Are the Reasoning Models Good at Automated Essay Scoring?

Anonymous ACL submission

Abstract

This study investigates the validity and reliability of reasoning models, specifically OpenAI's o3-mini and o4-mini, in automated essay scoring (AES) tasks. We evaluated these models' performance on the TOEFL11 dataset by measuring agreement with expert ratings (validity) and consistency in repeated evaluations (reliability). Our findings reveal two key results: (1) the validity of reasoning models o3-mini and o4-mini is significantly lower than that of a non-reasoning model GPT-4o mini, and (2) the reliability of reasoning models cannot be considered high, with Intraclass Correlation Coefficients (ICC) of approximately 0.7 compared to GPT-4o mini's 0.95. These results demonstrate that reasoning models, despite their excellent performance on many benchmarks, do not necessarily perform well on specific tasks such as AES. Additionally, we found that few-shot prompting significantly improves performance for reasoning models, while Chain of Thought (CoT) has less impact.

1 Introduction

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; Pack et al., 2024; Kim and Jo, 2024; Yoshida, 2024; Lee et al., 2024; Tate et al., 2024). For instance, a study by Pack et al., (2024) evaluates the validity and reliability of LLMs for AES in language education, finding GPT-4 exhibited the best performance with excellent intra-rater reliability and good validity.

Meanwhile, recent advancements in LLMs include reasoning models, which have been enhanced through reinforcement learning and demonstrate superior performance across various benchmarks (OpenAI, 2024; OpenAI, 2025a; OpenAI, 2025b; DeepSeek-AI, 2024). However, while these models are expected to show improved capabilities in AES, their actual performance in this specific domain remains unclear.

We investigate validity and reliability in AES for reasoning models, specifically OpenAI's o3-mini and o4-mini. Through evaluation using six prompting strategies combining zero-shot and few-shot prompting with Chain of Thought (CoT), we ensure robust findings. For validity, we measure agreement between expert and model ratings on the TOEFL11 dataset. For reliability, we evaluate intra-rater consistency through repeated evaluations. Our findings demonstrate two key results: (1) reasoning models o3-mini and o4-mini show significantly lower validity than non-reasoning model GPT-4o mini, and (2) reasoning models exhibit moderate rather than excellent reliability. These results reveal that reasoning models, despite excellent performance on many benchmarks, do not necessarily perform well on different tasks such as AES.

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 the 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 o3-mini (o3-mini-2025-01-31) and o4-mini (o4-mini-2025-04-16), with GPT-4o mini (gpt-4o-mini-2024-07-18) serving as our reference model. We focused on these models due to cost constraints and for rapid preliminary analysis. We accessed these models through Microsoft Azure OpenAI Service API. The reasoning models used their default parameters, while GPT-4o mini had the temperature set to 0 with other parameters at default values.

2.3 Prompt

We developed several types of prompts based on previous research (Yancey et al., 2023; Naismith et al., 2023; Yoshida, 2024). To evaluate the influence of representative prompt engineering techniques such as CoT and few-shot prompting, we created three prompt types with varying CoT degrees: S (score only, no CoT), SR (score before rationale), and RS (score after rationale). For each type, we prepared both zero-shot and few-shot versions, creating a total of six prompts (Sz, SRz, RSz, Sf, SRf, RSf). Since few-shot examples can influence scoring ability (Yoshida, 2024), to reduce bias, we randomly selected three expert evaluations for each essay as few-shot examples. This design assesses both CoT's isolated effect (comparing S with SR/RS) and CoT-few-shot interaction across models.

The prompts comprised several components: Instruction, Essay Prompt, Response, Rubric, Expert Examples (in the case of few-shot), and Output Format. For the Rubric section, we employed the original rubric used in TOEFL. Lastly, in the Output Format section, for S we output only the Rating, for SR we output the Rating followed by the Rationale, and for RS we output the Rationale followed by the Rating. Figure 1 shows an example template for a prompt of RSf.

2.4 Validity Evaluation

To evaluate the validity of reasoning models in AES, 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. To test significant differences in QWK across models and prompts, 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.

