The Effects of AI Model Choice and Prompting on the Consistency of Essay Scores

April 1 2025: Preliminary version

Prof.dr. Elbert Dijkgraaf[1]

Erasmus School of Economics, Center for Innovative Learning

Summary

This study investigates the consistency and effectiveness of artificial intelligence (AI) grading systems compared to traditional human evaluators, examining variations among different AI models and the impact of various prompt designs.

The correlation between human and AI grading shows substantial variation contingent upon the specific prompt and AI model employed. For example, basic prompts with ChatGPT-4 exhibit markedly low correlations, while tailored prompts with ChatGPT-4.5 and Grok 3 exhibit substantially higher correlations. However, the "Deep Research" mode can negatively impact grading accuracy.

Modifications to the prompts can result in significant positive and negative changes in grading correlation. Adding specific details can enhance correlation with human grading for some models but not for others. The correlation is noticeably higher for questions that are more factual.

Discrepancies between AI and human grading levels are particularly evident when human grades are low, leading to a wider variability in human grading results.

The internal consistency of AI grading varied also notably among models and prompts. Grok 3 and ChatGPT-4.5 exhibited exceptional internal consistency. In contrast, older AI models demonstrated far lower internal consistency, highlighting the importance of model sophistication in ensuring grading reliability. Similarly, the 'Deep Research' mode yields suboptimal performance.

Overall, this study highlights the potential efficacy of sophisticated AI systems to provide reliable and fair academic assessments, emphasizing the necessity of carefully selecting AI models and meticulously designing prompts to maximize their accuracy and consistency. However, even with the best available models large discrepancies with human graders are still present.

Keywords: Artificial Intelligence Grading, AI Model Consistency, Essay Evaluation, Prompt Design Influence, Human vs. AI Scoring, Educational Assessment Technology, Grading Reliability

Highlights:

- Assesses human versus AI grading consistency.
- Examines effect of prompt design for essay grading.
- Examines effect of varying AI models for essay grading.

Acknowledgements: The author thanks Marloes Nederhand, Jason Pridmore, Milan Vidakovic, Jonathan Flores Minuesa and Joao Goncalves for valuable comments.

[1] Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, dijkgraaf@ese.eur.nl

1. Introduction

In education, grading essays is an essential but time-consuming task. Teachers often face the challenge of delivering accurate and unbiased evaluations, a process that is not only labor-intensive but can also be subject to human bias. Literature indicates that this bias can undermine the consistency and fairness of assessments, leading to variability in grading. Moreover, manual grading is time-consuming, and many educators find it an unappealing part of their work.

The rise of artificial intelligence (AI) offers a potentially promising solution. AI-based systems, such as ChatGPT and Grok, have the potential to not only speed up the grading process but also standardize it, reducing the possibility of human errors. This research focuses on whether AI grading can provide consistency comparable to, or better than, human graders. The ultimate goal of this study is to demonstrate under what circumstances and with what design of prompts AI grading systems can offer reliable consistency in their assessments. This would represent a significant step forward in increasing the efficiency and fairness of academic evaluation processes.

2. Literature

In the era of digital educational innovations, artificial intelligence (AI) offers potentially promising solutions for evaluating student work. Grading student work is generally seen as a time-consuming and unpleasant task that often leads to high workloads among teachers. For example, Erturk et al. (2022) show that boredom arises during the grading of essays and increases as the grading process is prolonged. This can lead to bias in human grading. Additionally, it is clear that it is time-consuming and thus detracts from more valued tasks such as teaching and research.

AI might contribute to reducing human assessment bias, such as emotional and cognitive biases. The use of AI systems, which may evaluate based on consistent criteria without the influence of human moods or personal preferences, could lead to a more standardized grading procedure (Nguyen et al., 2023). This suggests that AI can help reduce the inherent human bias in traditional grading processes. Recent studies, such as those by Nguyen et al. (2023), have indeed shown that AI, like ChatGPT, can help increase the consistency and objectivity of assessments. Although AI has proven effective at lower cognitive levels, it shows less consistency at higher levels that require more analytical thinking (Nguyen et al., 2023). This raises questions about the capacity of AI to generate more complex evaluative judgments, often required in academic settings.

Various articles assume that the structure and clarity of prompts can play a crucial role in the effectiveness of essay assessments by AI. Detailed and carefully designed prompts could improve the performance of AI in grading by clearly delineating the criteria and expected responses, thereby reducing ambiguity and increasing consistency. However, current studies do not provide direct empirical evidence confirming this relationship (Kooli and Yusuf, 2024; Misgna et al., 2024).

Jackaria et al. (2024) used human and AI rating of 20 essays using ChatGPT 3.5 and one prompt. They found poor consistency between human and AI rating, good consistency between human raters and moderate consistency between different AI ratings.

Tate et al. (2024) used human and AI rating of several hundreds of essays using ChatGPT 3.5 and 4 and one prompt per essay sample. They find that humans are more consistent than ChatGPT

3.5, but ChatGPT 4 was more consistent than humans. The AI systems assigned fewer extreme scores, both high and low, with the differences generally not reaching statistical significance.

In summary, the literature indicates that, although AI has the potential to increase the efficiency and objectivity of assessments, it is essential to explore the limitations and variability in AI performance. A notable gap in current research is the lack of direct empirical evidence demonstrating the influence of specific prompt designs and AI model choice on the consistency of AI evaluations.

3. Methodology and data

This study involves a quantitative analysis comparing the consistency of AI evaluations with human evaluations. Specifically, ChatGPT and Grok 3 are used to evaluate 18 essays, each written by third-year bachelor's economic students, five times per prompt. The AI evaluations are then compared with those of a human grader, a professor with 30 years of experience, to assess differences in grades and the ranking of students. The human grader was during grading not aware of this project. Nine versions of ChatGPT were used, specifically 4, 4o, o3-mini, o3-mini high, o pro, the last three also with the deep research modus and 4.5. Also Grok 3 and Perplexity are used.

The study involves 18 third-year bachelor's students from Erasmus University who each wrote an essay of approximately 400 words with human supervision and no internet access about economics of climate change. These essays serve as the primary material for both AI and human evaluation. The essays covered various aspects of the EU Emission Trading System (ETS) for CO2. The specific questions asked to the students included:

1. What is the basic principle of ETS and how is it intended to reduce greenhouse gas emissions?
2. How does ETS compare to carbon taxes in terms of effectiveness and efficiency?
3. Discuss the main challenges encountered in the implementation of ETS in the EU.
4. What are the critical design features that determine the success of an ETS in achieving its environmental goals?
5. Evaluate the potential long-term impact of ETS on industrial innovation and sustainable economic development.

The collection of assessment data included the traditional method of essay grading by an experienced professor and the evaluations by ChatGPT and Grok 3. Each essay was independently evaluated five times by the AI to explore variability in the assessments. Internal consistency is the stability of the AI's grades when the grading process is repeated multiple times under the same conditions such as equal prompts and essays.

The data analysis involves calculating the correlation between the first round AI and human grader's evaluations, and between successive AI evaluations (first-second, second-third, etc.) for internal consistency. This was done both for the absolute grades and for the ranking of the students based on the grade. The ideal outcome is a correlation of 1, which would indicate perfect consistency.

