scientific reports

Check for updates

OPEN

Examining the responsible use of zero-shot AI approaches to scoring essays

Matthew Johnson✉ & Mo Zhang

The promise of AI to alleviate the burdens of grading and potentially enhance writing instruction is an exciting prospect. However, we believe it is crucial to emphasize that the accuracy of AI is only one component of its responsible use in education. Various governmental agencies, such as NIST in the US, and non-governmental agencies like the UN, UNESCO, and OECD have published guidance on the responsible use of AI, which we have synthesized to come up with our principles for the responsible use of AI in assessments at ETS. Our principles include fairness and bias mitigation; privacy & security; transparency, explainability, and accountability; educational impact & integrity; and continuous improvement. The accuracy of AI-scoring is one component of our principles related to educational impact & integrity. In this work, we share our thoughts on fairness & bias mitigation, and transparency & explainability. We demonstrate an empirical evaluation of zero-shot scoring using GTP-4o, with an emphasis on fairness evaluations and explainability of these automated scoring models.

Keywords AI, Scoring, Fairness, Explainability, Educational Measurement

Constructed-response (CR) items, also known as open-ended questions, is a common question type in educational assessment. Some examples include writing an essay, summarizing information, or providing explanations or evidence reasoning. CR items allow learners to demonstrate a spectrum of writing competencies ranging from low-level skills, such as grammar, word usage, and sentence structure, to high-level skills including argument articulation and critical thinking. The key difference of CR items from multiple-choice (MC) or selected-response (SR) items is that the scoring of CR items is generally not amenable to exact-matching approaches because the specific form(s) and/or content of the correct answer(s) are not known in advance[1]. This study concerns of the most common CR item type - essays.

Essay task is considered an integral component of writing assessments under a wide range of assessment contexts, from low-stakes classroom-based writing assessment to large-scale, high-stakes assessments. An essay task can measure important skills that are not adequately assessed by MC or SR items and thereby encouraging teachers to include essay writing in daily classroom practice[2]. Traditionally, if essays were included in an assessment, they were scored by human raters. However, evaluating thousands of essays can be labor-intensive, time-consuming, and expensive[3,4]. A possible solution to this efficiency problem is to score essays automatically using computers. The rapid advance of artificial intelligence (AI) technology, in particular the past 20 years, has made machine scoring of text a realistic option.

Earlier days of AI scoring

AI scoring (also called automated scoring) of text responses in educational assessments can be dated back to the 1960s[5] for the benefits of quick score turnaround, instant feedback, cost-saving related to human scoring, and arguably greater scoring accuracy and score reliability[6–9]. Earlier work formulated the AI scoring task as a prediction of human ratings based on text features or a comparison of text similarity between student-submitted answers and reference answers (e.g.,[10–12]). Responses are represented as vectors consisting of hand-crafted features such as lexical complexity, syntactic relations, and character or word n-grams; those vectors are then fed into various prediction or classification models to generate a score that best predicts what a human rater would assign to a given response[13–17].

Bejar et al.[18] argued that, for automated scoring, explicit evidence must be identified, quantified, and then aggregated by the systems. In the early days of automated scoring, several systems were well-known, such as Project Essay Grade (PEG), Intelligent Essay Assessor, and e-rater, which we will describe next. Some of the systems are still being used operationally today but might have been updated to the latest AI techniques. Those systems, in their earlier versions as reported in the literature, tended to extract well-defined NLP features to

Educational Testing Service, Research Division, 08541 Princeton, New Jersey, USA. ✉email: msjohnson@ets.org

clearly quantify the construct of measurement (e.g., academic writing skills) and/or apply mathematical or statistical models that are easy to interpret.

Take Project Essay Grade (PEG) as the first example. It is one of the earliest automated scoring systems, initially developed by Page[5] during the 1960s in the U.S. In developing PEG, Page[19] coined two important terminologies: trins and proxes. Trins are intrinsic variables that cannot be directly computed from an essay (e.g., fluency, diction, sentence structure, grammar, punctuation, construction); proxes are observable variables that can be computed to approximate the trins. For example, fluency can be approximated by the number of words, diction can be approximated by the variance of word length, and sentence structure can be approximated with counts of prepositions and relative pronouns[19,20]. Although most features used by PEG were originally surface features[21], since its introduction in 1966, PEG's developers made progress on narrowing the gap between the proxes and the trins[20]. A multiple linear regression model was trained using pre-scored essays that can be used to predict human ratings[20], and a final scoring model often contained 30 to 40 proxes[21].

Intelligent Essay Assessor (IEA), as reported for its earlier version, applied a matrix-based latent semantic analysis (LSA) method to detect the lexical semantic similarity between texts[22,23]. Assuming that the meaning of the text is derived solely from the words, LSA analyzed the similarity of words, terms, passages, and essays by placing them in a latent semantic space[24]. When the IEA system received a to-be-scored essay, the system evaluated the essay in terms of its location in a reduced multi-dimensional semantic space. The essay was first converted to a vector in the space. Next, the IEA system searched for ten pre-scored essays that had the smallest angle to the to-be-graded essay at the origin[24,25]. The angle reflected the similarity of information in the to-be-scored essay to those reference essays, which was an indication of the quality regarding content-relevance. Then, the average of the human scores on the ten pre-scored essays was weighted by the cosine value of the magnitude of the average angle, which constitutes an essay's content score[22]. In addition to the content score, the IEA system also extracted two other essay feature scores: "corpus-statistical writing-style" (e.g., grammar) and "mechanics" (e.g., punctuation)[26]. The final score of a to-be-graded essay was computed using a linear regression in which the human ratings were regressed on the content, style, and mechanics scores, with the content feature always being constrained to have the highest weight in the model[24].

