Are the Reasoning Models Good at Automated Essay Scoring?

Anonymous ACL submission

Abstract

This study investigates the performance of reasoning models (OpenAI's o1-mini and o3-mini) in automated essay scoring (AES) tasks. While these models demonstrate superior performance across various benchmarks, their effectiveness in AES applications remains unexplored. We conducted two experiments using the TOEFL11 dataset: (1) examining scoring consistency by having models evaluate identical essays 50 times, and (2) comparing their scoring accuracy against human expert assessments using Quadratic Weighted Kappa (QWK). Our results reveal that conventional models like GPT-4o mini outperform newer reasoning models in AES tasks, achieving significantly higher QWK scores (0.619 vs 0.454 and 0.442). Additionally, we found that reasoning models show scoring inconsistencies. These findings suggest that benchmark performance improvements may not translate directly to specialized tasks like essay evaluation, highlighting the importance of task-specific assessment in model selection for practical applications.

1 Introduction and related work

The development of Large Language Models (LLMs) has marked a significant breakthrough in artificial intelligence, showing remarkable progress and versatility across various fields (Brown et al., 2020; Wei et al., 2022; Kojima et al., 2023; OpenAI, 2023). These advances have made substantial impacts in education, where LLMs are being actively adopted and tested in different learning contexts (Kasneci et al., 2023; Yan et al., 2024; Jeon and Lee, 2023). One of the notable applications in this domain is automated essay scoring (AES). AES represents a well-established research field with over fifty years of continuous development and improvement (Page, 1966; Hussein et al., 2019; Ke and Ng, 2019; Ramesh and Sanampudi, 2022). In recent years, fine-tuned deep neural networks, especially those based on BERT architectures, have shown superior performance in this task, setting new standards for automated assessment accuracy.

The application of LLMs in AES has gained significant attention from researchers worldwide (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Naismith et al., 2023; Kim and Jo, 2024; Yoshida, 2024; Lee et al., 2024; Tate et al., 2024). For instance, a study by Yancey et al. (2023) examined the performance of GPT-3.5 and GPT-4 in evaluating short essays, showing that GPT-4 could match human expert ratings with high accuracy even without specific training examples. Furthermore, Kim and Jo (2024) developed an innovative approach using GPT-4 for zero-shot comparative assessment, which yielded promising results in their experimental evaluation.

Meanwhile, recent advancements in LLMs include reasoning models such as OpenAI's o1 and o3, which have been enhanced through reinforcement learning and demonstrate superior performance across various benchmarks (OpenAI, 2024; OpenAI, 2025; DeepSeek-AI, 2024). However, while these models are expected to show improved capabilities in the context of AES, their actual performance in this specific domain remains unclear. Additionally, these new models no longer allow users to adjust parameters like temperature that were previously available in conventional LLMs to control output randomness. This raises questions about both the consistency of their outputs and their practical effectiveness in AES applications.

In this study, we investigate the essay evaluation capabilities of reasoning models, specifically OpenAI's o1-mini and o3-mini which are representative reasoning models. Specifically, we assess the consistency of these models by having them evaluate the same essays using identical

prompts 50 times. Furthermore, we evaluate their essay scoring performance using the TOEFL 11 dataset, which is widely used in AES research, by calculating the agreement rate between expert assessments and model-generated scores.

The key contributions of this study are twofold: (1) An investigation of the consistency of reasoning models (o1-mini and o3-mini) in essay evaluation through repetitive testing, providing insights into their reliability for AES applications. (2) A comprehensive evaluation of these models' scoring capabilities using the TOEFL 11 dataset, offering empirical evidence of their performance compared to human expert assessments.

2 Methods

2.1 Dataset

We used TOEFL11 (Blanchard et al., 2013) as the essay dataset, which was designed to support research in natural language processing. The dataset contains 12,100 English essays with expert ratings on a three-point scale (low, medium, and high). These ratings were initially evaluated by multiple experts using a 5-point rubric and subsequently compressed to a 3-point scale following a standardized set of rules. The original rubric ratings are not included in the dataset.

In our evaluation process, we first had AI models score essays on a five-point scale using rubric, then classified the scores following the original methodology: scores below 2.5 as low, between 2.5 and 3.5 as medium, and above 3.5 as high. For quantitative analysis, we converted the low, medium, and high to 1, 2, and 3, respectively.

