

Can GPT-4 Replace Human Examiners? A Competition on Checking Open-Text Answers

Abdullah Al Zubaer

`abdullah1.zubaer@uni-passau.de`

University of Passau

Michael Granitzer

University of Passau

Stephan Geschwind

University of Passau

Johann Graf Lambsdorff

University of Passau

Deborah Voss

University of Passau

Article

Keywords:

Posted Date: July 11th, 2024

DOI: <https://doi.org/10.21203/rs.3.rs-4431780/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 8th, 2025.
See the published version at <https://doi.org/10.1038/s41598-025-21572-8>.

Can GPT-4 Replace Human Examiners? A Competition on Checking Open-Text Answers

Abdullah Al Zubaer^{1,*}, Michael Granitzer¹, Stephan Geschwind², Johann Graf Lambsdorff², and Deborah Voss²

¹Department of Computer Science and Mathematics, University of Passau, Passau, Germany

²Department of Business Administration, Economics and Business Information Systems, University of Passau, Passau, Germany

*abdullahal.zubaer@uni-passau.de

ABSTRACT

Under what circumstances can GPT-4 be relied upon as a safe and efficient decision-maker? To address this question we explore the performance of GPT-4 as an examiner of answers to open text questions. We formulate questions and sample solutions in the subject area of macroeconomics and collect answers from cohorts of students. Subsequently, we conduct a fair competition between GPT-4 and human experts, employing their expertise to assess the quality of the answers. We observe that the substitution of GPT-4 for a human examiner does not decrease inter-rater reliability. We run checks on potential biases (whether GPT-4 prefers AI-generated or lengthy answers). We find no consistent evidence of such biases. Our findings are robust to tilting the competition to one side's advantage, by using inferior or advanced prompting strategies. These findings suggest a broad range of potential applications for GPT-4 in university education. Overall, our study cautiously supports the utilization of GPT-4 as an assistant for automated grading systems.

Introduction

Generative AI models have taken the world by storm. ChatGPT¹, for instance, surpassed the mark of 1 million users within five days of its launch in December 2022 and reached 100 million users within two months. Its rapid spread can be attributed to the high quality of human-like conversations. It has the potential of broad repercussions for society in general and the educational system in particular. Especially the fourth version of GPT (Generative Pre-trained Transformer)² launched in March 2023, is widely acknowledged for its ability to understand and produce text. One of GPT's fundamental properties is its capacity to perform in-context learning³ in which it "learns" from a prompt containing examples or instructions to perform specific tasks relying entirely on its linguistic capability⁴. This allows GPT models, particularly GPT-4⁵, to be highly adaptable in various scenarios despite being a generalized language model. GPT-4, therefore, has a multitude of potential use cases in the education system. It can be used both by students and teachers to assist or even replace them in writing and grading exams and papers. Accordingly, studies on the usage of generative AI models in education have mushroomed after the release of ChatGPT.

When it comes to using GPT as a student, Herbold et al.⁶ demonstrate the capability of GPT-3.5 and GPT-4 to write high-quality essays. In a large-scale project, they have more than 100 high school teachers compare the performance of essays written by advanced cohorts of high school students to essays written by GPT and find that the latter surpasses humans in terms of quality. Noy and Zhang⁷ conducted an experiment involving 453 professionals assigned to writing tasks. Participants were randomly divided, with half required to register for and use ChatGPT, based on GPT-3.5, for their tasks. This intervention substantially improved their productivity, decreasing the time taken and increasing output quality, as assessed by three evaluators. Similarly, Yeadon et al.⁸ find that short essays in Physics written by text-davinci-003 (predecessor of GPT-3.5) are on par with Physics students in their second year of university education and could easily be given the highest classification at UK universities. GPT-4 has been shown to perform remarkably well and, in some aspects, even better than humans in the United States Medical Licensing Examination⁹⁻¹¹. Stribling et al.¹² evaluate GPT-4's effectiveness in tackling graduate biomedical science tests and discover that it frequently surpasses the average student's performance, especially in responding to fill-in-the-blank, short-answer, and essay questions. However, its ability to interpret figures is weaker, and it occasionally exhibits behaviors indicative of plagiarism and hallucination. These findings suggest early evidence that generative AI technologies have the potential to substitute for students' efforts in academic settings.

We differ from this literature by taking the teacher's perspective: When it comes to using GPT as a teacher, several studies indicate a positive effect of using generative AI to complement teaching activities^{13,14}. Different from this literature, we investigate the potential of generative AI to replace – and not merely complement – teaching activities, in particular as a rater of

open-text answers. In that, we are more closely aligned to studies that compare the performance of human and AI teachers and tutors. For instance, Phung et al.¹⁵ find that GPT-4 is on par with human tutors in multiple programming education schemes but significantly needs to improve when grading a student's code. Nilsson and Tuvstedt¹⁶ run an experiment similar to our study in that they let GPT-4 evaluate student answers based on a prompt that includes the assignment, the grading criteria, the student submission and instructions on how to grade. They find that GPT-4 performs rather well in grading elementary programming assignments submitted by students of undergraduate computer science courses. GPT-4's overall accuracy was 75% compared to grades (pass or fail) given by teaching assistants (ground truth), hinting at the feasibility of adopting GPT-4 for grading. Different from our study design, there is however no level playing field between teaching assistants and GPT-4. While teaching assistants were "instructed to be kind and understanding", GPT-4 was prompted to "[act] as a very diligent and harsh teacher's assistant". On another downside, taking teacher assessment as ground truth is a strong assumption that disregards human fallibility, for example, due to subjective bias, cognitive limitations, fatigue, or inattention. The comparison between GPT-4 and fallible human examiners on a level playing field remains a research gap that still needs to be well addressed.

We thus contribute to the existing literature by taking a more systematic approach and comparing the performance of human examiners and GPT-4 on a level playing field. We achieve this by distinguishing between the role of the teacher, who provides a sample solution to an open-text question and the instructions, and that of the raters, who assess the answers. This allows us to use fallible human examiners and GPT-4 as our raters, provide them with identical instructions and compare their subsequent performance. We employ two methods to arrive at such comparisons, one involving relative ranks and another on absolute points awarded for student answers. Our quantitative approach involves comparing the inter-rater reliability between three human examiners and teams of two human examiners and GPT-4.

