Investigating the Potential of Large Language Models for Automated Writing Scoring

Shan Wang

School of Foreign Languages and Literature, Beijing Normal University 100875, Beijing, China

Boobi2000@163.com

Abstract. This study investigates the potential of large language models (LLMs), specifically GPT-4, for automated writing scoring and feedback generation. Employing a mixed-methods approach, the research evaluates the accuracy and reliability of GPT-4 in predicting essay scores and the quality of its generated feedback. The results demonstrate a high level of agreement between GPT-4 scores and human raters, as evidenced by the confusion matrix and Quadratic Weighted Kappa metric. Qualitative analysis of GPT-4 feedback suggests its ability to provide constructive and comprehensive suggestions for improving student writing. However, there are still limitations surrounding LLM-based automated scoring and feedbacks. Thus, this study proposes the use of LLM-based systems as formative assessment tools to complement human judgment.

Keywords: Automated Writing Scoring, Large Language Models, GPT-4, Feedback Generation, Writing Assessment.

1 Introduction

Automated writing scoring (AWS) systems have attracted considerable interest in recent years due to their capacity to alleviate the burden of manual essay grading. These systems employ various computational methods to evaluate textual features and estimate essay scores [1]. They offer the potential for consistent and efficient scoring and thus they can reduce the time and resources necessary for manual grading.

Early AWS systems employed simple linear regression models using surface text features like essay length, word counts, punctuation, and vocabulary to predict scores [2] [3]. However, Critics argued that these early systems failed to capture the essence of writing competence and focus only on superficial aspects rather than the substantive qualities of good writing [4] [5]. The recent emergence of large language models (LLMs) has opened up new possibilities for AWS research. LLMs, such as GPT-4, have exhibited impressive capabilities in natural language understanding and generation. The potential of LLMs for AWS resides in their ability to take into account a broad range of linguistic features, contextual information, and domain knowledge when assessing essays [6].

Y. Kuang et al. (eds.), *Proceedings of the 2024 5th International Conference on Education, Knowledge and Information Management (ICEKIM 2024)*, Atlantis Highlights in Computer Sciences 22,
https://doi.org/10.2991/978-94-6463-502-7_116

This study aims to explore the potential of LLMs, particularly GPT-4, for automated writing scoring and feedback generation. We concentrate on two primary aspects: (1) the accuracy and reliability of GPT-4 in predicting essay scores, and (2) the quality and utility of GPT-4 generated feedback for improving student writing. To accomplish these objectives, we adopt a mixed-methods approach, including quantitative evaluation of scoring performance with qualitative analysis of feedback quality.

2 Research Methodologies

2.1 Data Collection

The dataset chose in this study is ASAP (Automated Student Assessment Prize) dataset This dataset consists of essays written by students spanning Grades 7 to 10, and each essay is accompanied by human-assigned scores for reference. To evaluate the performance of LLMs in automated scoring, we employed the GPT-4 model to generate scores for 600 essays randomly selected from the dataset. The essays were chosen to ensure a representative distribution across the eight writing tasks and score ranges.

In this research, The GPT-4 model was prompted using meticulously crafted instructions that included the essay text, the corresponding writing task, and the scoring rubric. The model's output was then compared to the human-assigned scores. In addition to the quantitative assessment of scoring accuracy, we also conducted a qualitative analysis of the GPT-4 generated feedback for a small sample of essays.

