

Is GPT-4 a reliable rater? Evaluating Consistency in GPT-4's Text Ratings ^{*}

Veronika Hackl[†] Alexandra Elena Müller[‡] Maximilian Sailer[§]
 Michael Granitzer[¶]

Abstract

This study investigates the consistency of feedback ratings generated by OpenAI's GPT-4, a state-of-the-art artificial intelligence language model, across multiple iterations, time spans and stylistic variations. The model rated responses to tasks within the Higher Education (HE) subject domain of macroeconomics in terms of their content and style. Statistical analysis was conducted in order to learn more about the interrater reliability, consistency of the ratings across iterations and the correlation between ratings in terms of content and style. The results revealed a high interrater reliability with ICC scores ranging between 0.94 and 0.99 for different timespans, suggesting that GPT-4 is capable of generating consistent ratings across repetitions with a clear prompt. Style and content ratings show a high correlation of 0.87. When applying a non-adequate style the average content ratings remained constant, while style ratings decreased, which indicates that the large language model (LLM) effectively distinguishes between these two criteria during evaluation. The prompt used in this study is furthermore presented and explained. Further research is necessary to assess the robustness and reliability of AI models in various use cases.

Keywords: GPT-4, Consistency, Higher Education, Feedback, Prompt Engineering, Large Language Models

1 Introduction

The integration of AI models, particularly LLMs, into the evaluation of written tasks within educational environments is a burgeoning trend. This trend is driven by the

^{*}This investigation served as a preliminary study preceding an extensive field study conducted as part of the BMBF-funded DeepWrite project at the University of Passau. The primary objective was to ascertain the consistency of GPT-4's assessments before their integration into authentic scenarios involving students within the realm of Higher Education. We extend our gratitude towards Johann Graf von Lambsdorff, Deborah Voss, and Stephan Geschwind for their contributions in designing the questions, sample solutions, and the field study associated with this investigation.

[†]University of Passau, Faculty of Social and Educational Sciences, Innstrasse 41, 94032 Passau, Germany, Veronika.Hackl@uni-passau.de. Main author.

[‡]University of Passau, Faculty of Law, Innstrasse 41, 94032 Passau, Germany. Research assistant.

[§]University of Passau, Faculty of Social and Educational Sciences, Innstrasse 41, 94032 Passau, Germany. Methodology contributor.

[¶]University of Passau, Faculty of Computer Science and Mathematics, Innstrasse 41, 94032 Passau, Germany. Hypotheses formulation contributor.

potential of these models to enhance learning outcomes by transforming traditional pedagogical methods.

As the use of these models becomes increasingly pervasive, it is imperative to thoroughly understand and quantify their reliability and consistency. *Elazar et al.* have defined consistency as 'the ability to make consistent decisions in semantically equivalent contexts, reflecting a systematic ability to generalize in the face of language variability' (*Elazar et al.*, 2021).

In the context of automated essay grading, inconsistent ratings could lead to unfair outcomes for students, undermining the credibility of the assessment process. Trust in the system 'is highly influenced by users' perception of the algorithm's accuracy. After seeing a system err, users' trust can easily decrease, up to the level where users refuse to rely on a system' (*Conijn et al.*, 2023, p.3). Similarly, in the context of personalized learning, unreliable predictions could result in inappropriate learning recommendations. Therefore, scrutinizing the consistency of AI models is a necessary step towards ensuring the responsible and effective use of these technologies in education (*Conijn et al.*, 2023).

GPT-4, through its emergent Automated Writing Evaluation capabilities, presents a significant advancement in overcoming traditional obstacles inherent in the evaluation of writing tasks. One such obstacle is discourse coherence, a fundamental aspect of writing that refers to the logical and meaningful connection of ideas in a text. In traditional manual grading, assessing discourse coherence can be a subjective and time-consuming process, often leading to inconsistencies in grading. However, GPT-4 with its advanced language understanding capabilities, can analyze the logical flow of ideas in a text, thereby providing a more objective and efficient evaluation of discourse coherence (*Naismith et al.*, 2023).