This approach also allows us to investigate whether GPT has a self-serving bias and prefers its own answers. Several studies have raised the question of whether GPT is prone to provide eloquent yet factually incorrect content^{17,18}. Such a tendency may concur with a propensity to regard persuasive but baseless assertions as true in open-text answers. In particular, such a bias might show if GPT is more inclined than human examiners to fall for 'eloquent bullshit', to prefer form over substance. In an educational context, this would be problematic as it disadvantages students who express their knowledge differently. If both students and examiners use GPT-like models for generating and grading answers, it could initiate a slippery slope. Answers generated by generative AI might be preferentially scored, thereby encouraging students to increasingly depend on these technologies for producing answers. This undermines the educational objective of fostering independent thinking and learning. Similarly, teachers might be misled by observing improvements in scores generated by GPT-like models, mistakenly interpreting them as genuine educational advancements. In the end, it would not be a generative AI producing output similar to that of humans; it would be humans aligning their output to that of generative AI.

To capture a potential confound to GPT's grading capabilities, we also look at whether GPT-4 has a bias towards longer text answers. The concern is that GPT is trained on a diverse internet corpus where detailed, comprehensive responses are often associated with higher quality information. It may then equate length with thoroughness or credibility, potentially favoring more extended, articulate responses even when they are not more accurate. We thus address the following two research gaps:

RQ1 How well does GPT-4 perform in ranking and assessing student answers as compared to fallible human examiners?

RQ2 Is GPT-4 biased by favoring GPT-generated answers or longer answers?

The answers to these questions allow us to draw cautious but far-reaching implications for the use of GPT-4 and similar AI models in the assessment and grading of student examinations in higher education.

Material and Methods

Data and Procedures

We address these research gaps based on student answers to open questions in an undergraduate course in macroeconomics collected prior to the conception of this study during exercise classes between April and July 2022 using classEx¹⁹. classEx is an online tool for interactive classroom experiments and exercises that allows students to anonymously participate using their personal mobile device (smartphone, tablet, laptop). Anonymity is guaranteed by programming protocols (e.g. data encryption, use of pseudo-IDs) when using classEx. Over the course of one semester, students were provided with six open questions that corresponded to the progress of the class. They had 5–10 minutes to write a text answer. Answers were written in German. Thereafter, 2–3 student answers were randomly drawn and shown on the lecturer screen. The lecturer discussed them openly based on a sample solution and selected the best of the displayed student answers to receive a prize of EUR 10.00 in a transparent and fair way. Hence, students were incentivized to write good answers. Participation was voluntary. Students were orally informed beforehand that their input would be stored anonymously and could be used for research purposes, to which they consented by participating in the exercise.

In total, 517 student answers were generated that way. The number of student answers per question varied strongly with the minimum being at 25 and the maximum at 126. Answers contain on average about 275 characters with large variations among students and questions. Further, they differ in their format, style and quality. An example of a question, sample solution and five student answers is given in Table 1 below (translated to English). It illustrates the complexity of the task. Assigning points and comparing the quality of answers requires a substantial knowledge of the underlying growth model, an identification of shortcomings and the steps of reasoning that students must undertake.

Question
How does the high population growth rate in Africa influence the level of the domestic product and its growth? Justify your answer with the help of the growth model!
Sample Solution
A high population growth rate, as is the case in many African countries, implies that the total capital stock is distributed among an increasing number of people. A low per capita capital stock results in a low level of per capita domestic product. However, the domestic product rises considerably due to the high population growth rate.
Student Answers
The high population growth rate in Africa influences the level of domestic product and its growth. This is confirmed by the correlation between birth rate and growth in per capita income. Causality can also run the other way round.
The higher the population growth rate, the lower the domestic product in Africa as the necessary per capita investments are higher.
A high population growth rate means, conversely, that the capital stock must now be divided among more people. A smaller per capita capital stock means that productivity falls and with it GDP. Therefore, the aim in an economy should be: $dN < dk$.
Growth rate n depends on the depreciation rate δ and the per capita capital stock k . If n increases, the per capita capital stock k automatically increases. A new steady state is formed in which the necessary investments are $>$ than the actual investments.
The high population growth rate in Africa has a negative impact on domestic product growth. As the necessary investments increase more strongly due to rapid population growth, the actual investments must increase all the more in order to achieve a steady state (necessary investments = actual investments).

Table 1. Example of Open Question, Sample Solution and 5 Student Answers

To compare the performance of three human examiners and GPT-4 in assessing student answers, we proceed as illustrated in Fig. 1: For each of the six questions, we randomly draw 10 sets of five student answers each (with replacement). Five answers per set keep comparisons cognitively manageable. We observed upfront that larger sets become too demanding to human examiners. Our choice of 10 sets per question ensures that each question is given equal importance, whilst limiting repeated selections of answers. To address RQ2 on the self-serving bias, we include further answers generated by GPT-3.5-turbo-instruct between March and May 2023. To achieve some variation in the quality and content of the GPT-generated answers, we varied the accuracy and relevance of the content provided to GPT. For full access to the prompts see Section S1 in the Supplementary Material. We set the temperature to zero and all other hyperparameters to the default values. The GPT-generated answers are randomly included in each set with a chance of 5/7. Their total number amounts to 39. Human examiners and GPT-4 are left ignorant about some of the answers being GPT-generated. They were led to believe that all the answers were student answers.

We provide the 10 sets along the open question, a respective sample solution and instructions to three human examiners and GPT-4. Their first task is to rank the five student answers in each set by quality based on the sample solution – from rank 1 (best answer of the set) to rank 5 (worst answer). Each rank can only be assigned once. The examiners work individually. Hence, for each set, we receive three rankings by each human examiner and one by GPT-4. Their second task is to award points for each answer based on the sample solution. In this task, the provided sample solution is amended by further information about the contents for which points are to be awarded – similar to solution templates used to correct exams. The maximum score differs depending on the question between 3 and 5. The sets are the same for both tasks and all four examiners. Task 1, the ranking, is always performed before Task 2, the point assessment. The human examiners performed both tasks between March and June 2023. GPT-4 performed the ranking task in June 2023 and the point assessment in August 2023.