3 LLMs' Capabilities in Automatic Scoring

3.1 Automatic Scoring Based on Rubric

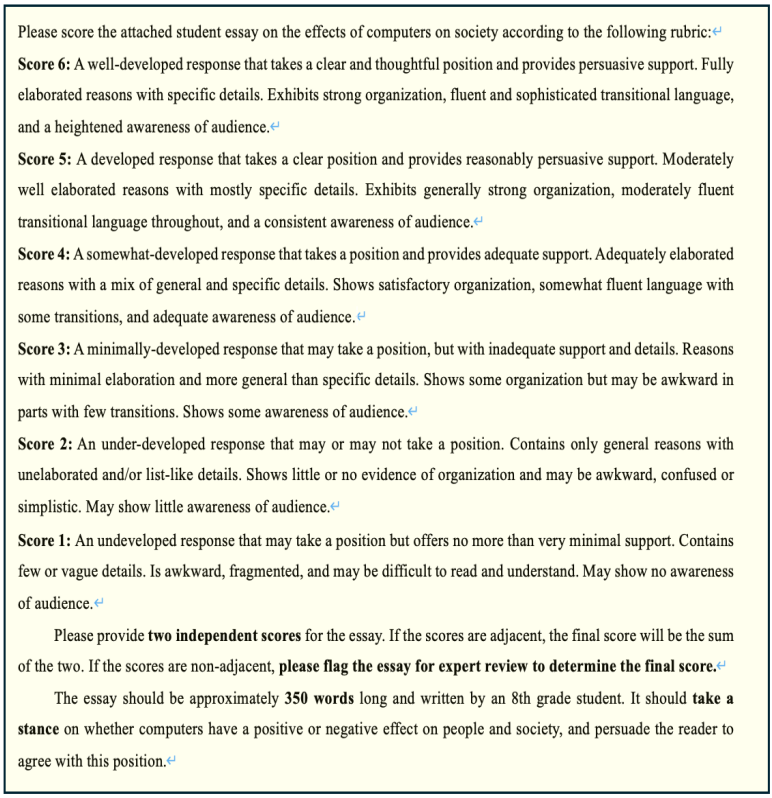
Previous Human essay scoring, although widely used in various writing assessments, has notable limitations. The subjectivity and inconsistency of raters can lead to score discrepancies for the same essay across different raters [7]. Moreover, rater fatigue and errors can also im-pact the quality of scoring [8]. AES systems can mitigate some limitations of human scoring by providing a more consistent, objective, and efficient approach to grading essays through the application of predefined criteria and algorithms [1] [9].

For the performance of large language model-based automated essay scoring systems, the quality of "prompting" is crucial. The key idea be-hind prompt engineering is to design carefully crafted prompts that guide the LLM to generate scores based on the specific criteria defined in the rubric. This approach allows for a more direct mapping between the scoring rubric and the generated scores, addressing the concern of construct underrepresentation in traditional automated scoring systems.

To illustrate how prompt engineering works in automatic scoring, we take the example of using GPT-4 to grade argumentative essays. First, it is crucial to clearly define the scoring criteria, which form the foundation for designing high-quality prompts. Scoring criteria typically encompass multiple dimensions, such as clarity of the essay,

logic of argumentation, relevance and sufficiency of evidence, organizational structure, accuracy and fluency of language, etc. Each dimension requires detailed descriptions that specify the characteristics of different performance levels.

When designing the prompts, we referred to the scoring rubric provided in the ASAP dataset. This rubric divide essay performance into six levels, from 1 to 6 points, with corresponding descriptions for each level. We input these descriptions into the prompts as the basis for GPT-4 to generate scores. In addition to the scoring criteria, the prompts should include key information, such as the writing prompt, task instructions, and the author's grade level. Figure 1 is a complete example of a prompt:


Please score the attached student essay on the effects of computers on society according to the following rubric:

Score 6: A well-developed response that takes a clear and thoughtful position and provides persuasive support. Fully elaborated reasons with specific details. Exhibits strong organization, fluent and sophisticated transitional language, and a heightened awareness of audience.

Score 5: A developed response that takes a clear position and provides reasonably persuasive support. Moderately well elaborated reasons with mostly specific details. Exhibits generally strong organization, moderately fluent transitional language throughout, and a consistent awareness of audience.

Score 4: A somewhat-developed response that takes a position and provides adequate support. Adequately elaborated reasons with a mix of general and specific details. Shows satisfactory organization, somewhat fluent language with some transitions, and adequate awareness of audience.

Score 3: A minimally-developed response that may take a position, but with inadequate support and details. Reasons with minimal elaboration and more general than specific details. Shows some organization but may be awkward in parts with few transitions. Shows some awareness of audience.

Score 2: An under-developed response that may or may not take a position. Contains only general reasons with unelaborated and/or list-like details. Shows little or no evidence of organization and may be awkward, confused or simplistic. May show little awareness of audience.

