frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 05 December 2023
DOI 10.3389/feduc.2023.1272229

OPEN ACCESS

EDITED BY
Martina Rau,
University of Wisconsin-Madison, United States

REVIEWED BY
Richard Segall,
Arkansas State University, United States
Oluwatosin Ogundare,
California State University, San Bernardino, United States

*CORRESPONDENCE
Veronika Hackl
veronika.hackl@uni-passau.de

RECEIVED 03 August 2023
ACCEPTED 13 November 2023
PUBLISHED 05 December 2023

CITATION
Hackl V, Müller AE, Granitzer M and Sailer M (2023) Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings.
Front. Educ. 8:1272229.
doi: 10.3389/feduc.2023.1272229

Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings

Veronika Hackl[1]*, Alexandra Elena Müller[2], Michael Granitzer[3] and Maximilian Sailer[1]

[1]Faculty of Social and Educational Sciences, University of Passau, Passau, Germany, [2]Faculty of Law, University of Passau, Passau, Germany, [3]Faculty of Computer Science and Mathematics, University of Passau, Passau, Germany

This study reports the Intraclass Correlation Coefficients of feedback ratings produced by OpenAI's GPT-4, a large language model (LLM), across various iterations, time frames, and stylistic variations. The model was used to rate responses to tasks related to macroeconomics in higher education (HE), based on their content and style. Statistical analysis was performed to determine the absolute agreement and consistency of ratings in all iterations, and the correlation between the ratings in terms of content and style. The findings revealed high interrater reliability, with ICC scores ranging from 0.94 to 0.99 for different time periods, indicating that GPT-4 is capable of producing consistent ratings. The prompt used in this study is also presented and explained.

KEYWORDS

artificial intelligence, GPT-4, large language model, prompt engineering, feedback, higher education

1 Introduction

The integration of AI models, particularly LLMs, into the evaluation of written tasks within educational settings is a burgeoning trend, driven by the potential of these models to enhance learning outcomes and transform traditional pedagogical methods. As the use of these models becomes increasingly pervasive, it is imperative to thoroughly understand and quantify the reliability and consistency of the outputs produced. Elazar et al. (2021) 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." 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 toward ensuring the responsible and effective use of these technologies in education (Conijn et al., 2023). Another obstacle is discourse coherence, a fundamental aspect of writing that refers to the logical and meaningful connection of ideas in a text. GPT-4 can analyse the logical flow of ideas in a text, providing an efficient evaluation of the coherence of the discourse (Naismith et al., 2023).

A key advantage of AI-generated feedback is its immediacy. As Wood and Shirazi (2020) noted, "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 (2023) observation that "The system performs best at the extreme ends of the grading spectrum: correct and incorrect solutions are generally reliably recognized [. . .]" 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 correct or incorrect solutions due to the sheer volume of work.

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, and significantly impacts learning across diverse educational settings, notably in higher education (Narciss and Zumbach, 2020). Regarding the development of writing skills, feedback on the text plays a crucial role, as it is nearly impossible to improve one's writing skills without such feedback (Schwarze, 2021). In the context of this study, the AI-generated feedback primarily focuses on the "Feed-Back" perspective (Hattie and Timperley, 2007), 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 help the learner determine 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 Higher Education (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 regarding higher-order learning outcomes" (der Kleij et al., 2015). To make the feedback comprehensive and adaptive, it is prompted to include comments on the student's performance, numerical ratings, and advice on how to improve.

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 student work is paramount. In this investigation, GPT-4 was used to assess responses to questions within the macroeconomics subject domain with a focus on both the content and the style of the responses. For content, the AI was prompted to evaluate how close the test response was semantically to the sample solution. A sample solution inserted as a demonstration on the prompt 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 was appropriate for an HE setting and if the response was logically structured and plausible. The responses in the test set were created by the authors and subject domain experts, imitating the differing quality of student responses.

The primary objective of this study is to evaluate the absolute agreement and consistency of the GPT-4 ratings in multiple iterations, time intervals, 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 periods, specifically within one week (short-term) and over several months (long-term).
H1.2: Different types of feedback do not affect the consistency of GPT-4's performance. In this context, types of feedback are categorized into two specific levels: content rating, which evaluates the substance of the work, and style rating, which assesses the stylistic quality of the written argumentation.
H2: There is a significant correlation between the content and style ratings in GPT-4's evaluations.