Feedback plays a crucial role in bridging the gap between a learning objective and the current level of competence and effective feedback, as outlined by Hattie and Timperley; encompassing three perspectives: Feed-Back, Feed-Up, and Feed-Forward. Feed-Back involves providing information about the current performance, Feed-Up clarifies the goals, and Feed-Forward gives guidance on how to improve. (*Hattie and Timperley*, 2007) 'Feedback is a core component of formative assessment processes and has been identified as a powerful factor influencing learning in various instructional contexts, including higher education' (*Narciss and Zumbach*, 2020). Regarding the development of writing skills, feedback on the text plays a crucial role, as it's nearly impossible to improve one's writing abilities without such feedback (*Schwarze*, 2021).

In the context of this study, the AI-generated feedback primarily focuses on the Feed-Back perspective, providing an analysis of the content and style produced by the student. In this scenario of analytic rating, 'the rater assigns a score to each of the dimensions being assessed in the task' (*Jonsson and Svingby*, 2007), in our case scores for style and content. The AI-generated feedback in this study is constructed to be adaptive and to assist the learner in figuring out options for improvement. This forms a contrast to non-adaptive or static feedback (e.g. the presentation of a sample solution) which is often used in HE scenarios due to its resource efficiency (*Sailer et al.*, 2023). Comprehensive feedback, which includes not only a graded evaluation but also detailed commentary on the students' performance, has been shown to lead 'to higher learning outcomes than simple feedback, particularly in regard to higher order learning outcomes' (*der Kleij et al.*, 2015). To make the feedback comprehensive and adaptive, it is prompted to include comments on the students' performance as well as numerical ratings and advice on how to improve.

A key advantage of AI-generated feedback is its immediacy. As noted by Wood and Shirazi (2020), 'Prompt feedback allows students to confirm whether they

have understood a topic or not and helps them to become aware of their learning needs.’(Wood and Shirazi, 2020, p. 24). This immediacy, which is often challenging to achieve in traditional educational settings due to constraints such as class size and instructor workload, can significantly enhance the learning experience by providing students with timely and relevant feedback (Haughney *et al.*, 2020). Kortemeyer’s observation that ‘The system performs best at the extreme ends of the grading spectrum: clearly correct and clearly incorrect solutions are generally reliably recognized [...]’ (Kortemeyer, 2023) further underscores the potential of AI models like GPT-4 in assisting human graders. This is particularly relevant in large-scale educational settings where human graders may struggle to consistently identify clearly correct or incorrect solutions due to the sheer volume of work.

2 Hypotheses

The stability of GPT-4’s performance is of significant interest given its potential implications for educational settings where the consistent grading of students’ work is paramount. In this investigation, GPT-4 was employed to assess responses to questions within the subject domain of macroeconomics with a focus on both the content and style of the responses. For content, the AI was prompted to evaluate how close the test answer is semantically to the sample solution. A sample solution inserted as demonstration in the prompt allows in-context learning and serves to control the quality of the output (Min *et al.*, 2022). For style, the AI was asked to check whether the language used in the test answer is appropriate for a HE setting and if the response is logically structured and plausible. The questions displayed different levels of complexity. The answers in the test set were created by the authors and subject domain experts, imitating the differing quality of student answers.

The primary objective of this study is to evaluate the consistency of ratings generated by GPT-4 across multiple iterations, time spans and variations. We demonstrate the agreement between raters and examine various dimensions of consistency. The term raters in our case refers to the different GPT-4 ratings. To provide a comprehensive analysis of GPT-4’s performance and application, we propose the following hypotheses:

- H1:** The ratings generated by GPT-4 are consistent across multiple iterations.
- H1.1:** The ratings generated by GPT-4 are consistent across different time spans, specifically within one week (short-term) and over several months (long-term).
- H1.2:** The complexity of the evaluated task does not influence the consistency of GPT-4’s ratings.
- H1.3:** Different types of feedback (e.g. style, content) do not affect the consistency of GPT-4’s performance.
- H2:** There is a significant correlation between the ratings for content and style in GPT-4’s evaluations.
- H3:** A certain type of prompt framework enables adaptation to new questions and answers, while maintaining consistency in the generated text.

While each of these hypotheses explores a distinct aspect related to GPT-4’s performance and application, collectively, they contribute to a comprehensive and multi-dimensional understanding of GPT-4’s potential and limitations in HE.