Finally, for *e-rater*, it was initially developed at Educational Testing Service in the late 1990s. *e-rater* relied heavily on natural language processing (NLP) techniques. NLP is an AI area concerned with processing human language using computers[27], by essentially translating ambiguous natural-language text into unambiguous, quantifiable internal measures. NLP allows for analysis of large bodies of text from which knowledge structures can be derived[28]. *e-rater* differed from other automated essay scoring systems in its use of a small set of linguistic features extracted using NLP techniques that allowed for constructing more understandable scoring models[29]. In their 2019 study, Yao et al.[9] reported nine e-rater features, which included measurement of grammar, mechanics, word usage, text organization, development, etc. Employing a corpus-based approach[6], e-rater scoring models were empirically trained on essays from a particular prompt or bundle of prompts within a genre. The system first processed a sufficient number of training essays and computed the feature values. Next, human ratings were regressed on the feature values, resulting in a multiple regression model that could be used to predict a score that a human grader would assign to a given essay[30]. The weights, or contributions, of each feature are traceable from the model and can be used for validating and interpreting the resulting automated scores.

As can be seen from the examples above, earlier days of AI scoring tended to apply white- or grey-box statistical or machine learning models to predict scores using construct-relevant, well-defined meaningful information extracted from the text responses. The landscape is changing with the rise of generative AI and transformer-like architectures.

Present AI scoring landscape

More recent work on AI scoring, which is still rapidly emerging as of this writing, started to adopt deep neural networks and transformer-based models for scoring (e.g.,[31–35]). Several well-attended public competitions on AI scoring (e.g., Automated Essay Scoring and Short-Answer Scoring Kaggle competitions organized by the Hewlett Foundation in 2012, NAEP Automated Scoring Challenge for reading assessments organized by the U.S. Department of Education in 2021, and Automated Essay Scoring 2.0 Kaggle competition organized by Learning Agency Lab in 2024) reflected this trend from using more explainable models such as linear regressions with a smaller number of hand-crafted meaningful features to using deep learning methods with LLMs that estimate billions of parameters simultaneously.

For an assessment, the choice of the modeling approach is highly contextually dependent on the purpose and characteristics of the assessment and the specific task to be scored. Researchers generally agree that while these very complex black-box models tend to outperform white- or grey-box models for prediction tasks in terms of accuracy, they are much less transparent and interpretable[36,37]. Figure 1, taken from[36], demonstrates the inverse relational pattern between model complexity and interpretability.

We could not find a study that compares white-box to black-box models using the same set, but we can perhaps gain some idea from different studies that all used K-12 students responses (similar to the ones used in this study). For essay scoring, Zhang and Deane[38] reported an R-squared of 0.694 (converted to a correlation coefficient of 0.833) between human and AI scores by using the (white-box) multiple linear regression model on 10 e-rater features (Table 8 in the referred publication); Sinharay et al.[39] reported a correlation coefficients of 0.838 and 0.772 between human and AI scores by using the (grey-box) gradient boosting with the same 10 features (Table 6 in the referred publication); two of the winners from the 2012 ASAP Automated Essay Scoring Kaggle competition disclosed their codes: one used K-Nearest Neighbors and the other used Neural Networks as scoring models. We are not able to pinpoint the exact places on the public leader board for those two winners, but the quadratically-weighted kappa (QWK) ranged from 0.783 to 0.801 among the winners. The gold winner[40] in the 2024 Kaggle essay scoring competition obtained a QWK of 0.772 between human and AI scores using

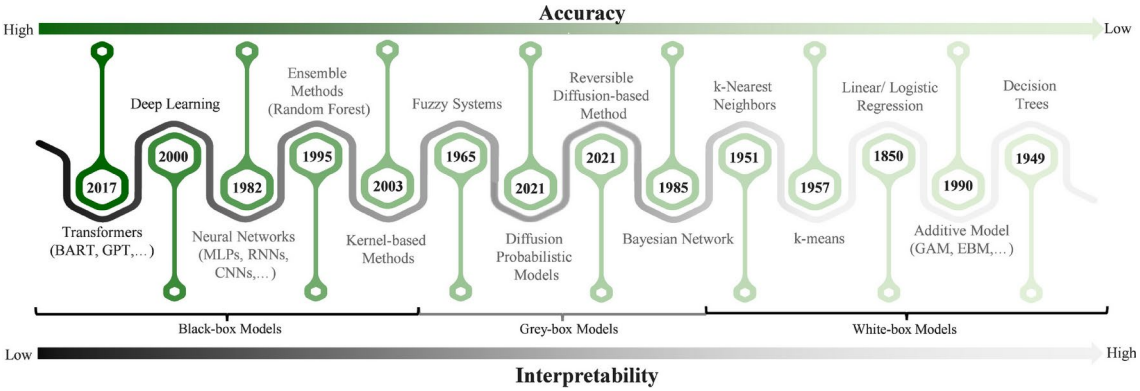

Figure 1. Illustration of the balance between accuracy and interpretability. The need for high model accuracy and interpretability is emphasized. Models with high accuracy need additional explainability, while models with low accuracy are simpler to comprehend but useless in many cases. The grey-box models represent the transition between black-box and white-box models. The number represents the year in which it first appeared in the field of AI research. (Graph and caption taken from[36] without modification.).

fine-tuned DeBERTaV3 model. It is noted that the data set in the 2024 Kaggle competition partially overlaps with the data set analyzed in this study. For scoring of short-response text, Heilman and Madnani[41] reported relatively low QWK values ranging from 0.438 to 0.708 across various scoring tasks from K-12 assessments using support vector regression models, which can be considered grey-box models, on extracted features such as word unigrams and word bigrams (i.e., sequence of adjacent words) (Table 4 in the referred publication). Later, using the same scoring model – support vector regression – with a similar feature set, Yao et al.[42] reported a 0.717 as the highest median QWK value the authors could obtain across different scoring tasks (Table 5 in the referred publication). The 1st place winner of the 2021 NAEP Automated Scoring Challenge reported a QWK of 0.888, and their model "extract features combined with Latent Semantic Analysis (LSA), the generation of N-gram vectors, Latent Dirichlet Allocation (LDA), passage similarity metrics, and deep neural network embeddings to train the best ensemble model from a series of classifier and regression models"[43].