Our human examiners are three student assistants employed at the Chair of Economic Theory who have successfully passed the course in macroeconomics with grades high above average. They can be considered trained experts in this field and are qualified to correct undergraduate exams in macroeconomics. The student assistants partook in the study not as regular experimental subjects but as employees of University of Passau. They were informed and consented that their work product would be used for the purposes of this study. GPT-4⁵, is accessed via chat completion API provided by OpenAI²⁰. The instructions provided to the human examiners are kept as similar as possible to the prompt provided to GPT-4 (see Section S2 of the Supplementary Material). This helps us level the playing field between the human examiners and GPT-4. We thus avoid any bias coming from different instructions or a prompt overfitting on the study. Nevertheless, we evaluate different prompting strategies for purposes of robustness in the Section *Robustness Checks and Extensions*.

Our entire study protocol has been reviewed and approved by the Research Ethics Committee of the University of Passau. We confirm that our study protocol adheres to all pertinent guidelines and regulations.

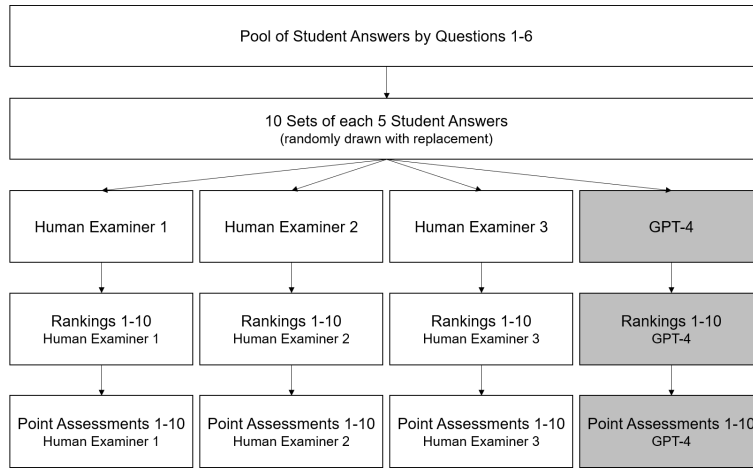


Figure 1. Procedure – Ranking and Point Assessment of Answers by Human Examiners and GPT-4

Experimental Design and Hypotheses

To address the question whether GPT-4 can substitute human examiners (RQ1), we employ a simple yet novel experimental design as illustrated in Fig. 2: We consider the ranking and point assessment by our three human examiners (HE) as the baseline scenario. Hence, the “All Human Team” serves as the ground truth. We compare this baseline scenario with three treatment scenarios. In each, one of the human examiners is replaced by GPT-4. These scenarios are classified as GPT-Human Team 1, 2, and 3. For comparison between baseline and treatment scenarios we use the inter-rater reliability (IRR) amongst the three examiners in each team. This novel approach mitigates potential biases and reduces noise that might be induced when basing the ground truth on one single examiner, as it is commonly done in annotation tasks or interactive assessments. Also, our approach is not prone to data-poisoning as the true rankings and scores have not been disclosed in the Web²¹.

This metric allows us to evaluate how consistent the members of each team are at ranking and assigning points. A high IRR implies that ratings are not significantly influenced by personal biases, errors, or subjective interpretations. They accurately capture the underlying reality. The metric thus offers a comparative measure of performance between All Human and GPT-Human Teams. Our novel design allows us to abandon simplifying assumptions on human examiners grading without errors. Instead, we can compare between examiners that have varying degrees of fallibility. A decrease in IRR when GPT-4 is introduced as the third examiner indicates that GPT-4’s judgment capability is less aligned with the remaining two human examiners than the replaced human examiner. GPT-4 would then perform worse than humans. In contrast, an increase in IRR demonstrates that GPT-4’s assessments perform better than those by humans.

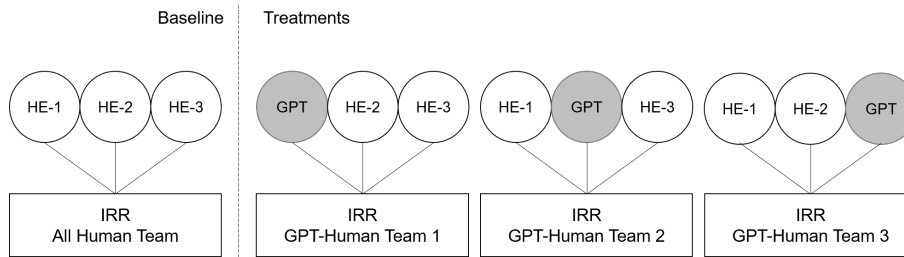


Figure 2. Experimental Design – Baseline and Treatment Scenarios

We investigate potential differences in the inter-rater reliabilities between baseline and treatment scenarios separately for the ranking and the point assessment task. This allows us to address RQ1 in its two relevant dimensions: the capability of GPT-4 to measure the relative quality of student answers as necessary to determine grades (ranking) as well as to distinguish between correct and incorrect content to award scores (point assessment). We expect the treatment effects to go in the same direction for both tasks since the challenges examiners face are similar: Ranking and point assessing answers to open questions requires an understanding of intricate reasoning, subtle nuances, and creative thinking that leave some room for inventive answers beyond the sample solution. Recent studies suggest that these challenges may be greater for GPT than for human examiners since

generative-AI models have shown to lack reasoning competencies^{5,15,18,22}. As GPT-4 reflects aggregate opinions present on the internet, we further asked GPT-4 itself to form expectations on its performance as compared to human examiners in our specific tasks. In line with the previous findings, GPT-4 replied: "In both tasks – ranking answers and awarding points – human examiners will likely outperform GPT-4 because they can better understand subtleties, context, and originality in student answers. While GPT-4 can handle structured tasks more effectively, its understanding of complex reasoning and unconventional solutions is limited." Similarly, we hypothesize that GPT-4, solely relying on the patterns it observed in its training data, cannot achieve the accuracy and comprehensiveness of the human mind:

H1 [Inter-rater reliability]: The All Human Team achieves a higher inter-rater reliability than teams that include GPT-4 (GPT-Human Teams).