3 Methods

The methods section of this study is designed to provide a comprehensive overview of the research process, detailing the steps taken to address the hypotheses. The research process involves a series of statistical analyses, with the data collection process specifically designed to evaluate the consistency of a LLM in providing feedback and rating students' responses within the subject domain of macroeconomics.

3.1 Data Collection

The data collection phase was conducted over a 14-week period from April 2023 to July 2023, with API calls being made at different times and on different days to mimic a realistic usage scenario. The assumption underlying this approach is that the behavior of the model changes over time (Chen *et al.*, 2023). The API was called through a key within the Audience Response System classEx, which was used to interface with the AI model (Giamattei and Lambsdorff, 2019).

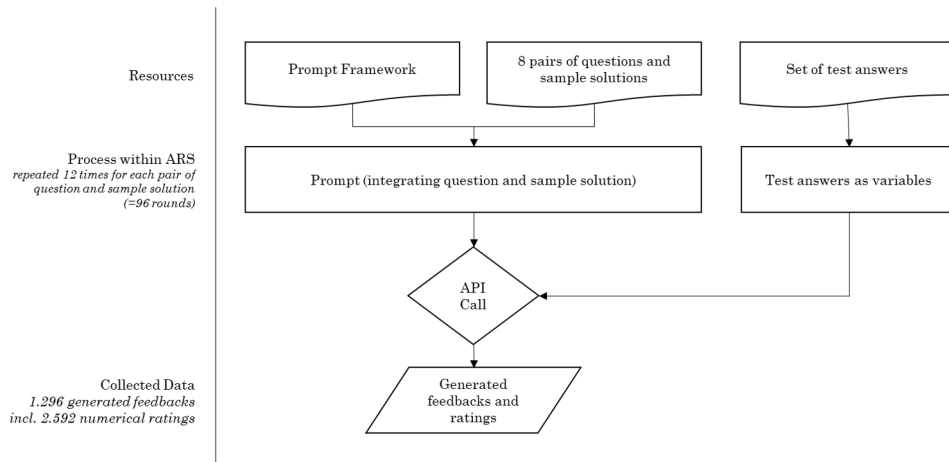


Figure 1: Flowchart: Generated texts

3.2 Prompt Framework and Test Responses

The first step in the research process involved the establishment of a prompt framework that serves as a universal structure within the context of this investigation. The goal was to insert new pairs of question and sample solutions without altering the consistency of the output, namely the LLM-generated feedback. Pairs of questions (Ruth and Murphy, 1988), along with corresponding sample solutions pertinent to macroeconomics, were prepared and integrated into the prompt framework. This integration set the stage for the model to assess students' responses and to generate feedback. First taxonomies aim at structuring prompt formulation approaches. The prompt used in this study would be a Level 4 on the Proposed Prompt Taxonomy TELeR (Turn, Expression, Level of Details, Role) by (Santu and Feng, 2023).

3.2.1 Establishing the prompt framework

The prompt framework was adapted to ensure consistency in the AI-generated feedback. A tight scaffold or rubric for rating to gain comparable results was used (Jonsson and Svingby, 2007). The system settings were adjusted to control the randomness of the model's responses, with a temperature setting of 0 used to minimize variability. (Si *et al.*, 2023; Schulhoff and Contributors, 2022). By forcing

the model into a deterministic behavior, it becomes more consistent in its outputs, while the chances to produce very good or very bad generations decrease. This is a brief documentation of the problems we encountered and the main changes we applied on the path to creating a prompt that works consistently for the use case:

Problem	Changes made in prompt
Output format varies	very clear instructions, ordinal numbers, examples
evaluations not strict enough	role prompting, clear evaluation criteria and application
robustness	shortening the prompt reduces calculation time, fewer outages
multiple identical inputs	different inputs can be tested at the same time, identical inputs must not be tested in one run as the parameters will then be passed incorrectly and/or the result is homogeneous
informal address with 'Du'	giving a clear instruction in the prompt with example
show star symbols	add the symbol in the prompt

This is the final scheme of the prompt framework used for the collection of data (shortened and translated, original language: German):