2.2 Models

To evaluate the essay assessment capabilities of reasoning models, we employed OpenAI's o1-mini (o1-mini-2024-09-12) and o3-mini (o3-mini-2025-01-31). For comparison in Experiment 2, we also utilized GPT-4o mini (gpt-4o-mini-2024-07-18). We accessed these models through Microsoft Azure OpenAI Service API. The reasoning models were used with their default parameters, while for GPT-4o mini, we set the temperature to 0 and kept all other parameters at their default values.

2.3 Prompt

We developed prompts based on those used by Yancey et al. (2023). Figure 1 shows a template for a prompt. The prompts comprised several components: Instruction, Essay Prompt, Response, Rubric, and Output Format. The Instruction section described the essay evaluation task, while the Essay Prompt section presented the prompt for the essay. The Response section contained the essay to be evaluated. For the Rubric section, we employed the original rubric used in TOEFL. Lastly, the Output Format section included evaluation rationale and an output format specification.

2.4 Experiment 1: Evaluation of fluctuations in scoring of reasoning models

To examine the variability in essay evaluation of reasoning models, we selected 50 essays from each expert-rated category (low, medium, and high) and had each model evaluate these essays 50 times. We calculated the mean score and standard deviation of each essay within each category. Furthermore, to examine whether there were differences in the variance of standard deviations between categories, we conducted Fligner-Killeen tests between each category. We applied Holm correction due to multiple comparisons. This allowed us to assess the consistency of evaluations of reasoning models.

2.5 Experiment 2: Evaluation of the ability reasoning models on AES

To evaluate the AES capabilities of reasoning models, we obtained AI ratings for all essays and calculated their agreement with expert ratings. We used Quadratic Weighted Kappa (QWK), a widely adopted metric in AES evaluation (Ke and Ng, 2019; Ramnarain-Seetohul et al., 2022), as our measure of agreement. The 95% confidence intervals for QWK were calculated using the bootstrap method with 1,000 resampling iterations.

To test significant differences in QWK across models, we conducted paired bootstrap tests with 1,000 resampling iterations at a 5% significance level. The p-values were adjusted using Holm's correction to account for multiple comparisons.

You are a rater for writing responses on a high-stakes English language exam for second language learners. You will be provided with a prompt and the test-taker's response. Your rating should be based on the rubric below, following the specified format. There are rating samples of experts so that you can refer to those when rating.

Prompt
"""*Essay prompt*"""

Response
"""*Essay to be evaluated*"""

Rubric
Rubric

Output format:
Rationale: [<<<Your rationale here.>>>]
Rating: [<<<Your rating here.>>>]

Figure 1: A template for a prompt. The parts where data should be inserted are in *italics*.

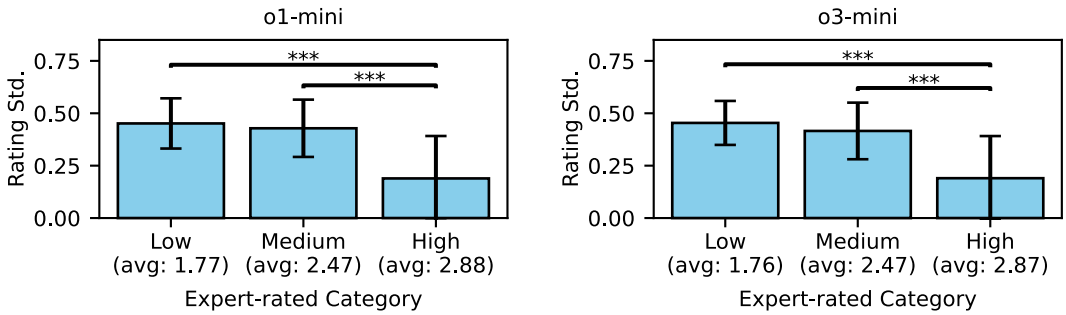

Figure 2: Standard deviations of repeated model ratings (50 times per essay) across 50 essays in each expert-rated category (low, medium, high). Error bars indicate standard deviations. Mean scores are shown in parentheses. Asterisks denote statistically significant differences between categories (***: p<0.001).

3 Results

3.1 Experiment 1: Evaluation of fluctuations in scoring of reasoning models

Figure 2 displays the standard deviations of repeated model ratings (50 times per essay) across 50 essays in each expert-rated category (low, medium, high). Both models exhibited similar average scores and standard deviations for each category. Regarding the average evaluation scores, both models tended to assign higher scores to essays rated as low or medium by experts, while assigning lower scores to essays rated as high. This tendency was particularly pronounced for essays rated as low.