Score 1: An undeveloped response that may take a position but offers no more than very minimal support. Contains few or vague details. Is awkward, fragmented, and may be difficult to read and understand. May show no awareness of audience.

Please provide **two independent scores** for the essay. If the scores are adjacent, the final score will be the sum of the two. If the scores are non-adjacent, **please flag the essay for expert review to determine the final score.**

The essay should be approximately **350 words** long and written by an 8th grade student. It should **take a stance** on whether computers have a positive or negative effect on people and society, and persuade the reader to agree with this position.

Fig. 1. A complete example of a prompt for automatic writing scoring.

To verify the accuracy and reliability of automatic scoring by large language models, we used 600 student argumentative essays from the ASAP dataset and compared the human scores with the machine scores generated by GPT-4. These essays are accompanied by scores provided by trained human raters, which can serve as the gold standard for evaluating the quality of automatic scoring.

Table 1. Confusion Matrix of human and GPT-4 scoring based on the provided rubric (n=600).

Human/GPT	1	2	3	4	5	6
1	0.85	0.12	0.03	0	0	0
2	0.08	0.78	0.12	0.02	0	0
3	0	0.15	0.7	0.13	0.2	0
4	0	0	0.1	0.75	0.14	0.01
5	0	0	0	0.08	0.8	0.12
6	0	0	0	0	0.05	0.95

The confusion matrix presented in Table 1 provides a detailed breakdown of the agreement between GPT-4 automated scoring and human scoring across six score points based on the given rubric.

The confusion matrix can be used to analyze in detail the scoring agreement and discrepancies between the GPT-4 and manual raters on each score level. Overall, the values on the diagonal of the matrix represent the percentage of perfect agreement between the GPT-4 and manual raters on each score level. In Table 1, the results of the study show a high degree of agreement, ranging from 70% to 95% across score levels.

Specifically, the highest agreement (95%) was observed on the 6-point scale, which suggests that the GPT-4 has a higher accuracy in identifying texts with the highest writing ability. Also noteworthy was the consistency of scores on the 1- and 5-point scales, which were 85% and 80% respectively. These findings suggest that the GPT-4 is also effective in distinguishing between essays in the lowest and highest score ranges.

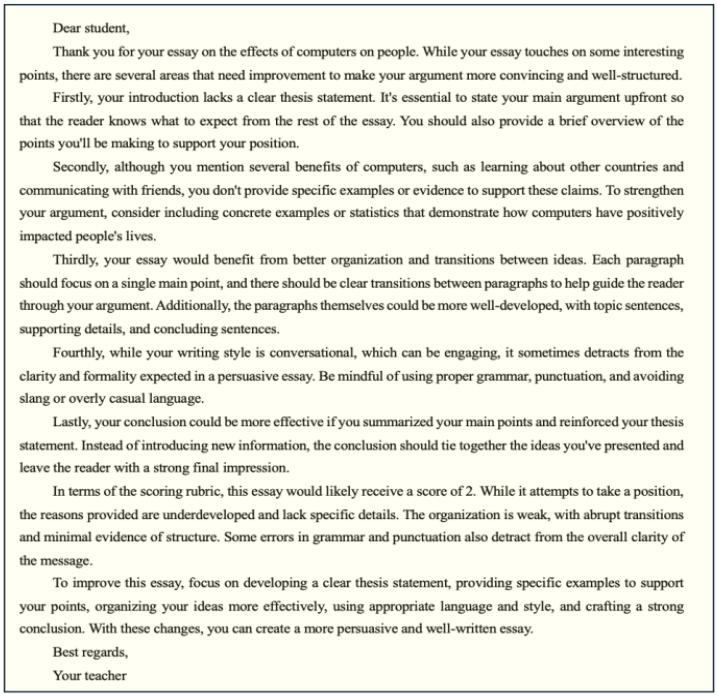
What's more, the values on the non-diagonal lines of the matrix represent the differences between the GPT-4 and manual scores. Notably, most of these differences are within adjacent score bands, suggesting that the GPT-4 rarely deviates substantially from the manually scored score bands.