Element/Function	Prompt Formulation
Role Prompting	You are a professor of macroeconomics and you pose this question to your students:
Variable	<Insert Question here>
Task Description	You evaluate the student's response based on the sample solution using the criteria of content and style, and provide suggestions for improvement. This is the sample solution. It is clearly structured and builds the argument coherently. This solution is both correct in terms of content and very good in terms of style. It would receive 5 out of 5 stars for content and style. Sample solution:
Variable	<Insert sample solution here>
Stepwise Task Description	Please evaluate the student's response based on the sample solution in three steps.
Set Behavior	Here are some general tips for evaluation: Good feedback is honest and motivating. Always address the student directly using "you," for example: "Your response." Explain or mention the relevant points you are referring to.
Step 1: Evaluation of content (text feedback)	Step 1: Provide feedback on the content. Answer the following questions: Is the student's response correct in terms of content? Orient yourself to the meaning of the sample solution, but do not mention the sample solution. Are there any areas for improvement? Use a maximum of 2 sentences for this feedback.
Step 2: Evaluation of style (text feedback)	Step 2: Provide feedback on the style: Is the language used by the student appropriate for the field of study? Is the response logically structured and does the argumentation make sense? Are there any areas for improvement? Use a maximum of 2 sentences for this feedback.
Step 3: Evaluation (numeric feedback)	Step 3: Evaluate the content and style of the response on a scale of 1 to 5 stars. The rating is based on the feedback on content and style. 1 star indicates a very poor performance. 5 stars indicate a very good performance. Only display the following for Step 3: Content: Number of stars (Please also provide the number of stars as a numeral in parentheses) Style: Number of stars (Please also provide the number of stars as a numeral in parentheses)
Set Format	You provide a concise evaluation divided into 1. to 3.. Always display the stars as follows: star. Output:

3.2.2 Test Responses

Following the establishment of the prompt framework, test responses were created by domain experts to mimic potential student answers to the given questions. The

test set included a variety of responses, ranging from very good answers to nonsense answers and potential prompt injections, to ensure a comprehensive evaluation of the model’s performance (Liu *et al.*, 2023). An initial set of ten test responses was prepared for the first question. Based on our experience from this initial set, we expanded the test response set to 14 for the subsequent seven questions. We included prompt injections in the test set of answers to fine-tune the prompt and increase the probability to withstand potential attacks (Perez and Ribeiro, 2022).

Type Number	Type of Test Answer
1	copy of sample solution
2	correct, bullet points
3	nonsense
4	very good answer
5	opposite of sample solution, but good style
6	solid answer
7	incorrect, average style
8	correct, slightly informal style
9	average answer
10	bad content, bad style
11	only symbols
12	copy paste of question
13	prompt injection 1
14	prompt injection 2

Table 3: Scheme of test set of answers

3.2.3 Format of the generated feedback

The AI model generated feedback texts for each student response, accompanied by two ratings on a scale of 1 to 5 - one for content ("Inhalt") and one for style ("Stil"). This process was repeated eleven times for each question, resulting in a robust dataset comprising 2.596 generated ratings.

This is the format of the AI-generated feedback:

<p>1. Your answer is factually correct and aligns with the essential points of the model solution. You have well explained the impacts of the high population growth rate on the level of the domestic product and its growth.</p> <p>2. Stylistically, your answer is well-formulated, and the argumentation is logically structured. However, you could explicitly mention the growth model to fully answer the question.</p> <p>3. Content: ★★★★★ (5) Style: ★★★★★ (4)</p>

3.2.4 Changing styles

We further sought to investigate the effects of stylistic changes on the evaluation of content. To serve this goal, the test set of answers was stylistically transformed,

keeping the content. Each original student response was rephrased in the style of Warren G., a US-American rapper, with the intention of maintaining the semantic content while changing the stylistic presentation. The corresponding ratings for content and style were again generated through GPT-4. To assess whether the rephrasing affected the content and style ratings, paired t-tests were conducted comparing the original and rephrased ratings for each response. Additionally, the mean and standard deviation of the rating differences were calculated to provide an overview of the magnitude and variation of the changes. This is an illustrative example of an answer out of the original test set (1) and a rephrased version (2). Both texts were translated, the original text is German. The original text, the rephrased text and the corresponding ratings for these examples were as follows:

Original answer, Feedback and Rating

(1) Macroeconomics as engineering starts from human moods and habits. Negative expectations or moods can lead to a decline in domestic production. This is due to fewer expenses being incurred. To reverse this trend, the government should increase its spending. As a result, the domestic product becomes more stable, and no underemployment occurs.