So, although not based on an ideal comparative analysis, gaining prediction accuracy with complex black-box models appeared to be the case in the AI scoring context for short-response items where the evaluation of a text response is focused on content, but not so much for long essays where the evaluation of a text response may be more about writing conventions such as grammatical accuracy and correct use of vocabularies. Even though the content certainly matters in long essays, such as the logical structure of an argument, unlike extracting features that indicate writing mechanics, it has been very challenging to design and extract manual features that can fully capture the content quality and the nuances in the content. Deep learning models, such as those based on transformer architecture, however, can generate extremely nuanced contextual representations of a text and can deal with the complexities and ambiguities in human language reasonably well[44]. Therefore, those black-box models may be more suitable and leading to higher prediction accuracy, when scoring for content-based short-response items. But, while one can potentially gain some prediction accuracy with the use of complex models, a lack of interpretability can be a serious issue when using these black box models in educational assessments especially for when bias in scoring is detected, it can be extremely difficult or even impossible to pinpoint what is causing the bias and to mitigate bias.

Presently, as mentioned earlier, with the rising popularity and promises of open-source large language models (LLMs), researchers around the world are conducting timely studies to investigate the potential of leveraging pre-trained LLMs for the purpose of AI scoring. Often researchers finetune pre-trained LLMs such as BERT, LLaMa or GPT for the downstream tasks. In terms of AI scoring, such downstream tasks typically involve predicting human ratings on a prompt, for which the process still requires labelled data (e.g., human ratings). LLMs have been shown to entail and criticized for "pervasive issue of bias"[45] for population groups differing in age, gender, color, ethnicity, geographic location, language and cultural background, political affiliation, disability status, etc. (e.g.,[46–49]). These biases are originated from the composition of the enormous training corpus of the LLMs[50–52]. For example, the original $BERT^{BASE}$ model was pre-trained on Wikipedia English data and BookCorpus data[53], and the pre-training of DeBERTaV3 model used data from Wikipedia, Bookcorpus, CC-News, Stories and OpenWebText combined[54]. As a result, a LLM would reflect the writing styles and linguistic characteristics demonstrated in its pre-training data. Previous research has cautioned that large corpora used to train language models "overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations"[55]. So in the context of scoring, these inherent biases in LLMs will be carried over and/or further exacerbated in the downstream tasks, for instance, if there exists algorithmic biases (of the same direction) in the downstream AI scoring applications.

Statement of research problem

The principle of fairness is foundational to responsible AI and a key standard of educational testing. As highlighted by researchers like[56], generative AI solutions like ChatGPT show some promise in terms of their

accuracy; however, there is still a risk that biases exist in their outputs. For example, differences in language usage or cultural references made by students in their writing could lead to biased scoring, disadvantaging certain groups of students while advantaging other groups. Various governmental agencies, such as NIST in the US, and non-governmental agencies like the UN, UNESCO, and OECD have published guidance on the responsible use of AI, which we have synthesized to come up with our principles for the responsible use of AI in assessments[57]. Our principles include fairness and bias mitigation; privacy & security; transparency, explainability, and accountability; educational impact & integrity; and continuous improvement. The accuracy of AI-scoring is one component of our principles related to educational impact & integrity.

Published research so far on using deep learning and transformer-based models has very little empirical work and discussion on fair use of AI for scoring; if at all, this type of discussion only appears briefly in the discussion section. Prediction accuracy tends to be prioritized in the use of LLM for AI scoring. All the aforementioned Kaggle competitions, including the most recent one that concluded in June 2024, only considered prediction accuracy as measured by quadratically-weighted kappa for ranking the competitors. We argue that it is critically important that AI systems in education evaluate, mitigate, and work to eliminate biases to foster an inclusive learning environment. This involves meticulous attention to the data used in training the generative AI models and ensuring the evaluations of the AI scoring represents diverse student backgrounds and abilities.

There is not much work that has been done in terms of using pre-trained LLMs under zero-shot condition for automated scoring purposes, with few exceptions (e.g.,[58–61]. A zero-shot scoring can better expose biases inherent in the pre-trained LLMs which might otherwise be hidden or overcome through finetuning. We can therefore better gauge their direct impact on scoring. Based on our knowledge, none of the published works analyzed the AI scoring potential under zero-shot condition from a fairness perspective. With fairness being a central concern in educational assessment, our work intends to fill this gap in the literature.

What if bias is detected? And how to mitigate bias? Explainability or interpretability is critical to ensure that an AI scoring model assigns accurate scores for valid reasons, which will, in turn, build people's trust and confidence in fair use of AI. Explainability is a serious challenge in developing and evaluating AI scoring capabilities. The problem gets even more complex when the model output is incorrect, biased, or not obviously wrong due to the size and complexity of the models. There is a growing body of research around explainable AI (XAI) that proposes methods to interpret such complex models as deep neural networks (see[37,62] for summaries). Many of those XAI techniques have achieved meaningful results in explaining the model outputs and have open-sourced their codes. For example, Modarressi et al.[63] proposed a method that works specifically well with transformer-based deep learning models. But, to apply any existing XAI methods requires knowing the actual scoring model, which is un-accessible in the case of GPT. To tackle this problem, we directly investigate the predictability of the sensitive attributes that we try to ensure fairness (i.e., race/ethnicity) using the same embedding input that is used to predict scores. We describe, in more details, the rationale of this approach in the next section.

To summarize, we attempt to address two specific research questions in evaluating zero-shot scoring using GTP-4o:

1. What is the scoring accuracy of zero-shot scoring using GPT-4o? And is it fair for different ethnicity groups?
2. If a lack of fairness is detected, can we explain it?For the remainder of this paper, we first describe our approach to evaluating AI scoring in terms of accuracy, fairness and explainability, followed by the actual zero-shot GPT-4o scoring experiment and results. We conclude with a discussion of the findings and the limitations of this study.

Our approach to evaluating AI scoring

How to evaluate AI scoring is generally well established in the literature, although new challenges do emerge with the use of LLMs. Overall, we followed the best practice guidance suggested in[64] to evaluate the scoring accuracy and fairness. Some specifics are given below.