Further, we are interested in the question of whether there is a self-serving bias in the rankings and point assessment of GPT-4 (RQ2). If a self-serving bias exists, GPT would rank and point assess the 39 GPT-generated answers systematically better than their human counterparts. This would pose a major problem for the use of GPT as an examiner since students who have their answers written by GPT-like models would receive better grades. There are two reasons why this bias might occur: First, GPT might be inclined to judge GPT-generated answers more favorably due to the linguistic proximity and correspondence with the training data. Second, and relatedly, GPT answers tend to be more sophisticated in terms of language than student answers. Yet Rudolph et al.¹⁸ emphasize that when it comes to critical, analytical thinking, AI lacks competence. If GPT prefers eloquent but less competent answers, this might likewise reflect in this bias. We hence hypothesize:

H2 [Self-serving bias]: GPT-4 judges GPT-generated answers more favorably than human examiners do.

Moreover, GPT-4 might be biased towards longer answers. First, this might be due to longer answers being more eloquent. Second, whereas the human mind recognizes repetitions of content and does not value them as additional input, GPT-4 might not be able to apply such critical assessment, confusing length with accuracy. Thus, we hypothesize:

H3 [Length bias]: GPT-4 judges long answers more favorably than human examiners do.

Prompting Strategy

For constructing the prompt, we followed the zero-shot learning technique³. It allows us to level the playing field between human examiners and GPT-4 and compare them in a one-to-one fashion without providing additional advantages to GPT-4³. Apart from some technical details, the prompt is thus entirely based on the instructions provided to our human examiners. We later abandon this setting and investigate whether other state-of-the-art prompt techniques demonstrate superior performance in contrast to zero-shot learning techniques (few-shot chain-of-thought prompting²³).

GPT-4 is accessed via chat completion API provided by OpenAI²⁰. For the hyperparameters, we set the temperature to zero to approximate a deterministic behavior of GPT-4. GPT-4 and other large language models are inherently non-deterministic, yet setting the temperature to zero allows control over GPT-4 behavior to a greater extent and increased replicability. OpenAI recently introduced new hyperparameters to improve replicability, including “seed”, which was unavailable during our study. Chat completion API also provides a placeholder for system messages to act as a guide to the overall generation of the model. We have kept the system message empty to avoid any external bias imposed on the model other than the prompt itself. This design decision directly reflects the objective of our study, where we aim to instruct human examiners and GPT-4 as closely as possible. All the remaining hyperparameters are kept unchanged as provided by the API. Our original prompts are in German. Original prompts and their translations into English can be found in Section S3 of the Supplementary Material.

Results

Inter-rater Reliability in Ranking and Point Assessment Task

We test Hypothesis 1 – if IRR drops when replacing a human examiner with GPT-4 – separately for our two tasks as different statistics are to be applied. For the ranking assessment, we use Kendall’s W. It quantifies the agreement between n raters, where $n \geq 2$, on a range from 0 to 1. It is the preferred metric for measuring IRR when data is ordinal. A score of 1 indicates complete agreement between the raters, while a score of 0 indicates no agreement. Figure 3 illustrates the agreement between two human examiners using a bubble chart. The 300 ranks assigned by Examiner 1 are on the x-axis and those assigned by Examiner 2 are on the y-axis. The size of the bubble indicates the frequency of the observations. The bubbles are visibly largest on the diagonal, diminishing in size as they diverge from this central axis. This shows considerable agreement between the examiners, but also the inevitable disagreement between well-trained humans. To illustrate subsequent data, the IRR for the two humans in Fig. 3 amounts to 0.83.

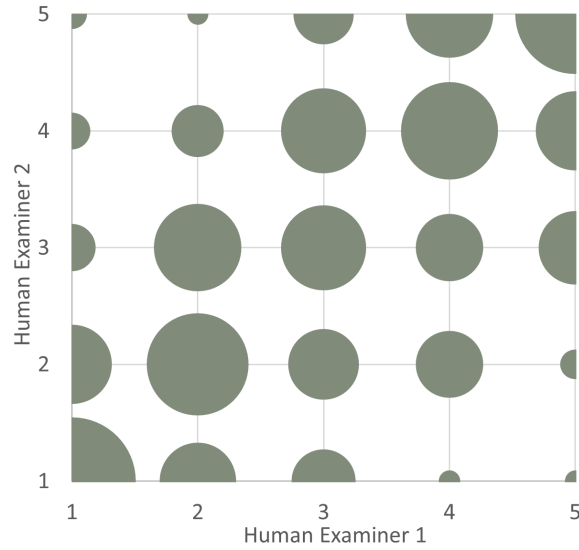


Figure 3. Bubble Chart of Ranks Assigned by Human Examiners 1 and 2

198 We calculate Kendall's W separately for the All Human and the three GPT-Human Teams and then compare the outcomes.
 199 If Kendall's W is significantly lower in the GPT-Human Teams, we find evidence in favor of Hypothesis 1. We illustrate our
 200 results in Fig. 4. As can be seen on the right, the data pooling all 300 observations does not show any significant differences.
 201 The IRR is 0.79 for the All Human Team (marked by a white bar) and either stays at that level for the GPT-Human Teams or
 202 even increases to 0.81 for the GPT-Human Team 2, marked by the darker bars (for details and summary statistics, see Section
 203 S4 in the Supplementary Material). The overlapping confidence intervals indicate that there are no significant differences. This
 204 can also be confirmed for the six questions separately, each referring to 50 observations. Overall, Kendall's W is higher in
 205 Question 2 and lower in Question 6, which could correspond to the quality of the students' answers or the examiners' ability
 206 to identify them. Both are likely to vary from question to question. However, this variation affects the All Human and the
 207 GPT-Human Teams equally. There is no question where the GPT-Human Teams perform significantly worse.