For this study, a total of 10 prompts were used to evaluate the consistency of the AI evaluations, conducted by ChatGPT and Grok 3. Each of the 18 essays was evaluated with each of these prompts. Below is a detailed description of the different prompts used to evaluate the essays:

1. Score per Sub-question between 0 and 5 points. NB: Sub-questions 1-5 mentioned above are included in the prompt.

2. Add to 1: Evaluate at the level of third-year bachelor students from Erasmus University Rotterdam (Economics).
3. Add to 1: You are a teacher.
4. Adjust 1: Adjusted Points Distribution: From the fourth prompt onwards, the score per question was increased from 5 to 20 points as this maybe makes more precision possible.
5. All of the aboven plus use the following criteria to guarantee consistency in the evaluations:
    a. Clarity of Understanding (0-5 points): Assesses whether the student's explanation is clear and correct.
    b. Depth (0-5 points): Evaluates whether the concepts are thoroughly analyzed.
    c. Comparison and Contrast (0-5 points): Analyzes how effectively the student compares ETS with other systems such as carbon taxation.
    d. Problem Analysis (0-5 points): Assesses whether the challenges and problems of ETS are thoroughly examined.
6. Add to 1: Assess the accuracy of the essays.
7. Instead of 1: give score of 0-100 points for whole essay.
8. Add to 1: base evaluation on answer model. NB: answer model for the five questions were included, e.g. for the first question we include “Caps total greenhouse gas emissions. Allows trading of emission allowances.”
9. Combination of Criteria: A combination of prompts 1, 2, 6, and 8.
10. Add to 1: assess based on improvement potential, the higher the potential the lower the grade.

4. Results

Table 1 shows the correlation between the evaluations of AI and those of an experienced human evaluator for different prompt and AI models. Table 2 shows the correlation of the grades for the second run compared to the first run. Table A3-A5 in the appendix do this for the other runs. Table 3 presents the average correlation of all 5 runs. Figure 1 gives the total scores for ChatGPT 4.5. Figure 2 gives the total scores for Grok 3.

4.1 Human versus AI

The correlation between human and AI grading depends on both prompting and model choice (see Table 1). The basic prompt for ChatGPT-4 (row 1) has a correlation of only 0.12, while there are three sub-questions with even a negative correlation. At the other hand prompt 9 for ChatGPT-4.5 and Grok 3 result in respectively a correlation of 0.76 and 0.73 and positive correlations for all sub-questions (row 42 and 55).

It seems that new models perform better for many, but not all prompts. For Grok 3 and ChatGPT-4.5 this is the case for all prompts except three (row 37, 48 and 52). In contrast, the introduction of complexity through the "Deep Research" mode was found to negatively affect grading accuracy, yielding poor or even negative correlations (row 32-34).

Refining the prompt influences the correlation of human and AI grading significantly. For ChatGPT-4 the correlation increases from 0.12 tot 0.69 if the answering model is included (row 8 versus 1). For ChatGPT-4.5 the correlation increases from 0.61 to 0.73 for prompt 9 (row 55).

However, refining does not always result in higher correlation. For ChatGPT-4o the correlation decreases with each addition to the basic prompt (rows 11-20), while for ChatGPT-o1 pro the correlation decreases for some additions and increases for others (rows 21-31). For Grok 3 nearly all additions perform better, while this is only the case for one addition (row 55) for ChatGPT-4.5.

It shows that the correlation is higher for Q1-Q3 compared with Q4 and Q5. It seems that AI is performing better with more factual questions. It could thus be the case that AI is performing better for lower levels of the Bloomsberg taxonomy.

In general the AI models perform better for the total score compared to the rank order, but the differences are not very big.

There are many cases where the level of the human grades are substantially lower than for the AI grading (see Figures 1 and 2). This is especially the case for low human grades resulting in a larger spread for human grading.

4.2 Internal consistency

The internal consistency between the five runs depends on both prompting and model choice (see Table 2 and A2-5). The basic prompt for ChatGPT-4 (row 1) has a correlation of only 0.29. At the other hand prompt 9 for ChatGPT-4.5 and GROK result in respectively a correlation of 0.96 and 0.97 and high correlations for all sub-questions (row 42 and 55).

New models perform far better for many, but not all prompts. For Grok 3 and ChatGPT-4.5 this is the case for all prompts except one (row 39). In contrast, the introduction of complexity through the "Deep Research" mode was found to negatively affect grading accuracy, yielding poor or even negative correlations (row 32-34).

Refining the prompt influences the correlation of human and AI grading significantly. For ChatGPT-4 the correlation increases from 0.29 tot 0.81 if the answering model is included or if grading at bachelor-3 level is added (row 2 and 8 versus 1).

However, refining does not always result in higher correlation. For Grok 3 and ChatGPT-4.5 the basic prompt results already in a correlation above 0.9.

It shows that the correlation is comparable between the sub-questions. Differences in the Bloomsberg taxonomy seems not to influence internal consistency.

For many students the AI grades are not very different between the five runs, but there are cases with a larger spread (see figures 1 and 2).

Table 1. Correlation grading AI compared to human grader

		Model	ScoreQ1	ScoreQ2	ScoreQ3	ScoreQ4	ScoreQ5	Total	Rank
1	Basis	4	-0.13	-0.12	-0.04	0.45	0.29	0.12	0.12
2	Bach-3	4	0.42	0.34	0.38	0.27	0.11	0.55	0.41
3	Teacher	4	0.58	0.13	0.30	0.30	0.30	0.30	0.12
4	20 points	4	0.04	-0.02	0.41	0.25	0.60	0.26	0.17
5	2+3+4+criteria	4	0.43	0.19	0.09	0.20	0.33	0.34	0.34
6	Rigthness	4	0.21	0.19	0.26	0.16	0.47	0.50	0.30
7	No sub-questions	4	-	-	-	-	-	0.19	0.24
8	Answering model	4	0.66	0.49	0.48	0.54	0.11	0.69	0.64
9	2+6+8	4	0.52	0.57	0.18	0.20	0.28	0.44	0.49
10	Improv. Potent.	4	0.45	0.14	0.37	-0.05	0.06	0.13	0.15
11	Basis	4o	0.56	0.15	0.12	0.55	0.15	0.67	0.51
12	Bach-3	4o	0.58	0.15	0.58	0.28	-0.14	0.58	0.46
13	Teacher	4o	0.58	0.15	0.58	0.28	-0.14	0.58	0.46
14	20 points	4o	0.76	0.00	0.43	0.27	0.21	0.48	0.37
15	2+3+4+criteria	4o	0.40	0.19	0.44	0.38	0.20	0.40	0.18
16	Rigthness	4o	0.45	0.25	0.10	0.45	0.53	0.44	0.33
17	No sub-questions	4o						0.07	0.01
18	Answering model	4o	0.44	0.48	0.37	0.30	0.16	0.50	0.25
19	2+6+8	4o	0.37	0.39	0.58	0.37	0.17	0.54	0.30
20	Improv. Potent.	4o	0.67	0.28	0.25	0.24	0.27	0.42	0.18
21	Basis	o1 pro	0.33	0.64	0.39	0.33	0.10	0.60	0.13
22	Bach-3	o1 pro	0.41	0.20	0.39	0.52	0.29	0.55	0.40
23	Teacher	o1 pro	0.50	0.46	0.40	0.36	0.19	0.52	0.29
24	20 points	o1 pro	0.09	0.20	0.35	0.08	0.00	0.07	-0.02
25	2+3+4+criteria	o1 pro	0.46	0.58	0.34	0.55	0.16	0.50	0.44
26	Rigthness	o1 pro	0.62	0.55	0.45	0.39	0.31	0.63	0.50
27	No sub-questions	o1 pro						-0.41	-0.37
28	Answering model	o1 pro	0.76	0.71	0.49	0.28	-0.10	0.54	0.46
29	2+6+8	o1 pro	0.72	0.43	0.66	0.16	0.28	0.64	0.49
30	Improv. Potent.	o1 pro	0.55	0.38	0.51	0.24	0.32	0.49	0.35
31	2+6+8 again	o1 pro	0.59	0.32	0.46	0.18	0.00	0.54	0.33
32	2+6+8	03-mini DR	0.13	0.02	-0.26	-0.09	0.13	0.00	-0.05
33	2+6+8	03-mini-high DR	0.08	0.11	-0.18	0.33	0.13	0.13	0.02
34	2+6+8	o1-pro DR	-0.45	-0.22	-0.17	-0.38	-0.22	-0.41	-0.49
35	2+6+8	03-mini	0.41	0.50	0.69	0.38	0.21	0.71	0.61
36	2+6+8	03-mini-high	0.30	0.44	0.65	0.45	0.35	0.68	0.58
37	Basis	Grok 3	0.65	0.40	0.49	0.46	0.28	0.58	0.41
38	Bach-3	Grok 3	0.61	0.28	0.46	0.40	0.49	0.63	0.50
39	Teacher	Grok 3	0.49	0.09	0.58	0.30	0.50	0.66	0.52
40	20 points	Grok 3	0.62	0.53	0.60	0.43	0.44	0.67	0.63
41	2+3+4+criteria	Grok 3	0.72	0.34	0.49	0.34	0.32	0.58	0.45
42	Rigthness	Grok 3	0.67	0.45	0.46	0.45	0.61	0.71	0.54
43	No sub-questions	Grok 3						0.40	0.25
44	Answering model	Grok 3	0.60	0.47	0.70	0.39	0.36	0.72	0.60
45	2+6+8	Grok 3	0.60	0.50	0.70	0.54	0.49	0.76	0.62