Evaluating accuracy

Haberman[65] defined prediction accuracy as measuring the ability to predict one variable by use of one or more other variables. Beyond the scope of discussion in current study, the author also made a distinction between prediction accuracy and agreement where Cohen's kappa[66], for example, is a form of agreement metric. In evaluating accuracy, we calculated percentage agreement between AI and human scores, Cohen's unweighted and quadratically weighted kappa, disattenuated correlation, as well as proportional reduction in mean squared error (PRMSE) for predicting the human true score[65,67].

Evaluating fairness

The most straightforward way to evaluate fairness is to compare the raw mean differences between human and AI scores for each demographic group. In this study, we also computed an adjusted mean score difference that conditioned on human true score. Technical details of this metric can be found in[68,69].

Explaining results

As mentioned earlier, without knowing the actual scoring model, we chose to use the embeddings to predict demographic group membership. This can be thought of as similar to the statistical concept of canonical correlation and a technique called "BIOT" (Best Interpretable Orthogonal Transformation[70]), where multidimensional text embeddings are on one side of the equation and external variables such as score or sensitive attributes to fairness (e.g., ethnicity) on the other side. The predictive relations are revealing from a fairness perspective in that, for AI scoring applications, it would not be desirable for the same input being predictive of not only performance, but

	GPT-4o Generated score					
Human	**1**	**2**	**3**	**4**	**5**	**6**
1	**43**	38	1	0	0	0
2	155	**1546**	283	2	0	0
3	52	2075	**1562**	70	0	0
4	10	1001	2470	**608**	15	0
5	1	158	1221	960	**118**	0
6	0	7	179	376	167	**3**

Table 1. Classification table comparing the scores assigned by humans and GPT-4o. Bold values indicate where there is exact agreement between humans and GPT-4o

Rater	Mean	Std.Dev
Human	3.69	1.12
GPT-4o	2.79	0.81

Table 2. Summary statistics for the human and GPT-4o assigned scores.

also one's group membership which is considered as construct-irrelevant. If so, it would suggest that the scoring model that takes the same input may capitalize on construct-irrelevant information that contains bias.

Results of an evaluation of zero-shot essay scoring

To evaluate the accuracy, fairness, and explainability of zero-shot generative AI scoring, we had GPT-4o score the entire database of 13,121 independent essays in the Persuade 2.0 database[71,72]. The essays were responses to written in response to one of eight prompts, and were answered by students in grades eight through twelve, with the vast majority of essays written by eighth (61.3%) and eleventh (23.5%) grade students. The gender of the writers was nearly evenly split between female (50.03%) and male (49.97%) writers. The race/ethnicity of writers is also included in the data set, with 42.8% identified as White, 23.0% identified as Hispanic/Latinx, 20.8% identified as Black/African American, and 9.0% identified as Asian/Pacific Islander. The essays ranged in length from 146 words to 4855 words, with a median of 407 words, and first and third quartiles of 286 and 564 words, respectively. Other investigators have used the same data set for research on AI scoring, such as[56] that sampled from the same corpus to create sample #3 in their paper.

To generate scores, we used OpenAI's Batch API service (OpenAI, 2024) to request GPT-4o to produce holistic ratings of each essay. We prompted the AI to score the essays by first passing the text of the "Holistic Rating Form" for independent essays available on the Persuade 2.0 GitHub page[71] as the system message and then asking GPT-4o to score the essay by passing the following user message,

> *score the essay that follows using the holistic scoring rubric: give only the numeric score. do not write anything other than the number. The essay is:*

Specific student essays were appended to this prompt after a line break. A prototype entry in the JSONL file passed to the Batch API is shared in Appendix A.

Accuracy results

We found that GPT-4o scored the essays poorly. On average, the AI scores were 0.9 points lower than the human ratings, matched only about 30 percent of the time, and were within one point of the human rating only 77 percent of the time. The classification table appears in Table 1 below.

It is clear from the classification table that our zero-shot prompt generated AI scores tended to be too low across much of the distribution. For example, essays where humans assigned scores of 3, were most likely to be scored as 2's by the GPT-4o. Similarly essays receiving human ratings of 4, 5, and 6 had modal generated scores of 3, 3, and 4, respectively. Indeed, the mean scores generated by GPT-4o, reported in Table 2, were on average 0.9 points lower than the human ratings on the same essays; the 95% confidence interval for the difference is (0.875, 0.922). The standard deviations of scores were also lower. The difference is drastic if we look at the marginal distributions. GPT-4o is much harsher than human raters: GPT-4o only assigned scores 6 and 5 to 3 and 300 responses, respectively, whereas in contrast humans assigned 6 and 5 to 732 and 2458 responses, respectively.

Cohen's unweighted kappa and quadratic weighted kappa (QWK) and their confidence intervals, calculated with the cohen.kappa function from the pscyh package[73] in R, are 0.082 (0.073, 0.091) and 0.437 (0.387, 0.487) respectively. These are both well below standard thresholds typically used to evaluate the agreement of parallel scores. In fact, the proportion reduction in mean squared error (PRMSE) for predicting the human true score[65,67] is:

$$PRMSE = 1 - \frac{E[(H-G)^2]}{\rho_H \cdot \mathrm{Var}(H)} = -0.617$$

where H is the random human score, G is the generated score, and ρ_H (set to 0.75 based on[71]) is the inter-rater reliability among randomly sampled human raters. This large negative value indicates that one would do substantially better assigning every essay the mean human score of 3.69 instead of the GPT-4o generated score, as the mean squared error for the GPT-4o generated scores is 61.7% larger than the mean squared error if we had simply assigned the mean human rating to all essays.

While the GPT-4o scores tended to be too low on average, they did tend to be higher when the humans gave higher scores. The disattenuated correlation (assuming human rater reliability of 0.75) is 0.757 and the squared disattenuated correlation is 0.574. The discrepancy between what the disattenuated correlation suggests, and what kappa, QWK, and PRMSE suggest is because correlation ignores the differences in means and the scaling of the AI-generated scores. So, if all one needs to do is rank order the essay writers, and the overall scale and location of the scores are not important (e.g., the holistic scoring guide descriptions are not important), the AI generated scores would be OK for low-stakes purposes of rank ordering students. However, if we want the scores to have some meaning, as suggested by the holistic scoring guide, then QWK and PRMSE should be metrics we use for evaluation, not disattenuated correlation.