2.5 Reliability Evaluation

We tested the intra-rater reliability of AES reasoning models by having each model evaluate 900 essays (300 from low, medium, high expert-rated categories) 20 times using best-performing prompts. We calculated Intraclass Correlation Coefficients (ICC) (3,1) values using a two-factor mixed effects model, with 95% confidence intervals via bootstrap resampling (1,000 samples). To test significant differences between ICCs, we employed nonparametric bootstrap testing with 1,000 samples based on distribution differences at a 5% significance level, applying Holm's method for multiple comparison correction.

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

Rating samples of experts:
Samples

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

Figure 1: An example template for a prompt of RSf. Data should be inserted in *italics*.

	zero-shot			few-shot		
LLM Model	**Sz**	**SRz**	**RSz**	**Sf**	**SRf**	**RSf**
o3-mini	0.440	0.460	0.442	0.539	0.542	0.542
o4-mini	0.447	0.457	0.445	0.536	0.537	0.539
GPT-4o mini	**0.628**	**0.618**	**0.621**	**0.595**	**0.575**	**0.555**

Table 1: QWK between expert and AI ratings for all essays evaluated using each model and prompt. Bold and underlined numbers indicate the highest QWK across models and prompts respectively.

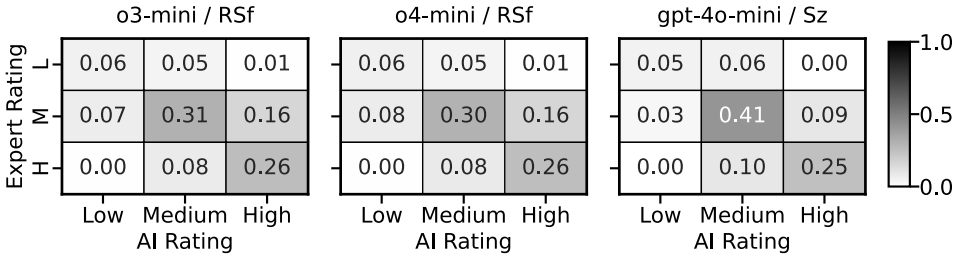

Figure 2: Confusion matrices between expert and AI ratings, featuring the prompt-model combinations that achieved the highest QWK for each model. H, M, and L indicate high, medium, and low. Each cell shows the proportion of cases where the expert rating (vertical axis) aligns with the AI rating (horizontal axis).

3 Results

3.1 Validity Evaluation

Table 1 shows QWK results for all essays across models and prompts. All test results are detailed in the Appendix, and we discuss QWK results in line with the main test findings. Surprisingly, GPT-4o mini demonstrated high QWK across all prompts between models. GPT-4o mini / Sz showed the highest QWK, which was significantly higher than QWK using other models and prompts except for GPT-4o mini / RSz.

Regarding few-shot, for o3-mini and o4-mini, the QWK of few-shot was significantly higher than that of zero-shot in all combinations within the same model. Additionally, there were no significant differences between any combinations of few-shot prompts within the same model. For GPT-4o mini, the QWK of zero-shot was significantly higher than that of few-shot in all combinations. Among zero-shot prompt combinations within the same model, the QWK of SRz was significantly lower than the QWK of Sz and RSz.

Figure 2 illustrates confusion matrices between expert ratings and AI ratings, which shows the combination of model and prompt with the highest QWK. For all three models, while the discrimination rates for low and high essays were not substantially different, for medium essays, both o3-mini and o4-mini demonstrated similar patterns, showing a general tendency to assign high ratings. In contrast, GPT-4o mini showed relatively higher agreement rates with expert evaluations.

3.2 Reliability Evaluation

Figure 3 shows the ICC results. While values of o3-mini and o4-mini were approximately 0.7, one of GPT-4o mini was approximately 0.95. For ICC value interpretation: values below 0.5 indicate poor reliability, 0.5-0.75 moderate reliability, 0.75-0.9 good reliability, and values above 0.9 excellent reliability (Koo and Li, 2016). Based on this criteria, o3-mini and o4-mini showed moderate reliability, while GPT-4o mini demonstrated excellent reliability.