46	Improv. Potent.	Grok 3	0.63	0.45	0.47	0.41	0.44	0.64	0.51
47	Basis	4.5	0.30	0.48	0.62	0.39	0.44	0.61	0.51
48	Bach-3	4.5	0.55	0.17	0.71	0.13	0.30	0.54	0.50
49	Teacher	4.5	0.56	0.23	0.59	0.24	0.33	0.53	0.44
50	20 points	4.5	0.64	0.35	0.57	0.08	0.28	0.60	0.41
51	2+3+4+criteria	4.5	0.67	0.25	0.56	0.28	0.09	0.56	0.45
52	Rigthness	4.5	0.68	0.34	0.45	0.24	0.32	0.55	0.41
53	No sub-questions	4.5						0.44	0.33
54	Answering model	4.5	0.36	0.52	0.68	0.20	0.40	0.59	0.52
55	2+6+8	4.5	0.62	0.56	0.73	0.30	0.31	0.73	0.67
56	Improv. Potent.	4.5	0.56	0.39	0.75	0.27	0.24	0.64	0.44
57	2+6+8	Perpl. 4.3	0.55	0.14	0.51	-0.08	0.19	0.36	0.21

Table 2. Correlation grading ChatGPT run 2 versus run 1

		Model	ScoreQ1	ScoreQ2	ScoreQ3	ScoreQ4	ScoreQ5	Total	Rank
1	Basis	4	0.15	0.18	0.08	0.66	0.44	0.29	0.27
2	Bach-3	4	0.09	0.18	0.43	0.76	0.64	0.81	0.60
3	Teacher	4	0.62	0.47	0.24	0.58	0.80	0.71	0.67
4	20 points	4	0.05	0.50	0.45	0.70	0.26	0.62	0.57
5	2+3+4+criteria	4	0.30	0.15	0.19	0.68	0.44	0.47	0.45
6	Rigthness	4	0.51	0.69	0.51	0.72	0.48	0.70	0.54
7	No sub-uestions	4	-	-	-	-	-	0.76	0.62
8	Answering model	4	0.58	0.77	0.60	0.49	0.41	0.81	0.51
9	2+6+8	4	0.60	0.37	0.57	0.57	0.17	0.63	0.60
10	Improv. Potent.	4	0.31	0.56	0.61	0.67	0.24	0.59	0.43
11	Basis	4o	0.73	0.27	0.58	0.92	0.58	0.77	0.78
12	Bach-3	4o	0.69	0.22	0.24	0.46	0.66	0.69	0.67
13	Teacher	4o	0.69	0.22	0.24	0.46	0.66	0.69	0.67
14	20 points	4o	0.50	0.52	0.60	0.65	0.61	0.68	0.59
15	2+3+4+criteria	4o	0.63	0.71	0.61	0.68	0.51	0.74	0.65
16	Rigthness	4o	0.68	0.60	0.59	0.71	0.64	0.80	0.67
17	No sub-questions	4o						0.91	0.89
18	Answering model	4o	0.83	0.67	0.66	0.80	0.50	0.77	0.64
19	2+6+8	4o	0.76	0.78	0.80	0.80	0.70	0.87	0.72
20	Improv. Potent.	4o	0.49	0.52	0.03	0.69	0.29	0.57	0.52
21	Basis	o1 pro	0.84	-0.06	0.59	0.67	-0.03	0.72	0.51
22	Bach-3	o1 pro	0.49	0.49	0.83	0.59	0.69	0.72	0.68
23	Teacher	o1 pro	0.54	0.58	0.60	0.62	0.53	0.90	0.78
24	20 points	o1 pro	0.41	0.48	0.81	0.60	0.40	0.61	0.50
25	2+3+4+criteria	o1 pro	0.29	0.78	0.72	0.75	0.65	0.83	0.73
26	Rigthness	o1 pro	0.62	0.77	0.70	0.68	0.64	0.90	0.74
27	No sub-questions	o1 pro						0.35	0.38
28	Answering model	o1 pro	0.64	0.73	0.82	0.58	1.00	0.91	0.89
29	2+6+8	o1 pro	0.86	0.85	0.83	0.83	0.61	0.93	0.91
30	Improv. Potent.	o1 pro	0.81	0.66	0.69	0.76	0.57	0.79	0.70
31	2+6+8 again	o1 pro	0.73	0.71	0.70	0.56	0.34	0.82	0.54

32	2+6+8	03-mini DR	0.13	0.28	0.24	0.32	0.15	0.23	0.26
33	2+6+8	03-mini-high DR	-0.09	-0.07	0.04	0.04	0.08	-0.08	-0.12
34	2+6+8	o1-pro DR	-0.15	0.03	-0.11	-0.17	-0.14	-0.10	-0.03
35	2+6+8	03-mini	0.91	0.95	0.87	0.80	0.58	0.88	0.87
36	2+6+8	03-mini-high	0.95	0.93	0.85	0.84	0.78	0.94	0.90
37	Basis	Grok 3	0.87	0.67	0.82	0.86	0.34	0.91	0.93
38	Bach-3	Grok 3	0.73	0.61	0.47	0.76	0.15	0.75	0.85
39	Teacher	Grok 3	0.72	0.63	0.57	0.87	0.89	0.87	0.90
40	20 points	Grok 3	0.89	0.90	0.85	0.94	0.79	0.94	0.92
41	2+3+4+criteria	Grok 3	0.97	1.00	0.75	0.95	0.87	0.96	0.98
42	Rigthness	Grok 3	0.90	0.73	0.87	0.89	0.87	0.94	0.91
43	No sub-questions	Grok 3						0.82	0.85
44	Answering model	Grok 3	0.85	0.80	0.70	0.84	0.89	0.94	0.90
45	2+6+8	Grok 3	0.95	0.85	0.96	0.92	0.79	0.96	0.93
46	Improv. Potent.	Grok 3	0.88	0.84	0.85	0.88	0.82	0.90	0.91
47	Basis	4.5	0.61	0.84	0.76	0.77	0.60	0.91	0.86
48	Bach-3	4.5	0.86	0.81	0.80	0.89	0.74	0.92	0.89
49	Teacher	4.5	0.80	0.91	0.89	0.95	0.78	0.96	0.94
50	20 points	4.5	0.87	0.90	0.88	0.88	0.86	0.97	0.95
51	2+3+4+criteria	4.5	0.86	0.87	0.81	0.88	0.78	0.95	0.93
52	Rigthness	4.5	0.91	0.94	0.83	0.92	0.75	0.95	0.92
53	No sub-questions	4.5						0.90	0.88
54	Answering model	4.5	0.89	0.91	0.78	0.81	0.91	0.92	0.91
55	2+6+8	4.5	0.97	0.96	0.91	0.93	0.90	0.97	0.94
56	Improv. Potent.	4.5	0.65	0.86	0.87	0.66	0.79	0.87	0.85
57	2+6+8	Perpl. 4.3	0.82	0.69	0.73	0.73	0.82	0.81	0.71