1. Your answer is substantively correct and covers the essential points of the question. However, you could delve a bit more into the role of the government and why it should act expansively.
2. The style of your answer is appropriate, and the argument is logically structured. However, you could try to elucidate the connections a bit more clearly and formulate the reasons more precisely.
3. Content: ★★★★★ (4)
Style: ★★★★★ (4)

Rephrased answer, Feedback and Rating

(2) Macroeconomics, baby, is like fine-tuning an engine block - it's all about the vibes and routine of the people. If folks have a negative vibe or their attitude isn't positive, it can delay the economic performance, the GDP, because they're simply spending less dough. But you know what, the government can change the game if they themselves start splashing more cash. This keeps the GDP in a stable zone, and we don't have people who have to sit at home twiddling their thumbs.

1. Your answer is substantively correct as you highlight the impacts of moods and expectations on domestic production and the role of government spending. However, you could define and explain the terms "domestic product" and "underemployment" more precisely.
2. Your style is informal and could be considered inappropriate in an academic context. Try to use a more formal language and structure your argument more clearly.
3. Content: ★★★ (3)
Style: ★★ (2)

3.3 Statistical Analysis

3.3.1 Intraclass Correlation Coefficient

The Intraclass Correlation Coefficient (ICC) is a statistical measure in order to assess the level of agreement or consistency among raters. A perfect ICC score of 1

indicates perfect agreement or consistency among raters, while a score of 0 indicates no agreement nor consistency. ICC estimates and their 95% confident intervals were calculated using RStudio based on a mean-rating, absolute-agreement, 2-way mixed-effects model. For the decision which ICC calculation to use the flowchart proposed by Koo et.al was used. The type of the reliability study is 'inter-rater reliability'. We assign the different iterations of GPT-4 the role of different raters and assume that the same set of raters (GPT-4 at different points of time) rates all subjects. The model chosen is the two-way mixed effects model as we assume to have a specific sample of raters. The model type decided for is based on the mean of multiple raters. Both the model definitions "absolute agreement" and "consistency" were chosen. This results in the 2-way mixed-effects model. The caveat in the ICC model chosen in the analysis is that it only represents the reliability of the specific raters involved in this experiment (Koo and Li, 2016). As Generative AI remains a "black box" system, this was considered to be the most suitable model (Cao *et al.*, 2023).

The extracted numerical ratings from the feedback texts formed the dataset for the statistical analyses and were utilized to calculate the ICC, providing a measure of the consistency of the ratings generated by the AI model.

3.3.2 Correlation Analysis and Rating Differences

In order to answer H2, a correlation analysis was conducted. This analysis involved calculating the correlation coefficient between the content and style ratings generated by the AI model. The correlation coefficient provides a measure of the strength and direction of the relationship between the content and style ratings, thereby providing insights into the model's grading criteria. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In this study, the skewness of the rating distributions was calculated to examine the symmetry of the data. The purpose of this analysis was to evaluate the extent to which the ratings deviated from a normal distribution.

4 Results

The results section of this study presents the findings from the statistical analyses conducted to address the hypotheses. The analyses include the computation of Intraclass Correlation Coefficients (ICCs), skewness measures for content and style ratings, and a correlation analysis between content and style ratings.

4.1 Intraclass Correlation Coefficients

The tables 4 and 5 present the ICCs for ratings on content (Inh) and style (Stil) based on two different measurements. Table 4 reports ICCs from the initial ten measurements conducted between April and June 2023. The ICC values for both absolute agreement and consistency for content and style are extremely high (0.999), suggesting almost perfect agreement and consistency among raters. The 95% confidence intervals (CI) are as well tight, ranging from 0.998 to 0.999, indicating if the study was replicated, the true ICC would be expected to fall within this range 95% of the time. The F-tests are significant ($p < 0.001$), providing statistical evidence that the raters are reliably consistent and in agreement with each other in their ratings.