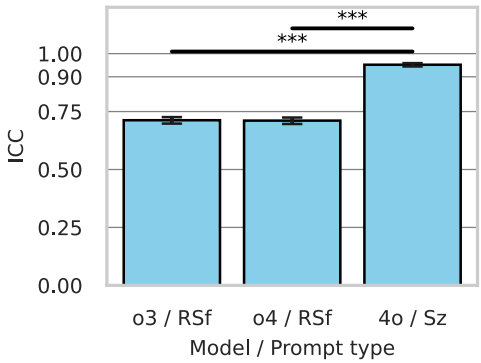

Figure 3: ICC in the model and prompt combination that had the highest QWK. Bars indicate 95% confidence intervals (***: p<0.001). o3-mini, o4-mini, and GPT-4o mini are abbreviated as o3, o4, and 4o, respectively.

4 Discussion

Our study found that newer models o3-mini and o4-mini performed significantly worse than GPT-4o mini. This underperformance persisted across all six prompting strategies tested, suggesting this stems from fundamental model characteristics rather than prompting techniques. The low QWK was caused by reasoning models rating expert-rated medium essays as high. Detailed analysis of reasoning processes in correctly versus incorrectly evaluated essays may help us understand these patterns.

In our experiments, the highest QWK of 0.628 achieved by GPT-4o mini falls short of the state-of-the-art QWK of 0.782 achieved by Cho et al. (2024), who introduced the Dual-scale BERT + CNN model for multi-trait AES. Although our prompts did not reach state-of-the-art performance with LLM-based AES, considerable potential remains for further improvement through fine-tuning, advanced prompt-engineering techniques, or ensemble methods.

For reasoning models, few-shot performance was significantly superior to zero-shot performance. While earlier research (DeepSeek-AI, 2024) suggests that few-shot approaches reduce performance in reasoning models, our results contradict this finding. We found that providing examples improves performance even in reasoning models, suggesting genuine potential for prompt engineering.

There were no significant differences in few-shot prompts regarding reasoning provision or placement (Sf, SRf, RSf), showing CoT has less impact than few-shot prompting. Since reasoning models perform reasoning internally, they could already use a form of CoT, explaining the lack of differences.

Reliability evaluation revealed reasoning models achieved moderate consistency (ICC≈0.7), considerably lower than GPT-4o mini's excellent reliability (ICC≈0.95). This 0.25-point ICC gap has important implications: reasoning models require 56% more evaluations to achieve equivalent reliability in high-stakes assessments. Score variability likely originates from reasoning models generating different reasoning paths for the same essay, causing fluctuations that do not occur in more deterministic non-reasoning models. Low reliability and limited validity reinforce each other: unstable scores naturally lead to worse agreement with human raters. This suggests unstable scoring in reasoning models contributes to lower validity scores, creating practical challenges for AES applications.

Our experiments found that reasoning models o3-mini and o4-mini, despite strong benchmark performance (OpenAI, 2025a; 2025b), showed lower validity and reliability in AES than GPT-4o mini. This demonstrates that benchmark performance doesn't guarantee effectiveness across all tasks. Previous research (Yoshida, 2024) made similar observations, and our results confirm this pattern applies to reasoning versus non-reasoning models. Therefore, evaluating each model on specific tasks is essential rather than assuming universal effectiveness. Additionally, reasoning models' improved performance with few-shot approaches shows potential for prompt engineering, representing one contribution of this research.

5 Conclusion

In this study, we investigated the validity and reliability of reasoning models (o3-mini and o4-mini) in AES tasks using the TOEFL11 dataset. Through comprehensive evaluation using QWK for validity and ICC for reliability, we found that reasoning models demonstrated lower performance compared to the non-reasoning model GPT-4o mini despite their superior performance on general benchmarks.

Our key findings revealed that (1) reasoning models achieved QWK values of 0.539-0.542, significantly lower than GPT-4o mini's 0.628, and (2) reasoning models showed moderate reliability (ICC ≈ 0.7) compared to GPT-4o mini's excellent reliability (ICC ≈ 0.95). Additionally, we discovered that few-shot prompting significantly improved reasoning model performance, while CoT had minimal impact.

These findings contribute to our understanding that state-of-the-art models may not universally excel across all tasks, highlighting the critical importance of task-specific evaluation rather than relying solely on general benchmark performance. This research provides essential guidance for practitioners in selecting appropriate models for AES and underscores the need for continued investigation into model performance across diverse applications.