3 Methods

The research process involves a series of statistical analyses, with the data collection process specifically designed to evaluate the consistency of GPT-4 in providing feedback and rating students' responses within the subject domain of macroeconomics.

3.1 Data collection

The data collection phase was conducted over 14 weeks from April 2023 to July 2023, with API calls made at different times and on different days to mimic a realistic usage scenario (see Figure 1). 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).

The dataset consists of multiple variables aimed at evaluating the quality of written responses in a macroeconomic context. The key variables include:

- MZP, Prompt, StudAnt, TypNr, AntwortTyp: These columns provide contextual information about the task, the type of answer, and other qualitative aspects.
- 1_*Inh*, 1_*Stil*, 2_*Inh*, 2_*Stil*, . . . , 12_*Inh*, 12_*Stil*: These columns capture the Intraclass Correlation Coefficient (ICC) related data. Specifically, these columns contain ordinal ratings that evaluate the content ("Inh") and style ("Stil") of the responses across multiple feedback cycles. Ratings range from 1 to 5, with one being the lowest and five the highest. We collected 2.592 numerical ratings in the ICC related columns.

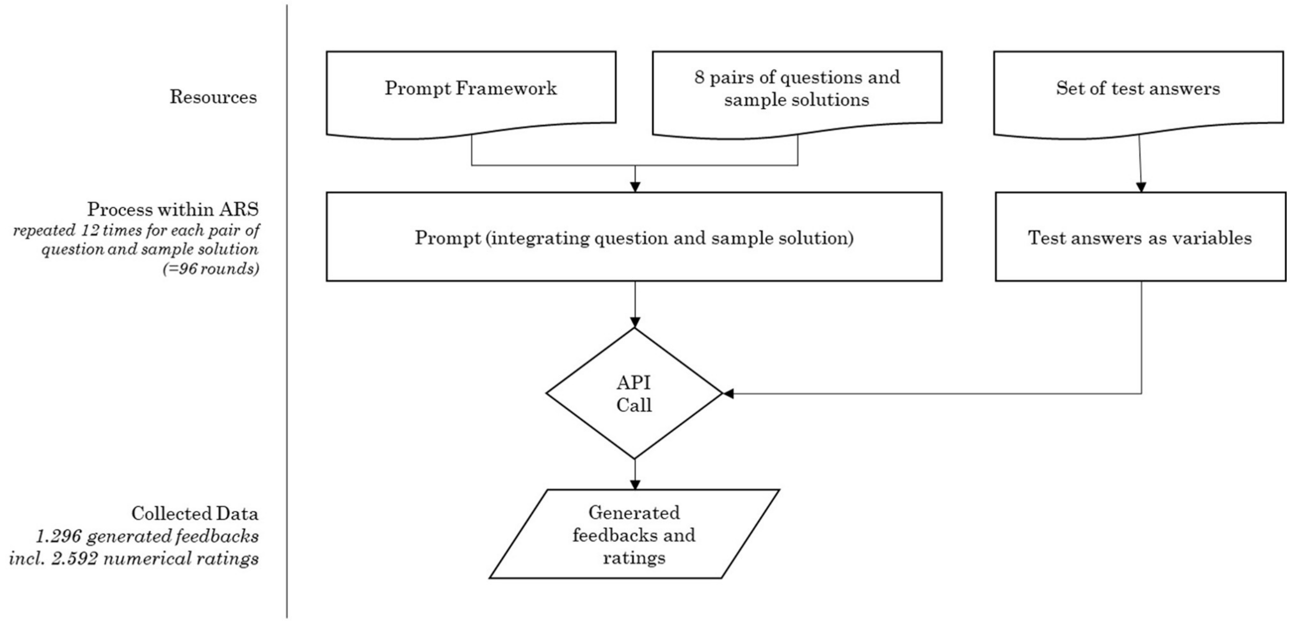

FIGURE 1
AI-generated feedback collection workflow.