Calculating the sum of exact agreements across all score point, the overall accuracy of GPT-4 scoring is 80.5%. This high level of accuracy is further corroborated by the Quadratic Weighted Kappa (QWK) coefficient of 0.85. The QWK is a widely used metric for evaluating the agreement between automated and human scoring [10] [11]. The typical threshold of 0.70 is considered as a standard for a reliable automated scoring system [11]. Thus, a QWK value of 0.85 indicates a strong agreement between GPT-4 and human raters.

3.2 Automatic Feedback

In addition to providing automated grading, the LLMs can also automatically generate personalized feedback. One of its main advantages over manual scoring is its ability to provide rich and relevant advice based on the specific strengths and weaknesses of each essay. Since traditional feedback systems tend to provide generic scoring criteria based on predefined rubrics [12], they often lack detailed and flexible advices. However, by harnessing the power of natural language processing and machine learning, LLMs can analyze more aspects of student writing, such as content, organization, style, and grammar usage, and provide targeted suggestions for improvement. As shown in

Figure 2, we investigate the ability of large-scale models to automatically generate writing feedback.



Dear student,

Thank you for your essay on the effects of computers on people. While your essay touches on some interesting points, there are several areas that need improvement to make your argument more convincing and well-structured.

Firstly, your introduction lacks a clear thesis statement. It's essential to state your main argument upfront so that the reader knows what to expect from the rest of the essay. You should also provide a brief overview of the points you'll be making to support your position.

Secondly, although you mention several benefits of computers, such as learning about other countries and communicating with friends, you don't provide specific examples or evidence to support these claims. To strengthen your argument, consider including concrete examples or statistics that demonstrate how computers have positively impacted people's lives.

Thirdly, your essay would benefit from better organization and transitions between ideas. Each paragraph should focus on a single main point, and there should be clear transitions between paragraphs to help guide the reader through your argument. Additionally, the paragraphs themselves could be more well-developed, with topic sentences, supporting details, and concluding sentences.

Fourthly, while your writing style is conversational, which can be engaging, it sometimes detracts from the clarity and formality expected in a persuasive essay. Be mindful of using proper grammar, punctuation, and avoiding slang or overly casual language.

Lastly, your conclusion could be more effective if you summarized your main points and reinforced your thesis statement. Instead of introducing new information, the conclusion should tie together the ideas you've presented and leave the reader with a strong final impression.

In terms of the scoring rubric, this essay would likely receive a score of 2. While it attempts to take a position, the reasons provided are underdeveloped and lack specific details. The organization is weak, with abrupt transitions and minimal evidence of structure. Some errors in grammar and punctuation also detract from the overall clarity of the message.

To improve this essay, focus on developing a clear thesis statement, providing specific examples to support your points, organizing your ideas more effectively, using appropriate language and style, and crafting a strong conclusion. With these changes, you can create a more persuasive and well-written essay.

Best regards,

Your teacher

Fig. 2. An example of a automatic writing feedback generated by GPT-4.

The AI-generated writing feedback provided by the language model demonstrates its ability to offer comprehensive and constructive feedback. Table 2 reflects the distribution of various evaluation aspects emphasized in the 600 writing feedback pieces generated by GPT-4.

Table 2. Components of AI-generated Feedback.

Feedback Evaluation Aspect	Frequency
Sufficient evidence and use of specific examples	311
Reasonable organization and coherence	296
Appropriate language and style for academic writing	275
Clear and compelling thesis statement	232
Strong and memorable conclusion	188

As shown in Table 2, the feedback touches upon crucial elements of effective writing: GPT-4 most frequently stressed the importance of using specific examples to support arguments [13]. In this point, GPT-4 guides students towards developing well-

substantiated arguments that can withstand scrutiny and contribute meaningfully to the scholarly discourse. The automated feedback also recognizes the importance of organization and smooth transitions in guiding the reader through the essay [14]. The appropriateness of language and style [15] was also a key focus, which helps students cultivate a mature and professional writing voice for academic discourse. Clear thesis statements were also emphasized [16]. GPT-4 encourages students to invest time and effort in crafting a precise and engaging thesis. A strong conclusion is also crucial in GPT's evaluation [4].