Table 5 reports ICCs from a control measurement. The ratings were obtained from two raters: the first being an average rating compiled from ten raters across the

months of April to June, and the second being a single rater evaluating in July. The result shows lower ICC values of 0.944 for both Inh and Stil. While these are still high values indicating good agreement, they are not as high as the ICC values in table 4. This implies that while a robust agreement persists between the mean rating and the July rater, it is not as pronounced as the concordance among the ten raters. This inference suggests a temporal evolution in the model’s behavior, necessitating diligent continuous assessment for its utilization in educational tasks.

ICC Type	ICC Value	95% CI	F-Test
Absolute agreement (Inh)	0.999	0.999 - 0.999	F(107,971) = 1332, p < 0.001
Absolute agreement (Stil)	0.999	0.998 - 0.999	F(107,971) = 689, p < 0.001
Consistency (Inh)	0.999	0.999 - 0.999	F(107,963) = 1332, p < 0.001
Consistency (Stil)	0.999	0.998 - 0.999	F(107,963) = 689, p < 0.001

Table 4: Reporting of Intraclass Correlation Coefficients (ICC)

ICC Type	ICC Value	95% CI	F-Test
Absolute agreement (Inh)	0.944	0.918 - 0.962	F(107,108) = 17.8, p < 0.001
Absolute agreement (Stil)	0.944	0.918 - 0.962	F(107,108) = 17.8, p < 0.001
Consistency (Inh)	0.944	0.918 - 0.962	F(107,107) = 17.8, p < 0.001
Consistency (Stil)	0.944	0.918 - 0.962	F(107,107) = 17.8, p < 0.001

Table 5: Reporting of Intraclass Correlation Coefficients (ICC) (mean rating of 10 raters from April to June, contrast rating of July)

4.2 Correlation between Content and Style Ratings

The relationship between the average content (Inh) and style (Stil) ratings was examined to assess the interplay between these two dimensions of evaluation. A correlation analysis was conducted, yielding a correlation coefficient of 0.87. This high value indicates a strong positive relationship between content and style ratings, suggesting that responses rated highly in terms of content were also likely to receive high style ratings, and vice versa.

This strong correlation underscores the interconnectedness of content and style in the evaluation process, suggesting that the AI model does not distinctly separate these two aspects but rather views them as interrelated components of a response’s overall quality. When the student answers were rephrased in a different style, we found that the average difference in content ratings before and after rephrasing was approximately 0.056 (stars rating), with a standard deviation of around 1.33. The paired t-test revealed no significant difference in content ratings between the original and rephrased responses ($t = 0.434$, $p = 0.665$). In terms of style ratings, the average difference before and after rephrasing was approximately 0.241, with a standard deviation of around 1.37. The paired t-test suggested a marginally significant difference between the original and rephrased style ratings ($t = 1.813$, $p = 0.073$).

The skewness of the content and style ratings was calculated to assess the distribution of these ratings. A positive skewness value indicates right-skewness, while a negative value indicates left-skewness. In this study, the positive skewness values for content suggest that the AI model tended to give higher ratings for content (see Table 6). Conversely, the majority negative skewness values for style suggest a

left-skewness, indicating that the model was more critical in its ratings for style (see Table 7).

These skewness values provide insights into the AI model’s rating tendencies. The right-skewness for content ratings suggests that the AI model may be more lenient in its content evaluations or that the student responses were generally of high quality. The left-skewness for style ratings, on the other hand, suggests that the AI model may have stricter criteria for style or that the style of the student responses varied more widely. These insights can inform future refinements of the AI model to ensure more balanced and fair evaluations.

Rater	Skewness
1_Inh	0.107009
2_Inh	0.080385
3_Inh	0.094007
4_Inh	0.116521
5_Inh	0.076956
6_Inh	0.096934
7_Inh	0.126752
8_Inh	0.089091
9_Inh	0.094007
10_Inh	0.090488
11_Inh	0.299014

Table 6: Skewness for Content Ratings

Rater	Skewness
1_Stil	-0.037198
2_Stil	-0.043986
3_Stil	0.029177
4_Stil	-0.017839
5_Stil	-0.047688
6_Stil	0.000873
7_Stil	-0.040248
8_Stil	-0.050956
9_Stil	-0.013981
10_Stil	-0.017839
11_Stil	-0.147365

Table 7: Skewness for Style Ratings

5 Discussion

The findings of this study provide insights into the potential of AI models, specifically GPT-4, in evaluating student responses in the context of macroeconomics.