Table 3. Average correlation

Nr	Prompt	Model	Correlation
45	2+6+8	Grok 3	0.85
55	2+6+8	4.5	0.84
40	20 points	Grok 3	0.83
41	2+3+4+criteria	Grok 3	0.81
53	No sub-questions	4.5	0.81
54	Answering model	4.5	0.81
46	Improv. Potent.	Grok 3	0.81
44	Answering model	Grok 3	0.80
52	Rigthness	4.5	0.80
42	Rigthness	Grok 3	0.80
50	20 points	4.5	0.79
35	29	03-mini	0.78
36	29	03-mini-high	0.78
56	Improv. Potent.	4.5	0.78
51	2+3+4+criteria	4.5	0.76
48	Bach-3	4.5	0.76
49	Teacher	4.5	0.75

29	2+6+8	o1 pro	0.75
38	Bach-3	Grok 3	0.75
47	Basis	4.5	0.74
43	No sub-questions	Grok 3	0.73
39	Teacher	Grok 3	0.71
37	Basis	Grok 3	0.68
31	29 again	o1 pro	0.68
25	2+3+4+criteria	o1 pro	0.66
18	Answering model	4o	0.66
19	2+6+8	4o	0.66
28	Answering model	o1 pro	0.65
57	2+6+8	Perpl. 4.3	0.63
17	No sub-questions	4o	0.62
26	Rigthness	o1 pro	0.59
22	Bach-3	o1 pro	0.59
14	20 points	4o	0.57
6	Rigthness	4	0.57
30	Improv. Potent.	o1 pro	0.56
15	2+3+4+criteria	4o	0.55
24	20 points	o1 pro	0.55
23	Teacher	o1 pro	0.54
16	Rigthness	4o	0.52
11	Basis	4o	0.52
9	2+6+8	4	0.50
7	No sub-questions	4	0.50
12	Bach-3	4o	0.48
13	Teacher	4o	0.48
8	Answering model	4	0.46
21	Basis	o1 pro	0.45
20	Improv. Potent.	4o	0.45
3	Teacher	4	0.43
27	No sub-questions	o1 pro	0.42
2	Bach-3	4	0.41
5	2+3+4+criteria	4	0.41
10	Improv. Potent.	4	0.39
4	20 points	4	0.34
1	Basis	4	0.34
33	29	03-mini-high DR	0.07
34	29	o1-pro DR	-0.02
32	29	03-mini DR	-0.02

Figure 1. Total scores ChatGPT 4.5, human grader A

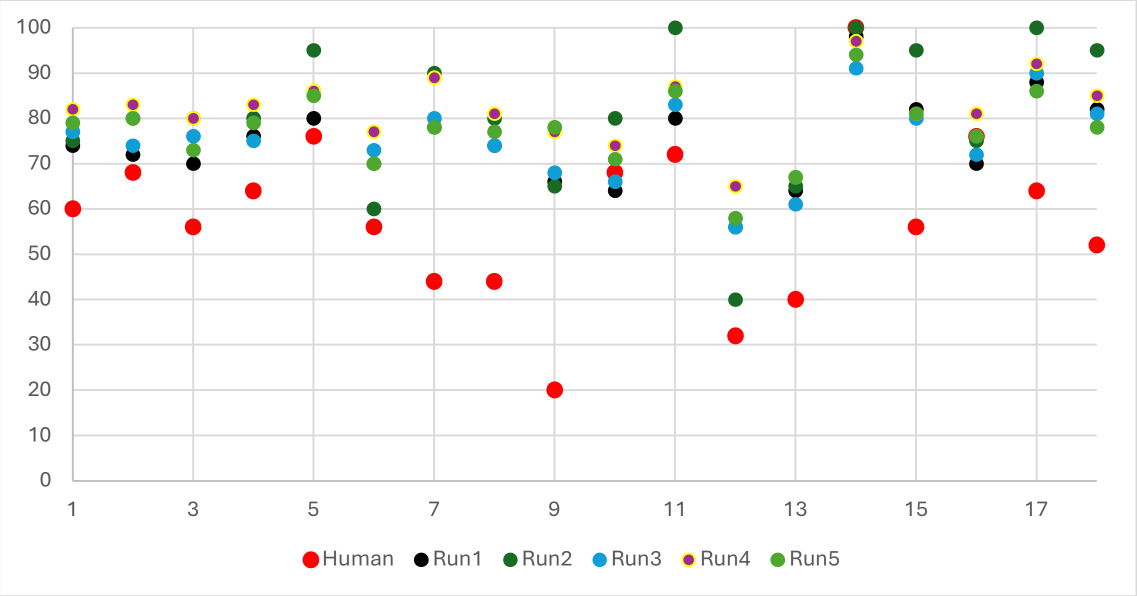
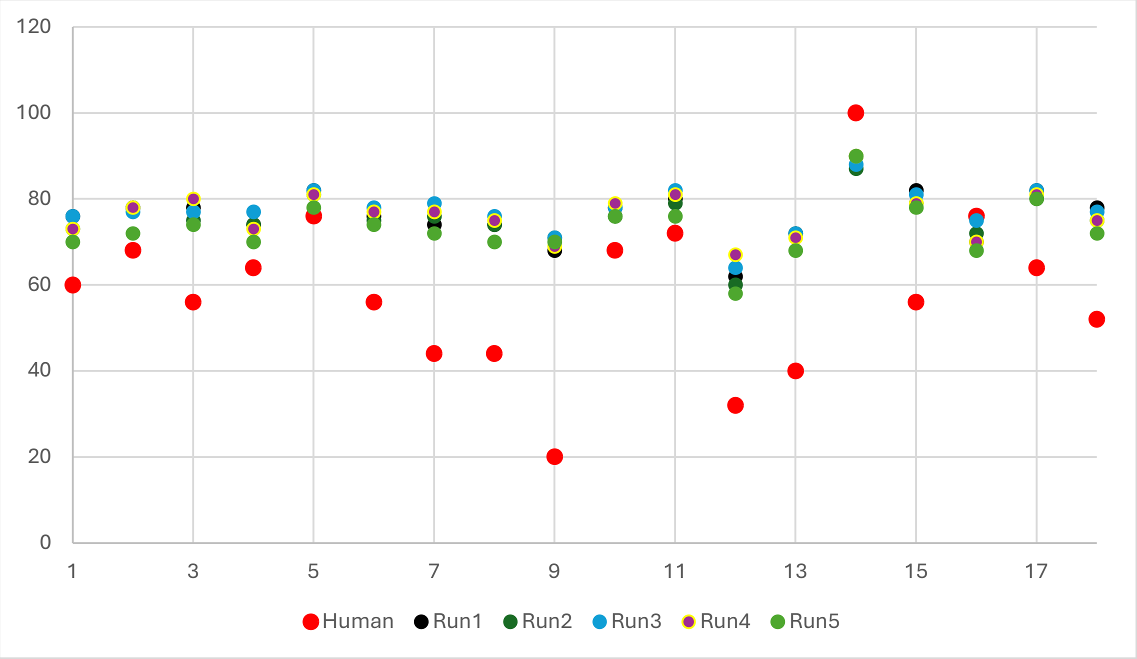