4 Evaluation

The results of this study demonstrate the remarkable potential of large language models, specifically GPT-4, in AWS and feedback generation. The high level of agreement between GPT-4 scores and human raters indicates that LLMs can effectively emulate human judgment in essay scoring and feedback generation. Moreover, the qualitative analysis of GPT-4 generated feedback suggests that LLMs can provide constructive suggestions for improving student writing. This feature of LLM-based AWS has the potential to positively impact students' writing development, particularly in L2 academic set-tings [17] [18].

However, it is essential to consider the limitations and concerns sur-rounding AWS systems, which may also apply to LLM-based approaches. One key issue is the potential discrepancy between the constructs and con-texts of high-stakes assessments and classroom writing [5] [19]. While GPT-4 has shown promising results in this study, further research is needed to examine its performance across a wider range of writing tasks and educational settings. Moreover, although LLMs have demonstrated the ability to assess semantic aspects of writing, their capacity to evaluate higher-order skills, such as creativity and critical thinking, remains uncertain. Furthermore, the "black box" nature of the scoring process in AWS systems [4] raises questions about transparency and interpretability. While the attention visualization technique used in this study provides some insights into the model's decision-making process, further efforts are needed to enhance the explainability of LLM-based AWS. Ethical considerations, such as data privacy, fairness, and potential biases, must also be addressed when implementing LLM-based AWS.

5 Conclusions

This study demonstrates the promising potential of large language models, particularly GPT-4, in automated writing scoring and feedback generation. The high agreement between GPT-4 scores and human raters and the qualitative analysis of GPT-4 generated feedback highlight its ability to effectively capture various aspects of writing quality and provide comprehensive suggestions for improvement. However, the limitations and ethical considerations surrounding LLM-based AWS systems must be carefully addressed. We propose the judicious use of LLM-based AWS as formative assessment

tools to complement human judgment, ultimately supporting student learning and growth.

References

1. Dikli, S.: An overview of automated scoring of essays. Journal of Technology, Learning, and Assessment 5(1), 1-35 (2006).
2. Page, E.: The imminence of grading essays by computer. The Phi Delta Kappan 47(5), 238-243 (1966).
3. Page, E.B., Petersen, N.S.: The computer moves into essay grading: Updating the ancient test. The Phi Delta Kappan 76(7), 561-565 (1995).
4. Ericsson, P.F., Haswell, R.H. (Eds.): Machine scoring of student essays: Truth and consequences. Utah State University Press (2006).
5. Condon, W.: Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? Assessing Writing 18(1), 100-108 (2013).
6. Nadeem, F., Nguyen, H., Liu, Y., Ostendorf, M.: Automated essay scoring with discourse-aware neural models. In: Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 399-409 (2021).
7. Wang, J., Engelhard, G., Raczynski, K., Song, T., Wolfe, E.W.: Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. Assessing Writing 18(1), 36-47 (2013).
8. Ling, G., Mollaun P., Xi, X. A study on the impact of fatigue on human raters when scoring speaking responses. Language Testing 31(4), 479-499 (2014).
9. Shermis, M.D., Burstein, J. (Eds.): Handbook of automated essay evaluation: Current applications and new directions. Routledge (2013).
10. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin 70(4), 213-220 (1968).
11. Williamson, D.M., Xi, X., Breyer, F.J.: A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice 31(1), 2-13 (2012).
12. Stevenson, M., Phakiti, A.: The effects of computer-generated feedback on the quality of writing. Assessing Writing 19, 51-65 (2014).
13. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. Behavior research methods, instruments, & computers 36(2), 193-202 (2004).
14. Crossley, S.A., Kyle, K., McNamara, D.S.: The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behavior research methods 48(4), 1227-1237 (2016).
15. McNamara, D.S., Crossley, S.A., McCarthy, P.M.: Linguistic features of writing quality. Written communication 27(1), 57-86 (2010).
16. Hyland, K.: Second language writing. Ernst Klett Sprachen (2003).
17. AbuSeileek, A.F.: Using track changes and word processor to provide corrective feedback to learners in writing. Journal of Computer Assisted Learning 29(4), 319-333 (2013).
18. Chen, C.F.E., Cheng, W.Y.E.: Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. Language Learning & Technology 12(2), 94-112 (2008).
19. Weigle, S.C.: Assessing writing. Cambridge University Press (2002).