.55	3.59	3.22	3.53	3.61
Astrobiology	Q1	3.35	3.39	3.92	3.62	**3.30**	3.57	3.47
	Q2	3.40	**2.93**	3.03	4.08	**2.93**	3.07	3.27
	Q3	2.32	3.06	3.27	4.57	**2.25**	2.51	2.49
History & Philosophy of Astronomy	Q1	**0.64**	0.94	1.50	0.87	0.66	0.86	0.78
	Q2	1.44	0.97	1.14	**0.94**	1.16	0.96	1.00
	Q3	1.28	0.82	1.74	1.32	1.02	**0.75**	0.77
	Q4	1.48	0.68	1.42	**0.56**	1.21	1.00	0.99
Overall Average	All	2.09	1.96	2.34	2.35	**1.92**	2.01	2.01

Table 5: p-values for results in Table 1, computed through bootstrap resampling with 10,000 iterations. None of the p-values exceed the 0.05 threshold for significance, implying no statistically significant difference in average grades is detected.

| Courses | Questions | Peer Grades | GPT-3.5 | | | GPT-4 | | |
			ZCoT W/ Answers	ZCoT W/ Answers & Rubrics	ZCoT W/ Answers & LLM Rubrics	ZCoT W/ Answers	ZCoT W/ Answers & Rubrics	ZCoT W/ Answers & LLM Rubrics
Introductory Astronomy	Q1	0.48	0.49	0.54	0.86	0.49	0.52	0.55
	Q2	0.52	0.57	0.78	0.64	0.49	0.78	0.60
	Q3	0.95	0.63	0.84	0.59	0.97	0.89	0.95
	Q4	0.98	0.64	0.80	0.67	0.96	0.74	0.79
	Q5	0.50	0.57	0.52	0.61	0.66	0.79	0.64
Astrobiology	Q1	0.66	0.69	0.98	0.89	0.65	0.87	0.86
	Q2	0.65	0.84	0.69	0.55	0.53	0.64	0.80
	Q3	0.50	0.51	0.52	0.50	0.87	0.71	0.72
History & Philosophy of Astronomy	Q1	0.83	0.59	0.61	0.52	1.00	0.61	0.57
	Q2	0.50	0.56	0.55	0.71	0.52	0.52	0.51
	Q3	0.58	1.00	0.49	0.48	0.55	0.57	0.55
	Q4	0.56	0.54	0.53	1.00	0.51	0.55	0.51

5 presents the p-values for the differences in average scores referenced in Table 1. Note that the significance of average differences is assessed using a threshold of 0.05. Based on the computed p-values, none of the differences in average scores is statistically significant.