Figure 2. Total scores Grok 3, human grader A



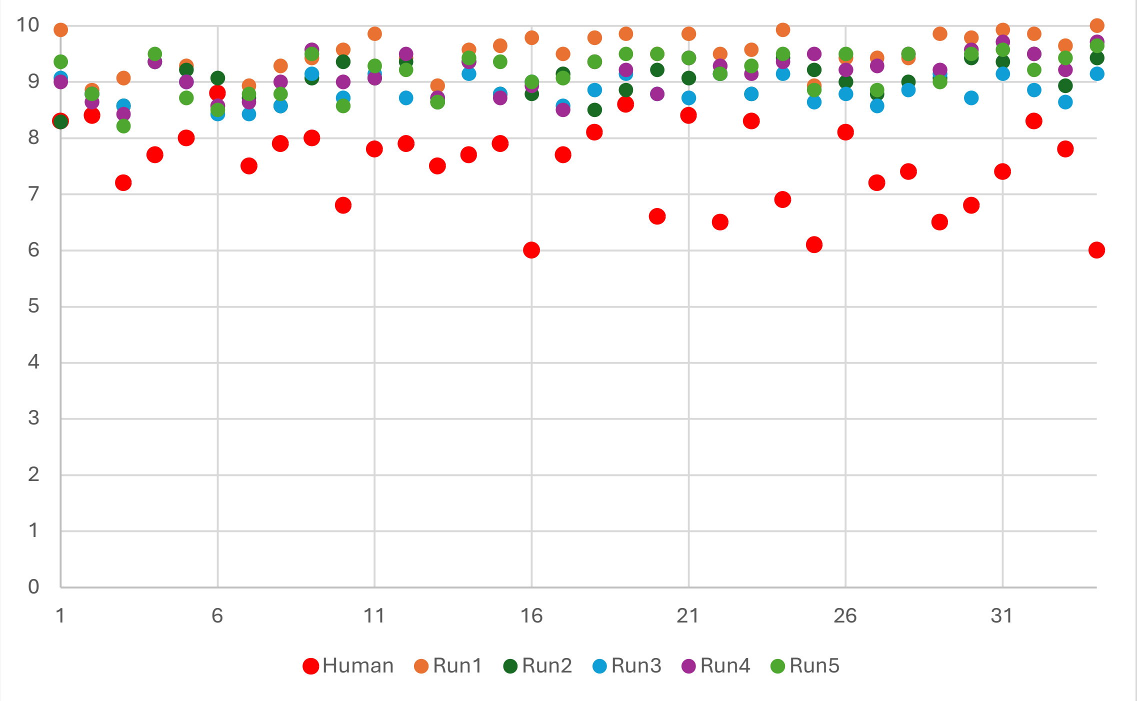
5. Sensitivity analysis

It could be the case that AI models' tendency to be cooperative and agreeable, even when performance is not so great, can explain that the human grader has in general lower grades. To test this we add prompts in ChatGPT-4.5 trying to compensate for this behavior. This was tested by incorporating elements into prompt 9 such as “I want you to be as critical as possible.”, “Be very critical”, “Only give high grades when answers are really good and low grades when they are not really good". The modifications to the prompts yielded results that did not meet the expected improvements. While the revised prompts resulted in lower grades for some essays, they frequently failed to replicate the lower grades assigned by the human grade. Interestingly, when after grading we commented that specific essays were given a far to high grade, ChatGPT immediately lowered the grades after re-evaluation with around 25%.

It could be that the result that the human grader has lower grades, especially for low grades is dependent on the human grader. We tested therefore with other essays from other human graders.

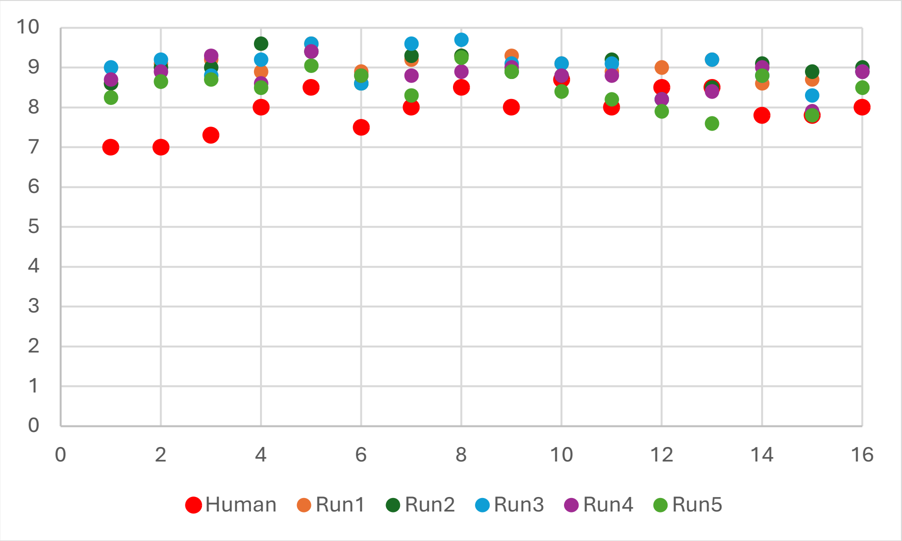
First, we use 34 essays from master students about the effects of climate change on labor outcomes, conflict or health outcomes in developing countries. These essays are around 5 pages. We use prompt 9 (replacing grading at bacherlor-3 for master level) and ChatGPT-4.5. Figure 3 presents the results. Now, the differences between the human grader and AI are even bigger. Many human grades are in the range 6-8, while all AI grades are above 9. A regression of the average of the five AI runs on human grades gives an insignificant coefficient and a R^2 of only 0.03, far lower than for grader A.