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 questions 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 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).

TABLE 1 Problems encountered and changes made in prompt.

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 clear instruction in the prompt with example
Show star symbols	Add the symbol in the prompt

3.2.1 Establishing the prompt framework

The prompt framework was adapted to ensure consistency in AI-generated feedback. A tight scaffold was used to obtain comparable results (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 (Schulhoff and Community Contributors, 2022; Si et al., 2023). 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. Table 1 is a brief documentation of the problems we encountered and the main changes we applied to create a prompt that works consistently in the use case. Table 2 is the final scheme of the prompt framework used for data collection (shortened and translated, original language: German).

3.2.2 Test responses

Following the establishment of the prompt framework, domain experts created test responses to mimic potential student responses to the given questions. The test set (see Table 3) included a variety of responses, ranging from very good responses 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 with this initial set, we expanded the test response set to 14 for the

TABLE 2 Prompt framework.

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 structured and builds the argument coherently. This solution is correct in terms of content and very good in terms of style. It would receive five out of five 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 "Your response." Explain or mention the relevant points to which you are referring
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 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 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 feedback on content and style. 1 star indicates a very poor performance. Five stars indicate 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:

TABLE 3 Scheme of test set of answers.

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 in 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

subsequent seven questions. We included prompt injections in the test set of answers to fine-tune the prompt and increase the probability of surviving potential attacks (Perez and Ribeiro, 2022).

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 is the format of the AI-generated feedback:

1. Your answer is factually correct and aligns with the essential points of the model solution. You have explained the impacts of the high population growth rate on the level of the domestic product and its growth.

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.

3. Content: ★ ★ ★ ★ ★ (5)

Style: ★ ★ ★★ (4)

In addition, we sought to investigate the effects of stylistic changes on content evaluation. 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, to maintain the semantic content while changing the stylistic presentation. The corresponding ratings for content and style were again generated through GPT-4. Both texts were translated, the original text is in German. The original text (1), the rephrased text (2), 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 because fewer expenses are incurred. To reverse this trend, the government should increase its spending. As a result, the domestic product becomes more stable and there is no underemployment.

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 response is appropriate, and the argument is logically structured. However, you could try to elucidate the connections more clearly and formulate the reasons more precisely.
3. Content: ★ ★ ★★ (4)
   Style: ★ ★ ★★ (4)

10.3389/feduc.2023.1272229

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, and the GDP, because they're simply spending less dough. But you know what, the government can change the game if they 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 more formal language and structure your argument more clearly.
3. 3. Content: ⋆ ⋆ ⋆ (3)
   Style: ⋆⋆ (2)

3.3 Statistical analysis

Descriptive statistics reveal that the mean scores for both content and style generally range between 2.6 and 2.7. Most of the standard deviations are 0.00, indicating that for many observations, all raters provided the same score or rating for "Inh." The highest standard deviation observed for "Inh" is 1.21. Just like "Inh," many observations for "Stil" also have a standard deviation of 0.00. The highest standard deviation observed for "Stil" is 0.67.

3.3.1 Intraclass Correlation Coefficient

The Intraclass Correlation Coefficient (ICC) is a statistical measure to assess the level of agreement or consistency among the raters. A perfect ICC score of 1 indicates perfect agreement or consistency among the raters, while a score of 0 indicates no agreement or consistency. ICC estimates and their 95% confident intervals were calculated using RStudio based on a two-way mixed effect model with mean rating and absolute agreement. To make the decision on which ICC calculation to use, the flow chart proposed by Koo et al. was used. The type of 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 chosen model 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 two-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 numerical ratings extracted from the feedback texts formed the data set for the statistical analyses and were used 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

To answer H2, a correlation analysis was performed. 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 insight 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 of the statistical analyses performed 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

Table 4, 5 present the ICCs for the content ratings (Inh) and style ratings (Stil). Table 4 reports the ICCs from the measurements conducted between April and June 2023. The ICC values for 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 also tight, ranging from 0.998 to 0.999, indicating that if the study was replicated, it would be expected that the true ICC would 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. Ratings were obtained from two raters: the first was an average rating compiled from ten raters across April to June and the second was a single rater evaluation in July. The result shows lower ICC values of 0.944 for both Inh and Stil. Although these are still high values indicating good agreement, they are not as high as the ICC values in Table 4. This implies that while 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.