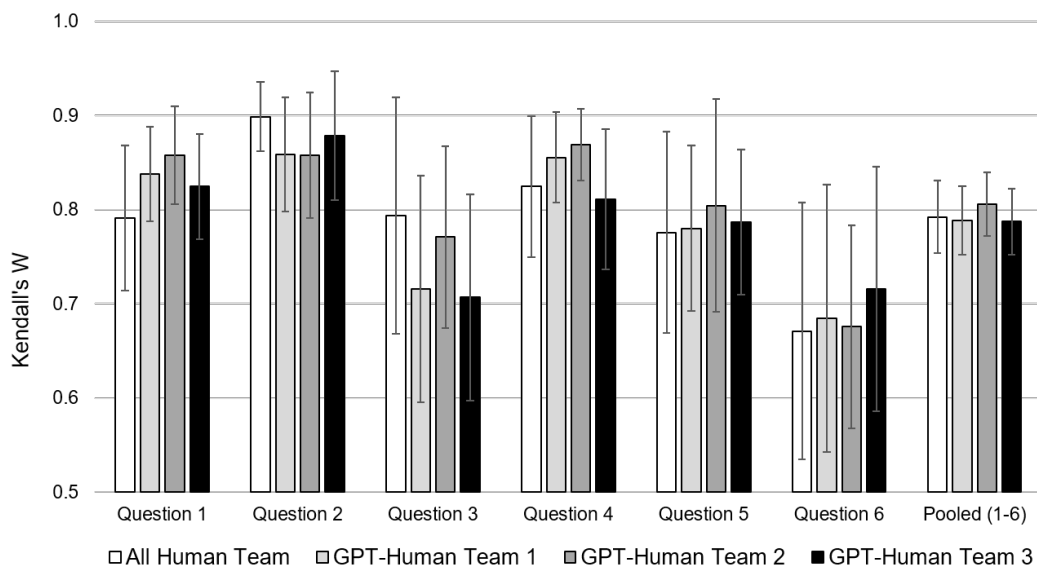


Figure 4. Results for the Ranking Task by Question with 95 Percent Confidence Intervals

The IRR for point assessments requires a slightly different approach. While ranks reveal ordinal information, point assessments are cardinal, meaning that a difference between 0 and 1 is equal to a difference between 1 and 2. The IRRs of our All Human Team and GPT-Human Teams should then be measured using Cronbach's Alpha. Figure 5 illustrates our results for the point assessment. As can be observed, we now find a slight drop in the IRR for all GPT-Human Teams pooling all questions, with coefficients around 0.933, compared to the All Human Team, with a value of 0.944. This drop however lacks statistical significance on any conventional level. Here, we report 95% confidence intervals. Looking at effect heterogeneity across Questions 1–6, it appears that this difference is mainly driven by Question 3 which presents somewhat of an outlier. Contrary to the other questions, we find a significant difference between the All Human Team and each of the three GPT-Human Teams for this question. This raises doubts about the capacity of GPT-4 to replace human examiners when it comes to point assessments.

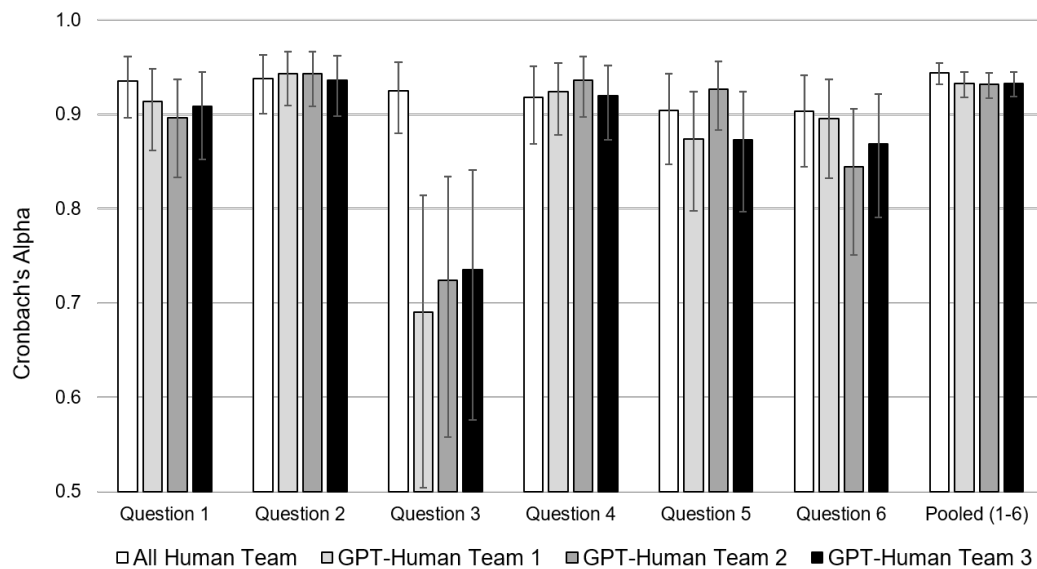


Figure 5. Results for the Point Assessment Task by Question with 95 Percent Confidence Intervals

Why does GPT-4 perform well on a ranking task and less well on a scoring task? A ranking task requires the comparison of five answers relative to each other. A higher rank for one means a lower rank for another. A scoring task allows more points to be assigned to the sum of all five answers, allowing one to perform better without downgrading the other. A scoring task thus gives examiners a degree of freedom in determining the level of tolerance. Some will be tolerant and show leniency in awarding points, while others will be stricter.

There are in fact major differences in tolerance between GPT-4 and the three human examiners. The human examiners award on average 1.30, 1.23, and 1.35 points, showing only small differences. GPT-4, on the other hand, stands out with an average of 1.81. The instructions, potentially supported by experience with related tasks, may have supplied human examiners with a similar level of tolerance. The data on which GPT-4 has been trained appears to generate a substantially higher level of tolerance. This difference is particularly large on Question 3, where GPT-4 awards twice as many points as human examiners. In the other five questions it awards only moderately more points – about 40% on average. The difference in tolerance thus seemingly accounts for the drop in IRR for Question 3. What is special about Question 3? It stands out as having by far the lowest scores of all questions – on average, 0.5 points out of 3. Comparable questions with the same maximum score have an average of 1.2 points. Our findings on tolerance are therefore similar to those of Nilsson and Tuvstedt¹⁶, who observe that GPT is particularly tolerant of low-scoring submissions.

This higher level of tolerance per se need not necessarily cast doubts about GPT-4's qualifications as an examiner. Given our assumption of human fallibility, we cannot exclude the possibility that our human examiners were excessively intolerant. We run a further test to check whether GPT-4 performs worse with its tolerance level. We do this purely based on the low-scoring submissions on which GPT-4 and human examiners disagree, mixing the many cases where humans are less tolerant with the fewer where their tolerance is higher. We first select all student responses to which neither of the human examiners assigned a point while GPT-4 assigned 1 point. We find 24 such answers. We mix these with all 17 answers where the human examiners were more tolerant and collectively assigned 2 or more points while GPT-4 gave none. We engage two advanced PhD candidates employed at the Chair of Economic Theory to serve as judges. The question is whether our two judges would assign a point or not while they are blind to the previous scores. Similar to the instructions provided to GPT-4 and the human examiners, they

receive the student answers in random order but sorted by question along with the respective question and sample solution to base their decision on. Would their scores correlate more with those of GPT-4 or with those of our human examiners? The benefit of this test is that the outcome does not depend on the judges' level of tolerance. This level might be identical to that of the human examiners while still siding with GPT-4 in selecting the deserving student answers.