Figure 3. Total scores ChatGPT 4.5, human grader B



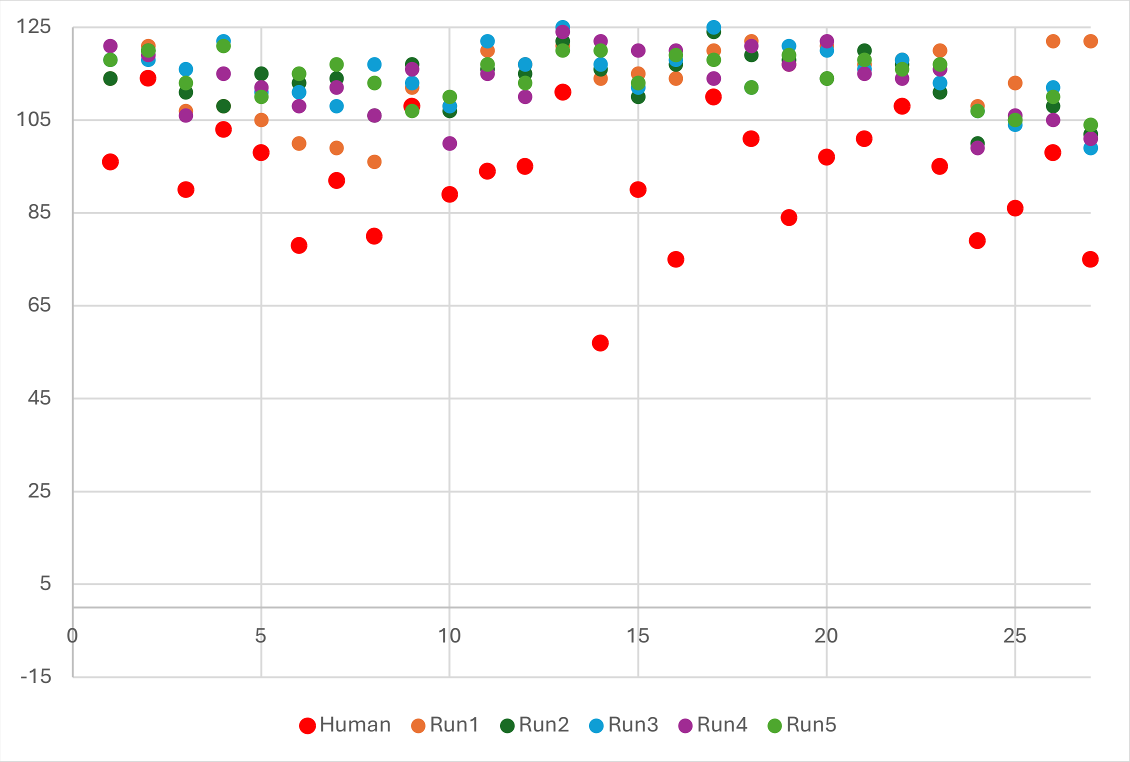
Second, we use 16 essays from master students about economics of migration. These essays are around 20 pages. We use prompt 9 (replacing grading at bacherlor-3 for master level) and ChatGPT-4.5. Figure 4 presents the results. Now, the differences between the human grader and AI are smaller. While often the human grade is on the bottom of the spread of the AI grades, they are often quite close. However, a regression of the average of the five AI runs on human grades gives an insignificant coefficient and a R^2 of only 0.01, far worse than grader A.

Figure 4. Total scores ChatGPT 4.5, human grader C



Third, we use 27 essays from bachelor students about reflections on the economic system. These essays are around 5 pages. We use prompt 9 and ChatGPT-4.5. Figure 5 presents the results. Now, the differences between the human grader and AI are larger again, especially for low grades. A regression of the average of the five AI runs on human grades gives a significant coefficient at 5% and a R^2 of 0.19, in between grader A versus B and C.

Figure 5. Total scores ChatGPT 4.5, human grader D



6. Conclusion

The alignment between AI-generated grades and human grading was found to be highly dependent on the AI model and prompt design, ranging from negligible correspondence under basic ChatGPT-4 prompts to much stronger correlations when using advanced models (like ChatGPT-4.5 or Grok 3) with carefully tailored grading prompts (see also Table 3).

Likewise, the internal consistency of AI grading across multiple runs varied by model sophistication: newer systems delivered far more stable results (with Grok 3 and ChatGPT 4.5 achieving near-perfect repeatability), whereas older models produced more variable scores upon repeated evaluation.

These findings highlight both the strengths and limitations of current AI grading systems: on one hand, state-of-the-art models can provide remarkably consistent and human-comparable assessments under optimal conditions, but on the other hand even the best AI graders still exhibit notable discrepancies from human evaluators—particularly for lower-scoring or complex responses—and certain elaborate prompting strategies (e.g. a “deep research” mode) can inadvertently diminish accuracy.

Therefore, while advanced AI has significant promise as a tool for reliable and fair assessment, careful model selection and prompt engineering are crucial to harness its benefits, and human oversight remains advisable to mitigate its shortcomings.

Future research should explore ways to further close the gap between AI and human grading, for example by testing AI performance across different cognitive task levels (e.g. varying Bloom’s taxonomy), refining prompt designs, and addressing cases where AI and human evaluations diverge, in order to guide the effective integration of AI into educational practice.

Given the fact that newer models seem to perform better on average, the gap between AI and human grading is likely to decrease further over time.

References

Erturk, S., van Tilburg, W. A. P., & Igou, E. R. (2022). Off the Mark: Repetitive Marking Undermines Essay Evaluations Due to Boredom. Motivation and Emotion, 46 (264–275).

Jackaria, P.M., B.H. Hajan & A.H, Mastul, A Comparative Aanalysis of the Rating of College Students'Essays by ChatGPT versus Human Raters, International Journal of Learning, Teaching and Educational Research, 23 (478-492).

Kooli, C., & Yusuf, N. (2024). Transforming Educational Assessment: Insights Into the Use of ChatGPT and Large Language Models in Grading. International Journal of Human–Computer Interaction.

Misgna, H., On, B.-W., Lee, I., & Choi, G. S. (2024). A Survey on Deep Learning-Based Automated Essay Scoring and Feedback Generation. Artificial Intelligence Review, 58(36).

Nguyen, M. N., Nguyen, B. T., Vo, D. T. H., Pham, T. T. T., Thai, H. T., & Ha, S. X. (2023). Evaluating the Efficacy of Generative Artificial Intelligence in Grading: Insights from Authentic Assessments in Economics. SSRN Electronic Journal.

Tate, T.P., J. Steiss, D. Bailey, S. Graham, Y. Moon, D. Ritchie, W. Tseng & M. Warschauer, Can AI provide useful holistic essay scoring? Computers and Education Artificial Intelligence, 7