The results presented offer partial support for Hypotheses 1, 1.1, and 1.2. Although ratings demonstrate short-term consistency, ICC values exhibit a marginal decline over an extended period. The consistency of GPT-4's performance remained unaffected by the varying feedback types, whether content or style, thereby corroborating Hypothesis 1.2.

TABLE 4 Reporting of Intraclass Correlation Coefficients (ICC) (mean rating of 10 raters from April to June, contrast rating of July).

ICC type	ICC value	95% CI	F-test
Absolute agreement (Inh)	0.999	0.999–0.999	$F_{(107, 971)} = 1{,}332$, $p < 0.001$
Absolute agreement (Stil)	0.999	0.998–0.999	$F_{(107, 971)} = 689$, $p < 0.001$
Consistency (Inh)	0.999	0.999–0.999	$F_{(107, 963)} = 1{,}332$, $p < 0.001$
Consistency (Stil)	0.999	0.998–0.999	$F_{(107, 963)} = 689$, $p < 0.001$

ICC estimates and their 95% confident intervals were calculated using RStudio based on a two-way mixed effect model with mean rating, and absolute agreement. The type of reliability study is "inter-rater reliability."

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

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	$F_{(107, 107)} = 17.8$, $p < 0.001$

ICC estimates and their 95% confident intervals were calculated using RStudio based on a 2-way mixed effect model with mean rating, and absolute agreement. The type of reliability study is "inter-rater reliability."

4.2 Correlation between content and style ratings

The relationship between the average content (Inh) and style (Stil) ratings was examined to assess the interaction between these two dimensions of evaluation. A correlation analysis was conducted, which yielded 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 ~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 ~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 scores for content (see Table 6). On the contrary, the majority of 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.

TABLE 6 Skewness for content ratings.

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

Hypothesis 2, positing a significant correlation between content and style ratings in GPT-4's evaluations, is therefore confirmed.

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 in time and

TABLE 7 Skewness for style 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

with variations of style and content. This means that the model could serve as a reliable automated tool for grading or assessing student work, thereby reducing the workload on human evaluators.

- The ICC values were lower when calculated with another set of feedbacks generated after a timespan of several weeks. The decline in ICC values may suggest that the model's evaluations are susceptible to "drift." This is crucial in longitudinal educational studies where consistency over time is vital. It may necessitate periodic recalibration or updating of the model to maintain reliable assessments.
- The positive correlation between content and style ratings suggests 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. This is particularly important in educational settings where assessment rubrics may weight content and style differently. It allows for a more nuanced evaluation that doesn't conflate the two factors.
- The ICC values show that forcing GPT-4 into a deterministic behavior through prompt- and system settings works. This is essential for educational assessments where fairness and consistency are required. Such deterministic behavior allows for the standardization of assessments, making it easier to compare results across different time points or student populations.

It is important to note the limitations of AI models, as their application in educational settings is not free of challenges. The decline in ICC values over time raises concerns about the temporal consistency of GPT-4's evaluations, particularly since the same test set was used throughout the study. If the decline is due to model drift—a phenomenon where the model's performance changes due to evolving data or internal updates—this could compromise the reliability of long-term educational assessments. Though making the model deterministic may ensurevad consistency, it can also limit the model's ability to adapt to different styles or levels of student responses. In education, adaptability to diverse learning styles is essential. 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. This is also relevant in practice when someone knows many things but does not know how to express them. Above that, the use of smaller models should be encouraged (Bursztyn et al., 2022) as well as the idea to evaluate AI-generated feedback 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. A hybrid approach where AI-generated evaluations are reviewed by human educators to ensure both reliability and validity is highly recommended. 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.

Data availability statement

The datasets, detailed case processing summary, reliability statistics and full descriptive statistics will be provided upon request by the corresponding author.

Author contributions

VH: Writing – original draft, Writing – review & editing. AM: Writing – review & editing. MG: Writing – review & editing. MS: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The report has been funded by the German Federal Ministry of Education and Research (BMBF) under the project DeepWrite (Grant No. 16DHBKI059). The authors are responsible for the content of this publication.