Race/ethnicity fairness results

When AI scores are not accurate, or at least highly correlated with human ratings, there is an increased chance that they are differentially biased against specific demographic groups. Table 3 reports the means of the human ratings, the means of the GPT-4o ratings, their differences, and adjusted differences based on errors-in-variables regression[74,75] that adjust for the randomness in the human ratings[68,69].

In our example, for instance, we found that GPT-4o scoring negatively impacted essays labeled as written by Asian/Pacific Islander students. The adjusted difference for Asian/Pacific Islander students was -1.16. Compared to the overall mean difference of -0.90, the essays written by Asian/Pacific Islander students were an extra one-quarter of a point lower on average, even when adjusting for overall performance (see Adj. Diff column). Clearly, this does not seem fair.

Transparency and explainability

So, why is GPT-4o giving scores that are too low overall and particularly low for essays written by Asian/Pacific Islander students? The truth is it is tough to say. AI systems like ChatGPT are huge black-box algorithms defined by billions or trillions of parameters. These algorithms operate in ways that are often not fully understood even by their own developers. This lack of clarity makes it immensely difficult to explain why AI makes certain decisions, thereby undermining trust in the system.

While gaining a deep understanding of why an AI system like ChatGPT assigns one score or another is probably not possible at this stage, we can carry out some explorations. Given that the GPT-4o scoring algorithm appeared to be treating different race/ethnicity groups differently, we wondered if it would be able to predict the race/ethnicity of the essay writers, and if that prediction was also related to the score it gave.

We prompted GPT-4o to predict the raceethnicity of each essay writer by selecting from a list of the six groups in the original database. Once again, we utilized the OpenAI Batch API service (OpenAI, 2024). Appendix B shows a prototype entry in the JSONL file. The following system message was passed:

> *The race ethnicity of the writer of the essay below is one of: (1) American Indian/Alaskan Native; (2) Asian/Pacific Islander; (3) Black/African American; (4) Hispanic/Latino; (5) Two or more races/Other; (6) White. I want you to predict the race ethnicity of the author. The essay is:*

The student's essay was then appended to the end of the system message, and a user message was issued to request the AI to do the following:

> *Provide your prediction of the race ethnicity using the numeric codes above. Do not write any words in your response to me.*

The classification table appears in Table 4. The first thing to note is that GPT-4o responded with a predicted race category of "(". This is likely an artifact of our specific prompt, where we labeled the racial groups with numeric codes in the format(n) and requested a single token in the GPT response. In these cases, the model likely predicted that a response in the same format (e.g., (3) instead of 3) would have been optimal. While this behavior could likely be corrected with further prompt engineering, we decided to keep it as is as a reminder of

	Sample size	Human	GPT-4o	Diff.	Adj. Diff
American Indian/Alaskan Native	**52**	**3.37**	**2.54**	**-0.83**	**-0.77**
Asian/Pacific Islander	1193	4.25	3.19	−1.05	−1.16
Black/African American	2725	3.55	2.62	−0.93	−0.90
Hispanic/Latino	3022	3.41	2.52	−0.89	−0.84
Two or more races/Other	516	3.94	3.04	−0.90	−0.95
White	5613	3.77	2.92	−0.85	−0.87

Table 3. Mean scores assigned by humans and GPT-4o and their differences by subgroup.

Reported Race/Ethnicity	(	2	3	4	5	6	Recall %
American Indian/Alaskan Native (1)	**2**	**0**	**7**	**3**	**6**	**34**	***0.0***
Asian/Pacific Islander (2)	81	62	54	35	108	853	*5.2*
Black/African American (3)	90	6	950	46	421	1212	*34.9*
Hispanic/Latino (4)	79	5	302	709	571	1356	*23.5*
Two or more races/Other (5)	8	3	55	9	39	402	*7.6*
White (6)	150	7	198	52	514	4692	*83.6*
Precision %	–	*74.7*	*60.7*	*83.0*	*2.4*	*54.9*	

Table 4. Classification of GPT-4o's race/ethnicity predictions.

GPT-4o Predicted race	Human	GPT-4o	Diff.	Adj. Diff
(	3.55	2.68	−0.87	−0.84
Asian/Pacific Islander (2)	4.13	2.82	−1.31	−1.42
Black/African American (3)	3.19	2.21	−0.98	−0.86
Hispanic/Latino (4)	3.22	2.14	−1.08	−0.97
Two or more races/Other (5)	3.06	2.33	−0.73	−0.58
White (6)	3.96	3.06	−0.90	−0.96

Table 5. Mean human and GPT-4o ratings by GPT-40 predicted race/ethnicity.

how prompts can affect results.In terms of exact agreement, GPT-4o did a better job predicting race/ethnicity (49% exact agreement, 51% when ignoring "(" responses), than it did scoring the essay (30% exact agreement). While this may seem impressive, it is mostly because a larger proportion of essays were predicted to have been written by White students, who make a large proportion of the sample. Cohen's kappa, which accounts for chance agreement, is 0.26 in this case after removing the "(" generated responses compared to 0.08 for the scores; the 95% confidence interval for kappa for the race guess is (0.25, 0.27) compared to (0.07, 0.09) for scores.

Furthermore, we found that, as shown in Table 5, even after adjusting for overall performance differences, essays that GPT-4o predicted were written by Asian/Pacific Islander students were scored half a point lower on average (adjusted difference -1.42 compared to average difference of -0.90); essays predicted to be Two or more races or Other were scored to three-tenths of a point higher (adjusted difference -0.58 compared to mean difference of -0.90).