Appendix

Table A3. Correlation grading ChatGPT run 3 versus run 2

		Model	ScoreQ1	ScoreQ2	ScoreQ3	ScoreQ4	ScoreQ5	Total	Rank
1	Basis	4	0.60	0.52	0.23	0.81	0.21	0.60	0.49
2	Bach-3	4	0.16	0.20	0.43	0.76	0.45	0.71	0.54
3	Teacher	4	0.35	0.66	0.41	0.56	0.45	0.65	0.54
4	20 points	4	0.27	0.47	-0.09	0.38	0.45	0.29	0.29
5	2+3+4+criteria	4	0.44	0.05	0.00	0.59	0.52	0.63	0.69
6	Rigthness	4	0.68	0.51	0.33	0.74	0.56	0.79	0.84
7	No sub-questions	4	-	-	-	-	-	0.58	0.48
8	Answering model	4	0.64	0.26	0.06	0.76	-0.02	0.51	0.49
9	2+6+8	4	0.82	0.55	0.39	0.65	0.26	0.68	0.53
10	Improv. Potent.	4	0.63	0.08	0.32	0.82	0.24	0.63	0.63
11	Basis	4o	0.82	0.84	0.62	0.66	0.41	0.80	0.64
12	Bach-3	4o	0.32	0.45	0.36	0.58	0.27	0.74	0.71
13	Teacher	4o	0.32	0.45	0.36	0.58	0.27	0.74	0.71
14	20 points	4o	0.72	0.63	0.55	0.88	0.76	0.83	0.57
15	2+3+4+criteria	4o	0.53	0.54	0.67	0.58	0.62	0.60	0.64
16	Rigthness	4o	0.85	0.72	0.64	0.73	0.42	0.85	0.69
17	No sub-questions	4o	-	-	-	-	-	0.86	0.80
18	Answering model	4o	0.61	0.74	0.78	0.84	0.49	0.79	0.70
19	2+6+8	4o	0.66	0.80	0.67	0.77	0.63	0.82	0.82
20	Improv. Potent.	4o	0.48	0.39	0.39	0.51	0.48	0.66	0.53
21	Basis	o1 pro	0.45	0.08	0.69	0.75	-0.01	0.66	0.65
22	Bach-3	o1 pro	0.69	0.60	0.67	0.60	0.40	0.77	0.72
23	Teacher	o1 pro	0.67	0.58	0.48	0.47	0.37	0.84	0.75
24	20 points	o1 pro	0.67	0.75	0.82	0.67	0.83	0.86	0.73
25	2+3+4+criteria	o1 pro	0.44	0.67	0.71	0.82	0.73	0.85	0.83
26	Rigthness	o1 pro	0.66	0.61	0.79	0.64	0.28	0.76	0.57
27	No sub-questions	o1 pro	-	-	-	-	-	0.79	0.74
28	Answering model	o1 pro	0.89	0.72	0.77	0.43	0.15	0.81	0.73
29	2+6+8	o1 pro	0.79	0.82	0.87	0.65	0.83	0.95	0.90
30	Improv. Potent.	o1 pro	0.70	0.66	0.60	0.32	0.61	0.70	0.56
31	2+6+8 again	o1 pro	0.93	0.91	0.82	0.66	0.77	0.95	0.76
32	2+6+8	03-mini DR	-0.04	-0.24	-0.20	-0.25	0.26	-0.10	-0.08
33	2+6+8	03-mini-high DR	0.30	0.10	-0.21	0.12	0.21	0.25	0.22
34	2+6+8	o1-pro DR	0.33	0.23	0.13	-0.10	-0.12	0.15	0.24
35	2+6+8	03-mini	0.85	0.86	0.84	0.83	0.66	0.89	0.88
36	2+6+8	03-mini-high	0.94	0.81	0.88	0.65	0.75	0.91	0.87
37	Basis	Grok 3	1.00	0.72	0.82	0.87	0.34	0.93	0.90
38	Bach-3	Grok 3	0.89	0.88	1.00	0.86	0.72	0.94	0.90
39	Teacher	Grok 3	0.77	0.75	0.62	0.79	0.63	0.85	0.82
40	20 points	Grok 3	0.90	0.85	0.86	0.93	0.84	0.97	0.97
41	2+3+4+criteria	Grok 3	0.91	0.90	0.76	0.95	0.79	0.92	0.92
42	Rigthness	Grok 3	0.90	0.75	0.89	0.91	0.93	0.95	0.97
43	No sub-questions	Grok 3						0.78	0.85

44	Answering model	Grok 3	0.94	0.78	0.68	0.86	0.81	0.92	0.87
45	2+6+8	Grok 3	0.96	0.92	0.92	0.90	0.87	0.97	0.97
46	Improv. Potent.	Grok 3	0.92	0.75	0.87	0.92	0.80	0.91	0.89
47	Basis	4.5	0.92	0.77	0.76	0.70	0.55	0.86	0.81
48	Bach-3	4.5	0.79	0.89	0.75	0.94	0.78	0.89	0.85
49	Teacher	4.5	0.88	0.78	0.79	0.91	0.61	0.91	0.84
50	20 points	4.5	0.89	0.83	0.78	0.89	0.88	0.96	0.92
51	2+3+4+criteria	4.5	0.77	0.85	0.67	0.91	0.85	0.91	0.89
52	Rigthness	4.5	0.88	0.86	0.85	0.88	0.89	0.93	0.88
53	No sub-questions	4.5						0.96	0.98
54	Answering model	4.5	0.92	0.96	0.91	0.90	0.93	0.97	0.92
55	2+6+8	4.5	0.91	0.95	0.91	0.87	0.90	0.95	0.92
56	Improv. Potent.	4.5	0.75	0.87	0.79	0.68	0.87	0.89	0.87
57	2+6+8	Perpl. 4.3	0.65	0.68	0.45	0.32	0.54	0.72	0.64

Table A4. Correlation grading ChatGPT run 4 versus run 3

		Model	ScoreQ1	ScoreQ2	ScoreQ3	ScoreQ4	ScoreQ5	Total	Rank
1	Basis	4	0.54	0.36	-0.03	0.91	0.39	0.64	0.65
2	Bach-3	4	0.30	0.23	0.46	0.71	0.39	0.62	0.54
3	Teacher	4	0.45	0.42	0.38	0.72	0.35	0.55	0.55
4	20 points	4	0.30	0.35	0.17	0.40	0.56	0.39	0.56
5	2+3+4+criteria	4	0.08	0.44	0.50	0.51	0.43	0.51	0.57
6	Rigthness	4	0.57	0.39	0.67	0.90	0.70	0.89	0.89
7	No sub-questions	4	-	-	-	-	-	0.58	0.58
8	Answering model	4	0.62	0.23	0.23	0.64	-0.18	0.39	0.32
9	2+6+8	4	0.66	0.56	0.34	0.45	0.46	0.56	0.42
10	Improv. Potent.	4	0.33	0.60	0.48	0.34	0.12	0.51	0.49
11	Basis	4o	0.41	0.46	0.38	0.46	0.02	0.49	0.74
12	Bach-3	4o	0.68	0.45	0.53	0.59	0.44	0.63	0.54
13	Teacher	4o	0.68	0.45	0.53	0.59	0.44	0.63	0.54
14	20 points	4o	0.62	0.42	0.68	0.64	0.86	0.80	0.79
15	2+3+4+criteria	4o	0.71	0.42	0.79	0.81	0.64	0.76	0.76
16	Rigthness	4o	0.39	0.54	0.07	0.53	0.14	0.37	0.35
17	No sub-questions	4o	-	-	-	-	-	0.81	0.71
18	Answering model	4o	0.78	0.77	0.70	0.85	0.75	0.81	0.71
19	2+6+8	4o	0.87	0.74	0.35	0.72	0.76	0.78	0.42
20	Improv. Potent.	4o	0.40	0.64	0.40	0.89	0.46	0.73	0.66
21	Basis	o1 pro	0.41	0.58	0.62	0.56	0.39	0.63	0.53
22	Bach-3	o1 pro	0.53	0.36	0.92	0.56	0.43	0.84	0.72
23	Teacher	o1 pro	0.46	0.54	0.54	0.48	0.40	0.75	0.70
24	20 points	o1 pro	0.54	0.73	0.63	0.39	0.83	0.78	0.65
25	2+3+4+criteria	o1 pro	0.81	0.70	0.80	0.68	0.51	0.77	0.72
26	Rigthness	o1 pro	0.70	0.83	0.50	0.61	0.43	0.69	0.33
27	No sub-questions	o1 pro	-	-	-	-	-	0.68	0.72
28	Answering model	o1 pro	0.83	0.82	0.68	0.73	0.37	0.87	0.89
29	2+6+8	o1 pro	0.96	0.57	0.81	0.71	0.77	0.90	0.82