With respect to standard deviations, none of the categories showed zero, indicating rating fluctuations even for the same essay. Furthermore, essays rated as high by experts demonstrated significantly lower standard deviations compared to those rated as low or medium. This suggests that essays rated as high by experts were evaluated more consistently, with less variation in the ratings by reasoning models.

3.2 Experiment 2: Evaluation of the ability of reasoning models on AES

Table 1 presents the evaluation results for all essays across three models. Surprisingly, o3-mini showed the lowest QWK, while GPT-4o mini achieved the highest QWK. Statistical analysis confirmed significant differences between these QWK scores.

Figure 3 illustrates confusion matrices between expert ratings and AI ratings. Both o1-mini and o3-mini demonstrated similar patterns, showing a general tendency to assign higher scores. This was particularly evident in their propensity to rate essays as high when experts had evaluated them as medium. In contrast, GPT-4o mini showed relatively higher agreement rates with expert evaluations across three models.

1. o1-mini	2. o3-mini	3. GPT-4o mini
$0.454^{2*,3***}$ [0.444-0.465]	$0.442^{1*,3***}$ [0.431-0.453]	$0.619^{2***,3***}$ [0.610-0.628]

Table 1: QWK values for all essays evaluated using each model. Asterisks next to model numbers indicate significant differences between corresponding models (*: p<0.05, ***: p<0.001). Values in parentheses represent 95% confidence intervals.

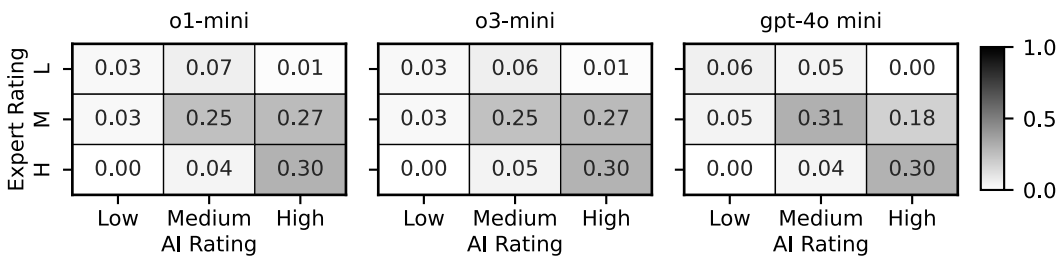

Figure 3: Confusion matrices between expert and AI ratings. H, M, and L indicate high, medium, and low.

4 Discussion

One of the most notable findings in our study was that the newer models, o1-mini and o3-mini, demonstrated lower performance compared to the conventional GPT-4o mini. While this result may appear counterintuitive, similar trends of conventional LLMs have been suggested in previous research (Yoshida, 2024). Specifically, Yoshida (2024) noted that newer models do not necessarily exhibit superior performance in AES tasks. However, to the best of our knowledge, this study is the first to show that a similar trend exists even in reasoning models. These findings provide important insights into how model evolution does not uniformly enhance performance across all tasks.

Another finding concerns the scoring consistency of reasoning models. Our analysis revealed a fundamental characteristic of these models: they exhibit inherent variability in their evaluations, even when repeatedly scoring the same essay. Furthermore, we found that this variability is not uniform across essay quality levels, with high-quality essays being evaluated more consistently than low or medium-quality ones. This systematic difference in scoring consistency was statistically significant, suggesting that the models' reasoning capabilities may be more stable when processing well-structured, high-quality content, possibly due to clearer patterns and more consistent features in such essays. The increased variability in scoring lower-rated essays could indicate that these models struggle to maintain consistent evaluation criteria when faced with more ambiguous or problematic writing. These findings are particularly important for educational applications, where scoring consistency is crucial for providing reliable feedback to learners.

These findings are related to significant disparities between benchmark tests and real-world tasks (Zhou et al., 2024; Banerjee et al., 2024). While recent language models have demonstrated remarkable improvements in standard benchmark tests (OpenAI, 2024; OpenAI, 2025; DeepSeek-AI, 2024), our results reveal that this superiority does not necessarily translate to tasks such as essay evaluation. This suggests that performance improvements in benchmarks might be the result of optimization for specific evaluation criteria. Conversely, our findings indicate that conventional models may sometimes perform better in complex, context-dependent tasks like essay evaluation.