In summary, the black-box model in GPT-4o can better predict the ethnicity of a writer than to predict a score on an essay. This suggests that we cannot rule out the possibility that GPT-4o infers, captures, and uses (construct-irrelevant) race/ethnicity information from the essays for the scoring task. Writer's ethnicity background should not be considered in determining essay quality. Given the biases reported in Table 3, it appears that whatever aspects of the essay that are used to predict race/ethnicity are also associated with the aspects that cause fairness issues. Unfortunately, it is difficult to determine exactly what those aspects are. Future research is needed.

Discussion

While AI technologies like ChatGPT hold great promise for easing the grading burden and enhancing educational practices, their responsible application requires a holistic approach. Ensuring fairness, transparency, explainability, and continuous improvement are critical to fostering an inclusive, reliable, and effective educational environment. The principles outlined in standards for responsible AI and in literature related to the problem of scoring, such as[1] and[64] offer valuable frameworks for addressing these challenges. Prior work[76,77] on formulating validity arguments for automated scoring in educational assessment is still relevant today that "it's not only the scoring." As we continue to integrate AI into education, it is our collective responsibility to safeguard these principles to truly benefit all students and educators.

In this work, we set out to examine the zero-shot scoring capabilities of GPT-4o with a primary focus on the three key facets of responsible AI: accuracy, fairness, and explainability. Our specific experiment looked at the viability of providing holistic scores for the sample of independent essays available in the Persuade 2.0 data set[71] and found a number of issues that should make one think twice about zero-shot approaches for scoring essays without careful evaluation.

We addressed two research questions. For the first research question, we found that the zero-shot AI generated scores were not accurate when compared to the human ratings in the data set. On average the AI-generated scores were almost a point lower and came nowhere close to meeting industry standards for quality of automated scores. In fact, the evaluation metrics PRMSE suggests you would be better off simply assigning the mean score to every student. Additionally, we found that scoring discrepancies were not consistent across race/ethnicity groups. Compared to other essays, those associated with Asian/Pacific Islander students were more biased (relative to human ratings) than the other race/ethnicity groups. For the second research question, we

also found that zero-shot AI was able to better predict one's race/ethnicity than it was to predict the correct score, and that those predicted races were associated with the scores that GPT-4o produced, even after accounting for true human score differences. This result strongly suggests the possibility of GPT-4o capturing and utilizing race/ethnicity information from the essays for the scoring task, for which information should be irrelevant to the construct of measurement.

These findings highlight the importance of rigorously evaluating automated scoring methods, no matter the approach taken. Automated scoring must be evaluated, not only for accuracy, but also for fairness and explainability. Explicit efforts are needed to mitigate bias in AI models. Using zero-shot generative AI approaches makes it difficult to mitigate bias, given one does not necessarily have access to the inner workings of the model. It may be possible to improve performance using clever prompt engineering approaches, but more research is needed.

Furthermore, transparency and explainability are essential. Stakeholders, including students, educators, and administrators, must understand how AI algorithms arrive at their conclusions. Transparency helps to foster trust and accountability. It also enables users to identify potential problems and biases. One potential approach to explainable AI in the context of generative AI automated scoring is the self-explanation, where, in addition to asking the AI for a score, you also ask for an explanation for the score. However, just as the scores need to be evaluated for accuracy and fairness, so too should these explanations. Future research is encouraged to explore how to leverage LLMs to extract meaningful features, which will improve the explainability of the AI scores. As a related possibility, one may combine a set of hand-engineered, construct-relevant features with a "score" from black-box models to generate a final writing score. The resulting scoring models not only will be more explainable but may also lead to comparable or even better prediction accuracy than pure black-box models. It is worth noting that if features (as implicit or explicit indicators of text quality) are problematic introducing construct-irrelevant information associated with group membership, the models will be biased whether it is a white-box or a black-box models. Related to bias detection and mitigation, previous research has proposed fairness-constrained models, e.g., in[68] and[9], where a penalty multiplier was introduced to algorithmically control the differences in scoring model performance on the subgroup population level. Though, those methods were developed for simpler scoring models (e.g., best linear predictor[78] or least square regression) that used a small set of theory-driven, hand-engineered features. More research will be needed to scale up those previous work to non-linear models with a very large number of parameters. Algorithms also exist in the broader AI field, such as[79], that deal with complex algorithms but more research will be needed to evaluate and adopt those methods for AI scoring under the educational assessment context.

While this study examined some of the critical issues related to the use of generative AI for scoring in educational assessment, it is not without limitations. For one, we did not do prompt engineering to find a prompt that produced the best results possible. We used a single prompt that we thought was reasonable and tweaked it a little to get out numeric scores to simplify analysis. It is perfectly possible that a different prompting strategy might produce different results. Second, we only focused on a single zero-shot generative AI approach. Few-shot generative AI approaches and/or fine-tuning approaches would likely yield better results. However, the goal of this paper was to focus more on how to evaluate scores, rather than how to best use AI for scoring and optimize performance. Finally, because we do not have access to the inner workings of GPT-4o, we were limited in what we could do to try and explain results. Attempting to predict the sensitive attribute that one wants to ensure fairness - in the current study, the race/ethnicity of the student, was one approach to try and see if the AI model were capturing features associated with race and ethnicity, but there could be others.

Data availability

The data set used in this research can be found at: https://github.com/scrosseye/persuade_corpus_2.0?tab=readme-ov-file

Received: 17 August 2024; Accepted: 6 November 2024
Published online: 03 December 2024