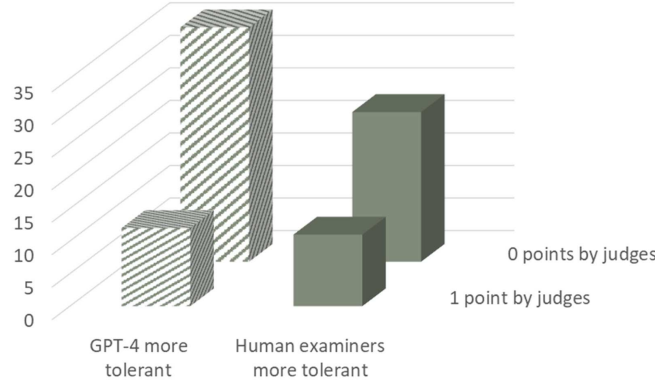


Figure 6. Resolution of Conflict by Judges

The results are shown in Fig. 6. With two judges per 41 student answers, we have 82 data points. As we can see, the bars are higher on the left when GPT-4 is more tolerant, showing that we had more observations there. They are also higher in the back, showing that the judges were more likely to assign 0 points. So the judges were less tolerant than GPT-4, just like the human examiners. The core question is whether the two judges' assessments would correlate with those of the other humans examiners, confirming that human judgment is superior to that of GPT-4. This hypothesis would suggest high bars on the second diagonal from the front-right observation to the back-left. While the back-left bar is the highest, the front-right bar is the smallest. Therefore, a visual inspection of the data does not suggest a correlation. A Fisher's exact test ($p=0.62$) does not allow us to reject the null hypothesis of no correlation. We can therefore conclude that the lower Cronbach's Alpha for GPT-4 is an artifact of its higher tolerance. It does not imply a lower quality in the scoring of student answers.

Self-serving Bias and Length Bias

Even if GPT-4 performs similarly to human examiners, there might be reservations against using it broadly for examination tasks. Its ranks and scores might show a problematic bias. GPT-4 could rate GPT-generated answers more favorably than humans would. Such a bias would imply a slippery slope with students using GPT to write answers and teachers simultaneously using GPT for grading, indicating an excessively favorable performance to both. A similar problem arises with lengthy answers, which might be indicative of "eloquent bullshit" performing too well, inducing students to write lengthy answers with little content. We examine the potential for such biases using an ordered probit regression:

$$GPT_i = \sum_{j=1}^3 \beta_j Human_{ji} + \gamma_1 GPTA_i + \gamma_2 Length_i + \sum_{k=1}^6 \delta_k Question_{ki} + \epsilon_i \quad (1)$$

where GPT and $Human$ refer to ranks and points assigned by GPT and our three human examiners respectively. The subindex i refers to our 300 observations. GPT_i is our dependent variable and we test how well it is explained by $Human_{ji}$ and potential biases. The first such bias is the dummy $GPTA_i$, which takes the value of 1 if a graded answer was written by GPT and 0 otherwise. $Length_i$ is the second potential bias, indicating the length of an answer in 10 characters. $Question_{ki}$ is also a dummy that takes the value 1 if the answer i refers to the question k and obtains the value 0 otherwise. We include this to account for unobserved heterogeneity between questions.

The positive and large coefficients on Human Examiners 1,2 and 3 corroborate our result that GPT ranks and awards points in line with our human examiners. Our coefficient of interest is on the $GPTAnswer$. If GPT has a bias towards preferring its own answer, we expect a negative coefficient for ranks and a positive coefficient for points. Indeed, we find point estimates for both coefficients that follow this pattern. However, both for ranking (Column 1) and points (Column 2), coefficients are very small and not significantly different from zero. We therefore find no evidence substantiating Hypothesis 2. We find that GPT does *not* judge GPT-generated answers more favorably than human examiners.

	(1) GPT-4 Rank	(2) GPT-4 Points
Human Examiner 1	0.36*** (0.07)	0.26** (0.1)
Human Examiner 2	0.38*** (0.06)	0.53*** (0.1)
Human Examiner 3	0.28*** (0.07)	0.78*** (0.1)
GPT Answer	-0.11 (0.2)	0.17 (0.2)
Length	-0.0055 (0.008)	0.032*** (0.008)
Question controls	Yes	Yes
Observations	300	300
Pseudo R^2	0.27	0.46

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2. Regression Results on Self-serving and Length Bias

When it comes to Hypothesis 3 and the question of whether GPT has a bias towards longer text answers, the coefficient on *Length* in Table 2 suggests that there is no bias towards preferring longer text answers when ranking answers (Column 1). While the coefficient is negative, it is very small in size and not significantly different from zero. When it comes to points, the picture looks somewhat different (Column 2). Here, we indeed find that GPT gives longer answers more points than our three human examiners. This suggests that GPT has a bias towards length when awarding points but not when ranking student answers. Why such a conflation of length and accuracy occurs in some tasks and not in others will require future research.

Robustness Checks and Extensions

Our focus on open-text answers in macroeconomics is quite specific. Whether our results are robust across further domains, disciplines, lengths of answers, or variations in style and content will have to be seen and is outside the scope of our study. Using our current dataset we can address another challenging question: To what extent do our findings depend on the current version of GPT-4 and variations in the design of the prompt? We also ask whether we can boost the performance of GPT-4 by using an advanced prompt strategy. Examiners who employ GPT-4 as an assistant are likely to use different prompts and approach the GPT-4 at different points in time. Would such variations matter to our results?