These findings provide valuable insights for the practical implementation of language models in real-world applications. They particularly emphasize the importance of carefully evaluating whether the latest models are genuinely suitable for specific tasks rather than assuming their superiority. Furthermore, our results suggest that model selection should not be based solely on benchmark scores but should include task-specific evaluations. These insights can serve as crucial guidelines for the practical deployment of LLMs.

5 Conclusion

Our study provides several important insights into the application of reasoning models in AES. First, contrary to expectations based on benchmark performances, new reasoning models (o1-mini and o3-mini) demonstrated significantly lower performance compared to conventional models like GPT-4o mini in essay evaluation tasks. This finding challenges the assumption that newer models inherently perform better across all applications.

Second, we discovered that reasoning models exhibit inherent scoring variability, even when evaluating the same essay multiple times. This variability was particularly pronounced in low and medium-rated essays, while high-quality essays received more consistent evaluations. This fundamental characteristic suggests that before implementing these models in educational settings, we must first acknowledge and account for their inherent scoring instability, rather than assuming they will provide consistent evaluations.

These results have significant implications for both research and practical applications of AES systems. They suggest that: (1) model selection for AES should prioritize task-specific performance over general benchmark results, (2) implementation of AES systems should include robust evaluation of scoring consistency, and (3) further research is needed to understand why newer models may underperform in specialized tasks despite their superior benchmark performance.

Limitation

While our findings provide valuable insights into AES using reasoning models, several limitations should be acknowledged. First, although the TOEFL11 dataset is widely recognized in AES research, our experiments were limited to this single dataset. To enhance the generalizability of our findings, future studies should consider evaluating model performance across multiple established datasets, such as the Automated Student Assessment Prize (ASAP) dataset or The Cambridge Learner Corpus-First Certificate in English exam (CLC-FCE), which represent different writing contexts and assessment criteria.

Second, our analysis focused primarily on the overall scoring patterns and consistency but did not investigate which specific aspects of essay evaluation (e.g., coherence, grammatical accuracy, or argument development) contributed most to the observed variability in scoring. Understanding these detailed patterns could provide more nuanced insights into the strengths and weaknesses of reasoning models in AES tasks.

Finally, our study was limited to OpenAI's reasoning models. Given the rapid development of reasoning-enhanced language models, such as Google's Gemini 2.0 Flash Thinking and DeepSeek's DeepSeek R1, future research should examine whether our findings regarding scoring consistency and performance generalize across different reasoning models. This broader investigation would help establish whether the observed characteristics are specific to certain model architectures or represent common traits across reasoning-enhanced models.

Acknowledgements

In preparing this manuscript, we utilized Claude Pro and ChatGPT Pro for language refinement and the generation of example Python code, in accordance with AI Writing/Coding Assistance Policy.

References

Sourav Banerjee, Ayushi Agarwal, and Eishkaran Singh. 2024. The Vulnerability of Language Model Benchmarks: Do They Accurately Reflect True LLM Performance? *arXiv:2412.03597*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A Corpus of Non-Native English. *ETS Research Report Series*, 2013(2):i–15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf.

Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. Automated language essay scoring systems: a literature review. *PeerJ Computer Science*, 5:e208.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 28(12):15873–15892.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*. pages 6300–6308.

Seungju Kim and Meounggun Jo. 2024. Is GPT-4 Alone Sufficient for Automated Essay Scoring?: A Comparative Judgment Approach Based on Rater Cognition. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pages 315–319.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35:2199-22213.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the Potential of Using an AI Language Model for Automated Essay Scoring. *Research Methods in Applied Linguistics*, *2*(2):100050.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building*

Educational Applications (BEA 2023), pages 394–403.

OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774*.

OpenAI. 2024. OpenAI o1 System Card. *arXiv:2412.16720*.

OpenAI. 2025. OpenAI o3-mini System Card. https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

Ellis B. Page. 1966. The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan*, 47(5):238–243.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Vidasha Ramnarain-Seetohul, Vandana Bassoo, and Yasmine Rosunally. 2022. Similarity measures in automated essay scoring systems: A ten-year review. *Education and Information Technologies*, 27(4):5573–5604.

Tamara P. Tate, Jacob Steiss, Drew Bailey, Steve Graham, Youngsun Moon, Daniel Ritchie, Waverly Tseng, and Mark Warschauer. 2024. Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824-24837.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating Short L2 Essays on the CEFR Scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.

Lui Yoshida. 2024. The Impact of Example Selection in Few-Shot Prompting on Automated Essay Scoring Using GPT Models. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky (AIED 2024)*, pages 61–73.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68.