References

1. Bennett, R. E. & Zhang, M. Validity and automated scoring. In Drasgow, F. (ed.) *Technology in Testing: Measurement Issues*, NCME Applications of Educational Measurement and Assessment Series, 142–173 (Taylors & Francis, 2016).
2. Weigle, S. C. *Assessing writing* (Cambridge University Press, New York, NY, 2002).
3. Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E. & Kukich, K. Comparing the validity of automated and human essay scoring. Tech. Rep. RR-98-08, Educational Testing Service, Princeton, NJ (1998).
4. Zhang, M. Contrasting automated and human scoring of essays. Tech. Rep. RDC-21, Educational Testing Service, Princeton, NJ (2013).
5. Page, E. B. The imminence of grading essays by computer. *Phi Delta Kappan* **47**, 238–243 (1966).
6. Dikli, S. An overview of automated essay scoring. *The Journal of Technology, Learning, Assessment***5** (2006).
7. Burrows, S., Gurevych, I. & Stein, B. The eras and trends of automatic short answer grading. *Int. J. Artif. Intell. Educ.* **25**, 60–117 (2015).
8. Wilson, J. et al. Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Comput. Educ.* **168**, 104208 (2015).
9. Yao, L., Haberman, S. & Zhang, M. Penalized best linear prediction of true test scores. *Psychometrika* **84**, 186–211 (2018).
10. Heilman, M. & Madnani, N. Ets: Domain adaptation and stacking for short answer scoring. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (Second Joint Conference on Lexical and Computational Semantics)*, 2, 275–279 (Association for Computational Linguistics, 2013).
11. McNamara, D. S., Crossley, S. A., Roscoe, R., Allen, L. K. & Dai, J. A hierarchical classification approach to automated essay scoring. *Assess. Writ.* **23**, 35–59 (2015).
12. Elalfi, A. E. E., Elgamal, A. F. & Amasha, N. A. Automated essay scoring using word2vec and support vector machine. *Int. J. Comput. Appl.* **117**, 20–29 (2019).

13. Haberman, S. & Sinharay, S. The application of the cumulative logistic regression model to automated essay scoring. *J. Educational Behav. Stat.* **35**, 586–602 (2010).
14. Basu, S., Jacobs, C. & Vanderwende, L. Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the ACL* (2013).
15. Foltz, P. W., Streeter, L. A., Lochbaum, K. E. & Laudauer, T. K. Implementation and applications of the intelligent essay assessor. In Shermis, M. & Burstein, J. (eds.) *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (Routledge, New York, 2013), 1st edn.
16. Lubis, F. F. M. et al. Automated short-answer grading using semantic similarity based on word embedding. *Int. J. Technol.* **12**, 571–581 (2021).
17. Uto, M., Xie, Y. & Ueno, M. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th international conference on computational linguistics* (Barcelona, Spain, 2020).
18. Bejar, I. I., Williamson, D. M. & Mislevy, R. J. Human scoring. In *Automated scoring of complex tasks in computer-based testing, 49–79 (Laurence Erlbaum Associates* (eds Williamson, D. M. et al.) (Mahwah, NJ, 2006).
19. Page, E. B. Computer grading of student prose using modern concepts and software. *J. Exper. Edu.* **62**, 127–142 (1994).
20. Page, E. B. The computer moves into essay grading. *Phi Delta Kappan* **76**, 561–566 (1995).
21. Ben-Simon, A. & Bennett, R. E. Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment***6** (2007).
22. Foltz, P. W., Laham, D. & Landauer, T. K. Automated essay scoring: applications to educational technology. In Collis, B. & Oliver, R. (eds.) *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (AACE, Chesapeake, VA, 1999).
23. Landauer, T. K., Foltz, P. W. & Laham, D. Introduction to latent semantic analysis. *Discourse Process.* **25**, 259–284 (1998).
24. Landauer, T. K., Laham, D. & Foltz, P. Automatic essay assessment. *Assess. Edu.* **10**, 295–308 (2003).
25. Srihari, S., Collins, J., Srihari, R., Babu, P. & Srinivasan, H. Automated scoring of handwritten essays based on latent semantic analysis. In *Document Analysis Systems VII, 7th International Workshop, DAS 2006, Nelson, New Zealand, February 13–15, 2006, Proceedings, 71–83 (Springer Verlag* (eds Bunke, H. & Spitz, A. L.) (Kassel, Germany, 2006).
26. Foltz, P. W., Gilliam, S. & Kendall, S. A. Supporting content-based feedback in online writing evaluation with lsa. *Interactive Learning Environ.* **8**, 111–129 (2000).
27. Mason, O. & Grove-Stephenson, I. Automated free text marking with paperless school. In Danson, M. (ed.) *Proceedings of the sixth International Computer Assisted Assessment Conference* (Loughborough, U.K., 2002).
28. Chowdhury, G. G. Natural language processing. *Ann. Rev. Inf. Sci. Technol.* **37**, 51–89 (2003).
29. Attali, Y. & Burstein, J. Automated essay scoring with e-rater v. 2.0. *The Journal of Technology, Learning, and Assessment***4** (2006).
30. Williamson, D. M., Xi, X. & Breyer, F. J. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice* 2–13 (2012).
31. Oka, R., Kusumi, T. & Utsumi, A. Performance evaluation of automated scoring for the descriptive similarity response task. *Nat. Sci. Rep.* **14**, 6228 (2024).
32. Thakur, N., Reimers, N., Daxenberger, J. & Gurevych, I. Augmented sbert: Data augmentation method for improving bi-encoder for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 296–310 (ACL, 2021).
33. Haller, S., Aldea, A., Seifert, C. & Strisciuglio, N. Survey on automated short answer grading with deep learning: from word embeddings to transformers. https://doi.org/10.48550/arXiv.2204.03503 (2022).
34. Amur, Z. H. & Hooi, Y. K. State-of-the-art: Assessing semantic similarity in automated short-answer grading systems. *Information Sci. Lett.* **1**, 1851–1858 (2022).
35. Zhang, M., Baral, S., Heffernan, N. & Lan, A. Automatic short math answer grading via in-context meta-learning. In *Proceedings of the 2022 Educational Data Mining Conference* (UK, 2022).
36. Ali, S. et al. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion* **99**, 101805 (2023).
37. Abuhmed, T. et al. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion* **99**, 101805 (2023).
38. Zhang, M. & Deane, P. Process features in writing: Internal structure and incremental value over product features. Tech. Rep. RR-15-27, Educational Testing Service, Princeton, NJ (2015).
39. Sinharay, S., Zhang, M. & Deane, P. Application of data mining methods for predicting essay scores from writing process and product features. *Appl. Measur. Educ.* **32**, 116–137 (2019).
40. awsaf et al. Aes 2.0: Kerasnlp starter. https://www.kaggle.com/code/awsaf49/aes-2-0-kerasnlp-starter (2024).
41. Heilman, M. & Madnani, N. The impact of training data on automated short answer scoring performance. In Tetreault, J., Burstein, J. & Leacock, C. (eds.) *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 81–85 (Association for Computational Linguistics, Denver, Colorado, 2015).
42. Yao, L., Cahill, A. & McCaffrey, D. F. The impact of training data quality on automated content scoring performance. In *Association for the Advancement of Artificial Intelligence* (New York, 2020).
43. Whitmer, J. *et al.* Results of naep reading item automated scoring data challenge (fall 2021). https://osf.io/preprints/edarxiv/2hevq (2021).
44. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
45. Navigli, R., Conia, S. & Ross, B. Biases in large language models: origins, inventory, and discussion. *J. Data Information Quality* **15**, 1–21 (2023).
46. Ayoub, N. F. et al. Inherent bias in large language models: A random sampling analysis. *Mayo Clinic Proc. Digital Health* **2**, 186–191 (2024).
47. Ma, W., Scheible, H., Wang, B. & Veeramachaneni, G. Deciphering stereotypes in pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 11328–11345 (2023).
48. Venkit, P. N., Srinath, M. & Wilson, S. A study of implicit language model bias against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, 1324–1332 (2022).
49. Kotek, H., Dockum, R. & Sun, D. Q. Gender bias and stereotypes in large language models. In *Collective Intelligence Conference (CI'23)*, 13 (ACM, Delft, Netherlands, 2023).
50. Mohan, G. B. et al. An analysis of large language models: their impact and potential applications. *Knowl. Inf. Syst.* **66**, 5047–5070 (2024).
51. Nadeem, M., Bethke, A. & Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 5356–5371 (online, 2021).
52. Kurita, K., Vyas, N., Pareek, A., Black, A. W. & Tsvetkov, Y. Measuring bias in contextualized word representations. In Costa-jussà, M. R., Hardmeier, C., Radford, W. & Webster, K. (eds.) *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172 (Association for Computational Linguistics, Florence, Italy, 2019).
53. Zhu, Y. *et al.* Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*, 19–27 (Santiago, Chile, 2015).