Acknowledgments

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 HE. We extend our gratitude toward 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.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

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: Association for Computational Linguistics), 1676–1686. doi: 10.18653/v1/2022.findings-emnlp.121

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., et al. (2023). A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. *arXiv* [preprint]. doi: 10.48550/arXiv.2303.04226

Chen, L., Zaharia, M., and Zou, J. (2023). How is ChatGPT's behavior changing over time? *arXiv* [preprint]. doi: 10.48550/arXiv.2307.09009

Conijn, R., Kahr, P., and Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *J. Learn. Anal.* 10, 37–53. doi: 10.18608/jla.2023.7801

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. *Rev. Educ. Res.* 85, 475–511. doi: 10.3102/0034654314564881

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., et al. (2021). Measuring and improving consistency in pretrained language models. *arXiv* [preprint]. doi: 10.48550/arXiv.2102.01017

Giamattei, M., and Lambsdorff, J. G. (2019). classEx-an online tool for lab-in-the-field experiments with smartphones. *J. Behav. Exp. Finance* 22, 223–231. doi: 10.1016/j.jbef.2019.04.008

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Haughney, K., Wakeman, S., and Hart, L. (2020). Quality of feedback in higher education: a review of literature. *Educ. Sci.* 10, 60. doi: 10.3390/educsci10030060

Jonsson, A., and Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educ. Res. Rev.* 2, 30–144. doi: 10.1016/j.edurev.2007.05.002

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Kortemeyer, G. (2023). Can an AI-tool grade assignments in an introductory physics course? *arXiv* [preprint]. doi: 10.48550/arXiv.2304.11221

Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., et al. (2023). Prompt injection attack against LLM-integrated applications. *arXiv* [preprint]. doi: 10.48550/arXiv.2306.05499

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., et al. (2022). Rethinking the role of demonstrations: what makes in-context learning work? *arXiv* [preprint]. doi: 10.48550/arXiv.2202.12837

Moghaddam, S. R., and Honey, C. J. (2023). Boosting theory-of-mind performance in large language models via prompting. *arXiv* [preprint]. doi: 10.48550/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: Association for Computational Linguistics), 394–403. doi: 10.18653/v1/2023.bea-1.32

Narciss, S., and Zumbach, J. (2020). *Formative Assessment and Feedback Strategies* (Cham: Springer International Publishing), 1–28. doi: 10.1007/978-3-030-26248-8_63-1

OpenAI (2023). GPT-4 technical report. *arXiv* [preprint]. doi: 10.48550/arXiv.2303.08774

Perez, E. et al. (2022). Discovering language model behaviors with model-written evaluations. *arXiv* [preprint]. doi: 10.48550/arXiv.2212.09251

Perez, F., and Ribeiro, I. (2022). Ignore previous prompt: attack techniques for language models. *arXiv* [preprint]. doi: 10.48550/arXiv.2211.09527

Ruth, L., and Murphy, S. M. (1988). *Designing Writing Tasks for the Assessment of Writing*. London: Bloomsbury Academic.

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., et al. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. *Learn Instr.* 83, 101620. doi: 10.1016/j.learninstruc.2022.101620

Santu, S. K. K., and Feng, D. (2023). TELeR: a general taxonomy of LLM prompts for benchmarking complex tasks. *arXiv* [preprint]. doi: 10.48550/arXiv.2305.11430

Schulhoff, S., and Community Contributors (2022). *Learn Prompting*. Available online at: https://github.com/trigaten/Learn_Prompting

Schwarze, C. (2021). Feedbackpraktiken im schreibcoaching: texte besprechen in der hochschullehre. *Coaching Theor. Prax.* 7, 117–134. doi: 10.1365/s40896-020-00045-x

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., et al. (2023). Prompting GPT-3 to be reliable. *arXiv* [preprint]. doi: 10.48550/arXiv.2210.09150

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent abilities of large language models. *arXiv* [preprint]. doi: 10.48550/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. *Comput. Educ.* 153, 103896. doi: 10.1016/j.compedu.2020.103896