We employ the following additional checks to answer these questions: We re-run our original prompt version P_{v1} about 9 months later. Additionally, we investigate the effect of four design variations to determine how robust our prompt is towards certain manipulations. Our first variation refers to the structure of the prompt. We modify its structure without altering the content, employing four different orders ($P_{v2} - P_{v5}$). The second refers to linguistic variability. $P_{v6.1}$ ($P_{v6.2}$) randomly introduces a considerable (limited) number of heavy (minor) spelling errors to the text. P_{v7} alters the tone of the prompt introducing substantial ambiguity. The third variation refers to roles. We alter the role of GPT-4 from student assistant to P_{v8} professor, P_{v9} expert student assistant, P_{v10} non-expert student assistant, and P_{v11} no role. The fourth variation, P_{v12} , involves introducing an advanced prompt strategy with proven superiority in solving reasoning tasks compared to traditional prompting strategy²³. We adopt a variation of chain-of-thought (CoT), i.e., zero-shot CoT²⁴, where we guide the model to generate reasoning behind the ranks and points given. The prompts utilized are provided in Section S5 of the Supplementary Material.

Figure 7 shows how our manipulations affected average Kendall's W and Cronbach's Alpha across three GPT-Human teams for the different variations (see Section S6 in Supplementary Material for extended results). Kendall's W and Cronbach's Alpha drop slightly but not substantially in the re-runs of the original prompt. Similarly, we observe that variations in the prompt design appear to play a subordinate role: Changes to the structure of the prompt (as employed in $P_{v2} - P_{v5}$) decrease IRRs in all treatments but only slightly. Likewise, spelling errors ($P_{v6.1}$ and $P_{v6.2}$) as well as suggestive and ambiguous instructions (P_{v7}) in the prompt have a dampening yet negligible effect. Variations in the role assigned and not assigned to GPT-4 as indicated by $P_{v8} - P_{v11}$ do not considerably alter our results. Finally, we find that introducing an advanced prompting strategy, P_{v12} , does not substantially bring any improvement over the original prompt P_{v1} . Using GPT-4 as an examiner generates findings that are persistently close to those of human examiners and leave virtually no room for hallucination or performance drop.

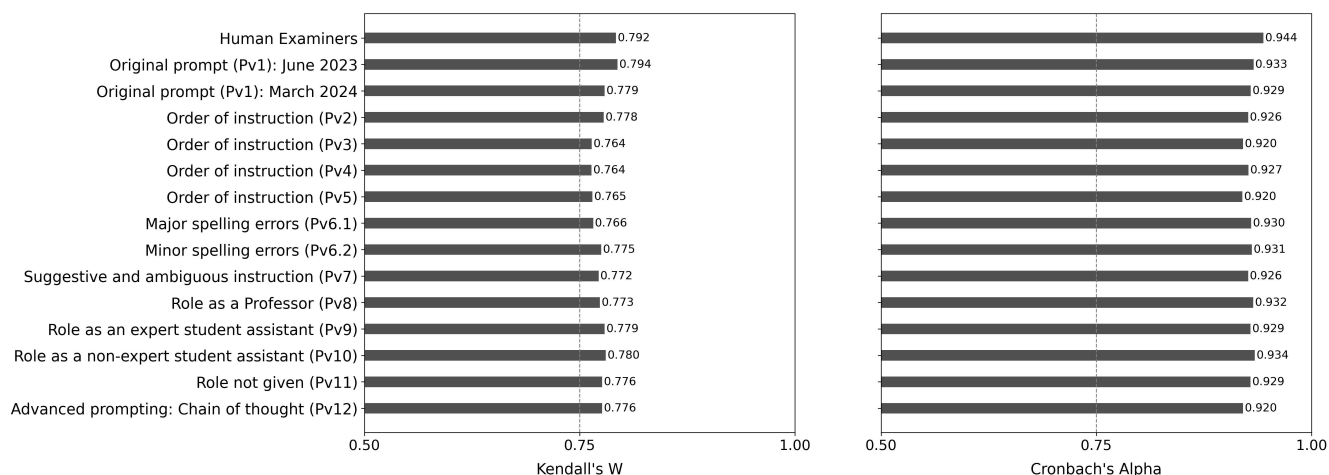


Figure 7. Robustness of GPT-4 Across Varied Run Times and Prompts in Ranking and Point Assessment

Discussion and Conclusions

Undoubtedly, GPT-4 poses major challenges to the education system, but it also offers big opportunities. The latter could lie in its potential to reduce teachers' workload – for example, by letting GPT-4 grade and assess student answers. This study investigates whether GPT-4 can actually replace human examiners in ranking and point-assessing student answers to open questions. Particularly for the ranking tasks, our results give overworked teachers cause for optimism: GPT-4 can recognize differences in the quality of student answers as reliably as trained human examiners. We do not detect a bias towards GPT-generated answers or the length of answers in the ranking assessment of GPT-4. Hence we can rule out that GPT-4 is inclined to fall for eloquent bullshit or linguistic proximity to its own model. Concerns that unethical and dishonest students using AI for their answers may gain an advantage seem unfounded.

Also, in the point assessment task, GPT-4's performance is promising: There is a slight yet insignificant decrease in IRR from the All Human Team to the GPT-Human Teams. This may be due to a greater tolerance in the point assessments of GPT-4, which is particularly visible for low-scoring student answers. In this regard, our findings are seemingly in line with Nilsson and Tuvstedt¹⁶, who found GPT to be inferior with failing submissions. Our additional test with human teachers to check whether GPT-4's tolerance of low-scoring answers systematically differs from the human examiners does however not indicate such an issue. Thus, we find that GPT-4 can indeed reliably replace human examiners in both ranking and point assessing student answers. Again, there is no bias towards GPT-generated answers in GPT-4's point assessments. Yet, the length of answers positively affects the number of points awarded by GPT-4, showing that concerns about a conflation of length with accuracy cannot be completely ruled out.

Our study has a few further caveats to consider when drawing implications: The open questions asked are not ambiguous – there is only one correct solution. Hence, the external validity of our findings is limited to similar types of questions. Potentially, GPT-4 is not as competent in differentiating between multiple correct answers. Future research should thus investigate if results hold for more complex questions allowing for more than one correct argumentation. Another regard in which the external validity might be limited is the language. We conducted our study in German. GPT-4's accuracy in German is at a quite similar level as in English⁵, yet its performance in ranking and point assessing student answers might be different in languages it is less proficient in. Also in this regard, further testing should be done. Relatedly, our results are limited to the use of GPT-4. We cannot draw any inferences on the use of other large language models in assessing student answers.