54. He, P., Gao, J. & Chen, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. https://arxiv.org/pdf/2111.09543 (2023).
55. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 14 (ACM, Virtual Event, Canada, 2021).
56. Tate, T. P. et al. Can AI provide useful holistic essay scoring?. *Comput. Edu. Artif. Intell.* **7**, 100255 (2024).
57. ETS. Responsible use of AI in assessment. Tech. Rep., Educational Testing Service, Princeton, NJ (2024).
58. Kortemeyer, G. Performance of the pre-trained large language model gpt-4 on automated short answer grading. *Discover Artif. Intell.* **4**, 47 (2024).
59. Lee, G.-G., Latif, E., Wu, X., Liu, N. & Zhai, X. Applying large language models and chain-of-thought for automatic scoring. *Comput. Edu. Artif. Intell.* **6**, 100213 (2024).
60. Tobler, S. Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments. *MethodsX* **12**, 102531 (2024).
61. Steiss, J. et al. Comparing the quality of human and chatgpt feedback of students' writing. *Learn. Instr.* **91**, 101894 (2024).
62. Arrieta, A. B. et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* **58**, 82–115 (2020).
63. Modarressi, A., Fayyaz, M., Aghazadeh, E., Yaghoobzadeh, Y. & Pilehvar, M. T. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)* (2023).
64. ETS. Best practices for constructed-response scoring. Tech. Rep., Educational Testing Service, Princeton, NJ (2021).
65. Haberman, S. Measures of agreement versus measures of prediction accuracy. *ETS Res. Report Series* **2019**, 1–23 (2019).
66. Cohen, J. Weighted kappa: Normal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**, 213–220 (1968).
67. Loukina, A. *et al.* Using prmse to evaluate automated scoring systems in the presence of label noise. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 18–29 (ACL, Seattle, WA, USA, 2020).
68. Johnson, M. S., Liu, X. & McCaffrey, D. F. Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *J. Educ. Meas.* **59**, 338–361 (2022).
69. Johnson, M. S. & McCaffrey, D. F. Evaluating fairness of automated scoring in educational measurement. In Yaneva, V. & von Davier, M. (eds.) *Advancing Natural Language Processing in Educational Assessment* (Routledge, New York, 2023), 1st edn.
70. Bibal, A., Marion, R., von Sachs, R. & Frenay, B. Biot: Explaining multidimensional nonlinear mds embeddings using the best interpretable orthogonal transformation. *Neurocomputing* **453**, 109–118 (2021).
71. Crossley, S. Persuade_corpus_2.0. github. https://github.com/scrosseye/persuade_corpus_2.0 (2024).
72. Crossley, S. A. *et al.* A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing***61**, 100865 (2024).
73. Revelle, W. *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, Illinois, USA.
74. Fuller, W. A. *Measurement Error Models* (John Wiley & Sons, 1987).
75. Iwata, S. Errors-in-variables regression using estimated latent variables. *Economet. Rev.* **11**, 195–200 (1992).
76. Bennett, R. E. & Bejar, I. I. Validity and automated scoring: It's not only the scoring. *Educ. Meas. Issues Pract.* **17**, 9–17 (1998).
77. Bennett, R. E. Moving the field forward: Some thoughts on validity and automated scoring. Research Memorandum RM-04-01, Educational Testing Service, Princeton, NJ (2004).
78. Haberman, S. J. Application of best linear prediction and penalized best linear prediction to ETS tests. Tech. Rep. RR-20-08, Educational Testing Service, Princeton, NJ (2020).
79. Bird, S. *et al.* Fairlearn: A toolkit for assessing and improving fairness in ai. Tech. Rep. MSR-TR-2020-32, Microsoft (2020).

Author contributions

Matthew Johnson conceptualized and conducted the study. Both authors contributed to the manuscript writing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-79208-2.

Correspondence and requests for materials should be addressed to M.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.