30	Improv. Potent.	o1 pro	0.72	0.66	0.41	0.09	0.43	0.63	0.49
31	2+6+8 again	o1 pro	0.94	0.68	0.76	0.78	0.67	0.91	0.74
32	2+6+8	03-mini DR	-0.31	-0.44	-0.20	0.09	-0.12	-0.21	-0.20
33	2+6+8	03-mini-high DR	-0.07	0.13	-0.25	0.02	-0.10	-0.03	0.05
34	2+6+8	o1-pro DR	-0.21	-0.02	-0.24	0.01	0.07	0.01	0.18
35	2+6+8	03-mini	0.84	0.82	0.84	0.74	0.85	0.87	0.85
36	2+6+8	03-mini-high	0.98	0.77	0.79	0.75	0.85	0.90	0.87
37	Basis	Grok 3	0.63	0.54	0.73	0.83	0.39	0.83	0.85
38	Bach-3	Grok 3	0.81	0.78	1.00	0.89	0.78	0.97	0.95
39	Teacher	Grok 3	0.70	0.78	0.53	0.82	0.63	0.81	0.81
40	20 points	Grok 3	0.94	0.93	0.88	0.86	0.90	0.98	0.98
41	2+3+4+criteria	Grok 3	0.98	0.98	0.86	0.91	0.75	0.97	0.97
42	Rigthness	Grok 3	0.91	0.72	0.67	0.75	0.85	0.89	0.85
43	No sub-questions	Grok 3						0.77	0.80
44	Answering model	Grok 3	0.92	0.95	0.88	0.96	0.79	0.95	0.89
45	2+6+8	Grok 3	0.89	0.91	0.94	0.86	0.85	0.92	0.94
46	Improv. Potent.	Grok 3	0.93	0.74	0.85	0.93	0.76	0.93	0.91
47	Basis	4.5	0.78	0.83	0.86	0.77	0.85	0.96	0.90
48	Bach-3	4.5	0.86	0.90	0.78	0.95	0.74	0.94	0.93
49	Teacher	4.5	0.67	0.83	0.69	0.91	0.72	0.88	0.84
50	20 points	4.5	0.87	0.85	0.79	0.90	0.80	0.93	0.85
51	2+3+4+criteria	4.5	0.83	0.87	0.71	0.85	0.83	0.89	0.83
52	Rigthness	4.5	0.91	0.87	0.88	0.83	0.89	0.93	0.86
53	No sub-questions	4.5						0.92	0.92
54	Answering model	4.5	0.85	0.94	0.82	0.92	0.88	0.95	0.93
55	2+6+8	4.5	0.94	0.93	0.83	0.85	0.82	0.94	0.94
56	Improv. Potent.	4.5	0.90	0.96	0.83	0.89	0.95	0.95	0.91
57	2+6+8	Perpl. 4.3	0.79	0.66	0.75	0.33	0.51	0.80	0.75

Table A5. Correlation grading ChatGPT run 5 versus run 4

		Model	ScoreQ1	ScoreQ2	ScoreQ3	ScoreQ4	ScoreQ5	Total	Rank
1	Basis	4	0.16	0.14	0.04	0.81	0.30	0.40	0.46
2	Bach-3	4	-0.19	0.53	0.43	0.52	0.11	0.31	0.29
3	Teacher	4	0.11	0.51	-0.04	0.51	0.28	0.19	0.20
4	20 points	4	-0.16	0.21	0.40	0.54	0.51	0.38	0.48
5	2+3+4+criteria	4	0.52	0.23	0.49	0.57	0.59	0.63	0.71
6	Rigthness	4	0.71	0.37	0.42	0.72	0.47	0.67	0.79
7	No sub-questions	4	-	-	-	-	-	0.60	0.40
8	Answering model	4	0.51	0.78	0.47	0.63	0.11	0.62	0.42
9	2+6+8	4	0.72	0.60	0.45	0.58	0.61	0.64	0.49
10	Improv. Potent.	4	0.54	0.44	0.05	0.56	0.22	0.42	0.40
11	Basis	4o	0.41	0.32	0.09	0.67	0.68	0.51	0.54
12	Bach-3	4o	0.22	0.37	0.45	0.65	0.52	0.64	0.63
13	Teacher	4o	0.22	0.37	0.45	0.65	0.52	0.64	0.63
14	20 points	4o	0.52	0.42	0.63	0.64	0.08	0.53	0.65
15	2+3+4+criteria	4o	0.49	0.62	0.26	0.56	0.40	0.54	0.54

16	Rigthness	4o	0.55	0.69	0.40	0.57	0.43	0.59	0.60
17	No sub-questions	4o	-	-	-	-	-	0.71	0.46
18	Answering model	4o	0.81	0.75	0.73	0.77	0.77	0.84	0.71
19	2+6+8	4o	0.89	0.72	0.61	0.71	0.61	0.80	0.67
20	Improv. Potent.	4o	0.20	0.24	0.50	0.56	0.35	0.40	0.53
21	Basis	o1 pro	0.55	0.69	0.47	0.29	0.19	0.53	0.41
22	Bach-3	o1 pro	0.51	0.53	0.84	0.45	0.78	0.78	0.63
23	Teacher	o1 pro	0.52	0.55	0.48	0.32		0.69	0.62
24	20 points	o1 pro	0.77	0.79	0.56	0.37	0.49	0.83	0.79
25	2+3+4+criteria	o1 pro	0.74	0.75	0.68	0.72	0.75	0.90	0.78
26	Rigthness	o1 pro	0.44	0.77	0.46	0.63	0.39	0.71	0.49
27	No sub-questions	o1 pro	-	-	-	-	-	0.66	0.65
28	Answering model	o1 pro	0.83	0.79	0.67	0.56	0.34	0.75	0.59
29	2+6+8	o1 pro	0.88	0.80	0.78	0.76	0.79	0.89	0.83
30	Improv. Potent.	o1 pro	0.80	0.70	0.54	0.59	0.26	0.74	0.64
31	2+6+8 again	o1 pro	0.86	0.68	0.84	0.79	0.65	0.91	0.79
32	2+6+8	03-mini DR	-0.07	0.02	-0.06	-0.02	-0.09	-0.08	0.06
33	2+6+8	03-mini-high DR	0.08	0.35	0.14	0.22	0.11	0.21	0.29
34	2+6+8	o1-pro DR	0.48	0.30	0.13	-0.13	0.04	0.42	0.37
35	2+6+8	03-mini	0.91	0.93	0.87	0.83	0.89	0.96	0.92
36	2+6+8	03-mini-high	0.88	0.81	0.89	0.76	0.79	0.88	0.82
37	Basis	Grok 3	0.68	0.67	0.60	0.78	0.45	0.74	0.75
38	Bach-3	Grok 3	0.92	0.80	0.89	0.92	0.67	0.94	0.91
39	Teacher	Grok 3	0.88	0.72	0.76	0.88	1.00	0.93	0.94
40	20 points	Grok 3	0.95	0.88	0.88	0.85	0.86	0.94	0.93
41	2+3+4+criteria	Grok 3	0.90	0.84	0.80	0.87	0.83	0.95	0.94
42	Rigthness	Grok 3	0.89	0.82	0.77	0.91	0.80	0.93	0.89
43	No sub-questions	Grok 3						0.89	0.90
44	Answering model	Grok 3	0.87	0.82	0.85	0.98	0.77	0.96	0.95
45	2+6+8	Grok 3	0.95	0.90	0.79	0.87	0.85	0.93	0.94
46	Improv. Potent.	Grok 3	0.96	0.93	0.95	0.96	0.85	0.97	0.95
47	Basis	4.5	0.79	0.87	0.86	0.89	0.72	0.95	0.93
48	Bach-3	4.5	0.59	0.93	0.69	0.89	0.81	0.89	0.86
49	Teacher	4.5	0.93	0.83	0.85	0.84	0.72	0.92	0.91
50	20 points	4.5	0.89	0.90	0.89	0.84	0.84	0.93	0.86
51	2+3+4+criteria	4.5	0.81	0.91	0.77	0.92	0.91	0.93	0.84
52	Rigthness	4.5	0.90	0.91	0.88	0.94	0.93	0.96	0.91
53	No sub-questions	4.5						0.91	0.87
54	Answering model	4.5	0.87	0.90	0.89	0.85	0.91	0.91	0.84
55	2+6+8	4.5	0.87	0.91	0.84	0.88	0.86	0.94	0.94
56	Improv. Potent.	4.5	0.93	0.90	0.85	0.86	0.88	0.92	0.87
57	2+6+8	Perpl. 4.3	0.90	0.89	0.87	0.86	0.85	0.94	0.93