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. Each dataset in AES research has unique characteristics including varying essay prompts, rubrics, and target student populations, requiring careful adaptation of prompting strategies and experimental methodologies. Given the scope and complexity of conducting rigorous multi-dataset evaluations, we focused our initial investigation on TOEFL11 to establish a detailed methodological framework that can be systematically applied to other datasets in future studies. 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, which represents different writing contexts and assessment criteria.

Second, while our study provides comprehensive evaluation through six different prompting strategies and reliability analysis, our analysis focused primarily on overall scoring patterns and consistency without investigating which specific aspects of essay evaluation (e.g., coherence, grammatical accuracy, or argument development) contributed most to the observed variability in reasoning model performance. Understanding these detailed evaluation patterns could provide more nuanced insights into why reasoning models tend to rate expert-evaluated "medium" essays as "high" (Figure 2) and why their evaluations show greater variability (ICC ≈ 0.7) compared to GPT-4o mini. This limitation represents an important direction for future research in understanding the fundamental differences between reasoning and non-reasoning models in AES tasks.

Finally, our analysis was limited to OpenAI's models despite the availability of alternative LLMs and pre-trained language models (PLMs) such as BERT-based systems. This decision was driven by cost constraints and the need for rapid preliminary analysis, leading us to focus on representative OpenAI models as benchmark examples. While this allowed for in-depth evaluation of reasoning versus non-reasoning approaches, future research should expand comparisons to include other major providers (e.g., Google's Gemini, Anthropic's Claude, DeepSeek's models) and established fine-tuned PLMs to provide a more comprehensive landscape view. Given the rapid development in this field, continuous evaluation with emerging reasoning models will be essential to determine whether our findings reflect universal characteristics of reasoning approaches or specific architectural traits of the models tested.

Acknowledgments

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

Minsoo Cho, Jin-Xia Huang, and Oh-Woog Kwon. 2024. Dual‐scale BERT using multi‐trait representations for holistic and trait‐specific essay grading. *ETRI Journal*, 46(1):82-95.

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.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155-163.

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

OpenAI. 2025b. OpenAI o3 and o4-mini System Card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

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.

Appendix

Table 2 shows the p-values for QWK between all combinations of models and prompts.

	o3/Sz	o3/SRz	o3/RSz	o3/Sf	o3/SRf	o3/RSf	o4/Sz	o4/SRz	o4/RSz	o4/Sf	o4/SRf	o4/RSf	4o/Sz	4o/SRz	4o/RSz	4o/Sf	4o/SRf	4o/RSf
o3/Sz	-	0.000	1.000	0.000	0.000	0.000	1.000	0.068	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o3/SRz	0.000	-	0.068	0.000	0.000	0.000	0.280	1.000	0.120	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o3/RSz	1.000	0.068	-	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o3/Sf	0.000	0.000	0.000	-	1.000	1.000	0.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.068
o3/SRf	0.000	0.000	0.000	1.000	-	1.000	0.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.324
o3/RSf	0.000	0.000	0.000	1.000	1.000	-	0.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.232
o4/Sz	1.000	0.280	1.000	0.000	0.000	0.000	-	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o4/SRz	0.068	1.000	0.000	0.000	0.000	0.000	1.000	-	0.364	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o4/RSz	1.000	0.120	1.000	0.000	0.000	0.000	1.000	0.364	-	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
o4/Sf	0.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	-	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
o4/SRf	0.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	1.000	-	1.000	0.000	0.000	0.000	0.000	0.000	0.000
o4/RSf	0.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	1.000	1.000	-	0.000	0.000	0.000	0.000	0.000	0.068
4o/Sz	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-	0.000	1.000	0.000	0.000	0.000
4o/SRz	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-	1.000	0.000	0.000	0.000
4o/RSz	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	-	0.000	0.000	0.000
4o/Sf	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-	0.000	0.000
4o/SRf	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-	0.000
4o/RSf	0.000	0.000	0.000	0.068	0.324	0.232	0.000	0.000	0.000	0.000	0.000	0.068	0.000	0.000	0.000	0.000	0.000	-

Table 2: P-values in the significance test results for QWK across all model and prompt combinations. The areas shaded in gray indicate combinations that were significant at the 5% level. o3-mini, o4-mini, and GPT-4o mini are abbreviated as o3, o4, and 4o respectively.