- The high ICC values for both content and style ratings suggest that the AI model was able to consistently apply well-defined evaluation criteria at different points of time and with variations of style and content.
- The ICC values were lower when calculated with another set of feedbacks generated after a timespan of several weeks.
- The high level of concurrence between ratings underlines the dependability of the evaluation method employed in this study.
- The positive correlation between content and style ratings underscores the interconnectedness of content and style in the evaluation process.
- Rephrasing the answers stylistically did not significantly affect the content ratings, implying that GPT-4 was able to separate content from style in its evaluations.
- The ICC values show that forcing GPT-4 into a deterministic behavior through prompt- and system settings works.

It is important to note the limitations of AI models as their application in educational settings is not free of challenges. As stated in this paper, the ICC values differ for ratings at different points of time. There are variations in consistency for different levels of question complexity. Other limitations are being mentioned in OpenAI’s technical report on GPT-4: AI models can sometimes make up facts, double-down on incorrect information, and perform tasks incorrectly (OpenAI, 2023). Another challenge is the "black box" problem, as discussed by (Cao *et al.*, 2023).

This refers to the lack of transparency and interpretability of AI models, which can hinder their effective use in educational settings. Further research is needed to address this issue and enhance the transparency and interpretability of AI models.

Despite these challenges, there are promising avenues for enhancing the capabilities of AI models. The provision of feedback to macroeconomics students can be characterized as an emergent capability of the AI model. Emergence is a phenomenon wherein quantitative modifications within a system culminate in qualitative alterations in its behavior. This suggests that larger-scale models may exhibit abilities that smaller-scale models do not, as suggested by (Wei *et al.*, 2022). However, a direct comparison with GPT-3.5 is needed to substantiate this claim. The potential of AI models in providing feedback can be further enhanced by improving their "Theory of Mind" or human reasoning capabilities, as suggested by (Moghaddam and Honey, 2023). This could lead to more nuanced and contextually appropriate feedback, thereby enhancing the learning experience of students. Furthermore, the study by (Fu *et al.*, 2023) points to the potential of using AI models to audit generative AI. This opens up new avenues for ensuring the quality and reliability of AI-generated content. Above that, the use of smaller models should be encouraged (Bursztyn *et al.*, 2022) as well as the idea to evaluate AI-generated feedbacks either by a human rater or an AI before shown to the student (Perez and *et al.*, 2022).

In conclusion, while the results of this study are encouraging, they underscore the need for further research to fully harness the potential of AI models in educational settings. Future studies should focus on addressing the long-term performance, but also the limitations of AI models and exploring ways to enhance their reliability, transparency, and interpretability.