Eventually, there are some ethical and legal considerations on the part of the teachers to take into account: We do not recommend that teachers pawn off grading exams and essays to GPT-4 without any control or supervision. Ultimately, grading and assessing student's work lies in the responsibility of the teacher. There is currently no legal basis to transfer this responsibility to the AI. The AI cannot be held responsible in case of misgrading. This falls under the teacher's purview. We thus only cautiously recommend using the AI's grading and assessment capabilities as a full replacement to a teacher's assessment in setting where grades carry little importance. When it comes to a complementary role to teacher's assessment, we can lift these restrictions and advocate for broader use of GPT-4: Pre-grading exams GPT-4 seems to do just as well as trained teaching assistants.

Data availability

The datasets generated during and/or analysed during the current study are available in Zenodo repository, <https://doi.org/10.5281/zenodo.11085379>.

Code availability

A link to a replication package that contains the data and the analysis code is also available in Zenodo repository, <https://doi.org/10.5281/zenodo.11085379>

References

1. OpenAI. Introducing ChatGPT. <https://openai.com/blog/chatgpt> (2022). Accessed: 2023-12-20.
2. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. *et al.* Improving language understanding by generative pre-training (2018).
3. Brown, T. *et al.* Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901 (Curran Associates, Inc., 2020).
4. Liu, P. *et al.* Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **55**, DOI: [10.1145/3560815](https://doi.org/10.1145/3560815) (2023).
5. OpenAI *et al.* Gpt-4 technical report (2023). [2303.08774](https://arxiv.org/abs/2303.08774).
6. Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z. & Trautsch, A. A large-scale comparison of human-written versus ChatGPT-generated essays. *Sci. Reports* **13**, 18617, DOI: [10.1038/s41598-023-45644-9](https://doi.org/10.1038/s41598-023-45644-9) (2023).
7. Noy, S. & Zhang, W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192, DOI: [10.1126/science.adh2586](https://doi.org/10.1126/science.adh2586) (2023).
8. Yeadon, W., Inyang, O.-O., Mizouri, A., Peach, A. & Testrow, C. P. The death of the short-form physics essay in the coming AI revolution. *Phys. Educ.* **58**, 035027, DOI: [10.1088/1361-6552/acc5cf](https://doi.org/10.1088/1361-6552/acc5cf) (2023).
9. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on medical challenge problems. arXiv: 2303.13375 (2023).
10. Yuan, Z., Yuan, H., Tan, C., Wang, W. & Huang, S. How well do large language models perform in arithmetic tasks? (2023). [2304.02015](https://arxiv.org/abs/2304.02015).
11. Gilson, A. *et al.* How does ChatGPT perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR Med. Educ.* **9**, e45312, DOI: [10.2196/45312](https://doi.org/10.2196/45312) (2023).
12. Stribling, D., Xia, Y., Amer, M. *et al.* The model student: GPT-4 performance on graduate biomedical science exams. *Sci. Reports* **14**, 5670, DOI: [10.1038/s41598-024-55568-7](https://doi.org/10.1038/s41598-024-55568-7) (2024).
13. Jeon, J. & Lee, S. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Educ. Inf. Technol.* **28**, 15873–15892, DOI: [10.1007/s10639-023-11834-1](https://doi.org/10.1007/s10639-023-11834-1) (2023).
14. Jauhiainen, J. S. & Guerra, A. G. Generative AI and ChatGPT in school children's education: Evidence from a school lesson. *Sustainability* **15**, DOI: [10.3390/su151814025](https://doi.org/10.3390/su151814025) (2023).
15. Phung, T. *et al.* Generative AI for programming education: Benchmarking ChatGPT, GPT-4, and human tutors (2023). [2306.17156](https://arxiv.org/abs/2306.17156).
16. Nilsson, F. & Tuvstedt, J. GPT-4 as an automatic grader: The accuracy of grades set by GPT-4 on introductory programming assignments. Dissertation (2023).
17. Geertsema, P., Bifet, A. & Green, R. ChatGPT and Large Language Models: What are the Implications for Policy Makers?, DOI: [10.2139/ssrn.4424048](https://doi.org/10.2139/ssrn.4424048) (2023).
18. Rudolph, J., Tan, S. & Tan, S. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *J. Appl. Learn. Teach.* **6** (2023).
19. Giamattei, M. & Lamsdorff, J. G. 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](https://doi.org/10.1016/j.jbef.2019.04.008) (2019).
20. OpenAI. Chat api reference. <https://platform.openai.com/docs/api-reference/chat> (2023).

21. Sainz, O. *et al.* NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Bouamor, H., Pino, J. & Bali, K. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10776–10787, DOI: [10.18653/v1/2023.findings-emnlp.722](https://doi.org/10.18653/v1/2023.findings-emnlp.722) (Association for Computational Linguistics, Singapore, 2023).
22. Espejel, J. L., Ettifouri, E. H., Alassan, M. S. Y., Chouham, E. M. & Dahhane, W. GPT-3.5, GPT-4, or BARD? evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Nat. Lang. Process. J.* **5**, 100032 (2023).
23. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 35, 24824–24837 (Curran Associates, Inc., 2022).
24. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. In Koyejo, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 35, 22199–22213 (Curran Associates, Inc., 2022).

Acknowledgements

We thank Sarah ben Messaoud, Joachim Schnurbus, the members of the project DeepWrite and the participants of the Annual meeting of the German Association for Experimental Economic Research e.V. (GfEW) 2023 in Erfurt for helpful advice and comments. We also thank our five human examiners and judges.

Author contributions statement

M.G. and J.G.L. conceived and designed the study, including the statistical strategy. S.G. and D.V. were responsible for setting up the data collection process, providing instructions and the original prompt for the point and rank assessment analysis, and investigating self-serving and length biases. A.A.Z. carried out the initial analysis for the ranking tasks and point assessments as well as robustness checks and extensions including prompt variations. All authors actively engaged in discussion of the results and implications and collaboratively contributed to the manuscript's writing and review process throughout all stages.

Competing interests

The authors declare no competing interests.

Funding

The report has been funded by the German Federal Ministry of Education and Research (BMBF) under the project DeepWrite (Grant. No. 16DHBKI059).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.pdf](#)