References

- BURSZTYN, V., DEMETER, D., DOWNEY, D. and BIRNBAUM, L. (2022). Learning to perform complex tasks through compositional fine-tuning of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 1676–1686.
- CAO, Y., LI, S., LIU, Y., YAN, Z., DAI, Y., YU, P. S. and SUN, L. (2023). A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. [arXiv:2303.04226](https://arxiv.org/abs/2303.04226).
- CHEN, L., ZAHARIA, M. and ZOU, J. (2023). How is ChatGPT’s behavior changing over time? [arXiv:2307.09009](https://arxiv.org/abs/2307.09009).
- CONIJN, R., KAHR, P. and SNIJDERS, C. (2023). The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. *Journal of Learning Analytics*, **10** (1), 37–53.
- DER KLEIJ, F. M. V., FESKENS, R. C. W. and EGGEN, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students’ learning outcomes: A meta-analysis. *Review of Educational Research*, **85** (4), 475–511.
- ELAZAR, Y., KASSNER, N., RAVFOGEL, S., RAVICHANDER, A., HOVY, E., SCHÜTZE, H. and GOLDBERG, Y. (2021). Measuring and Improving Consistency in Pretrained Language Models. [arXiv:2102.01017](https://arxiv.org/abs/2102.01017).
- FU, J., NG, S. K., JIANG, Z. and LIU, P. (2023). GPTScore: Evaluate as You Desire. [arXiv:2302.04166](https://arxiv.org/abs/2302.04166).
- GIAMATTEI, M. and LAMBSDORFF, J. G. (2019). classEx – an online tool for lab-in-the-field experiments with smartphones. *Journal of Behavioral and Experimental Finance*, **22**, 223–231.
- HATTIE, J. and TIMPERLEY, H. (2007). The power of feedback. *Review of Educational Research*, **77** (1), 81–112.
- HAUGHNEY, K., WAKEMAN, S. and HART, L. (2020). Quality of feedback in higher education: A review of literature. *Educ. Sci.*, **10** (3), 60.
- JONSSON, A. and SVINGBY, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, **2** (2), 130–144.
- KOO, T. K. and LI, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, **15** (2), 155–163.
- KORTEMAYER, G. (2023). Can an AI-tool grade assignments in an introductory physics course? [arXiv:2304.11221](https://arxiv.org/abs/2304.11221).
- LIU, Y., DENG, G., LI, Y., WANG, K., ZHANG, T., LIU, Y., WANG, H., ZHENG, Y. and LIU, Y. (2023). Prompt Injection attack against LLM-integrated Applications. [arXiv:2306.05499](https://arxiv.org/abs/2306.05499).
- MIN, S., LYU, X., HOLTZMAN, A., ARTETXE, M., LEWIS, M., HAJISHIRZI, H. and ZETTLEMOYER, L. (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? [arXiv:2202.12837](https://arxiv.org/abs/2202.12837).

- MOGHADDAM, S. R. and HONEY, C. J. (2023). Boosting Theory-of-Mind Performance in Large Language Models via Prompting. [arXiv:2304.11490](#).
- NAISMITH, B., MULCAIRE, P. and BURSTEIN, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Toronto, Canada: Association for Computational Linguistics, pp. 394–403.
- NARCISS, S. and ZUMBACH, J. (2020). *Formative Assessment and Feedback Strategies*, Cham: Springer International Publishing, pp. 1–28.
- OPENAI (2023). GPT-4 Technical Report. [arXiv:2303.08774](#).
- PEREZ, E. and ET AL. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. [arXiv:2212.09251](#).
- PEREZ, F. and RIBEIRO, I. (2022). Ignore Previous Prompt: Attack Techniques For Language Models. [arXiv:2211.09527](#).
- RUTH, L. and MURPHY, S. M. (1988). *Designing Writing Tasks for the Assessment of Writing*.
- SAILER, M., BAUER, E., HOFMANN, R., KIESEWETTER, J., GLAS, J., GUREVYCH, I. and FISCHER, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers’ diagnostic reasoning in simulation-based learning. *Learning and Instruction*, **83**, 101620.
- SANTU, S. K. K. and FENG, D. (2023). TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks. [arXiv:2305.11430](#).
- SCHULHOFF, S. and CONTRIBUTORS, C. (2022). Learn Prompting.
- SCHWARZE, C. (2021). Feedbackpraktiken im Schreibcoaching: Texte besprechen in der Hochschullehre. *Coaching Theor. Prax.*, **7**, 117–134.
- SI, C., GAN, Z., YANG, Z., WANG, S., WANG, J., BOYD-GRABER, J. and WANG, L. (2023). Prompting GPT-3 To Be Reliable. [arXiv:2210.09150](#).
- WEI, J., TAY, Y., BOMMASANI, R., RAFFEL, C., ZOPH, B., BORGEAUD, S., YOGATAMA, D., BOSMA, M., ZHOU, D., METZLER, D., CHI, E. H., HASHIMOTO, T., VINYALS, O., LIANG, P., DEAN, J. and FEDUS, W. (2022). Emergent Abilities of Large Language Models. [arXiv:2206.07682](#).
- WOOD, R. and SHIRAZI, S. (2020). A systematic review of Audience Response Systems for Teaching and Learning in Higher Education: The Student Experience. *Computers & Education*, **153**.