A PILOT STUDY OF RLM-ASSISTED GRADING FOR AI-HUMAN CO-EVALUATION IN ASSESSMENT

Sean TSOI* [a] and Sin Kei CHEUNG [b]

[a] Centre for Learning and Teaching, Vocational Training Council, Hong Kong
[b] Centre for Learning and Teaching, Vocational Training Council, Hong Kong

Sean TSOI* (mrseantsoi@vtc.edu.hk)

Reasoning Language Models (RLMs) have gained growing significance in the Generative AI (GenAI) market owing to its "thinking" ability in addition to content generation. While research have showed Large Language Models' (LLMs) satisfactory performance in essay grading with zero-shot prompting, little research has been conducted on the grading capability of RLMs. It is intriguing to examine whether RLMs would further improve the grading ability of GenAI. This study compares the marking and grading of essays by human and RLM markers to explore and recommend feasible and effective AI-human co-evaluation approaches with RLM assistance in assessments.

Four RLMs, namely DeepSeek r1 (Guo et al., 2025), Grok 3 (xAI 2025), OpenAI o3-mini (OpenAI 2025) and Claude 3.7 Sonnet (Anthropic 2025), were selected to generate 20 essays. Human and RLM markers assigned grades according to a grade descriptor. Apart from grading, the human markers also conducted grade review, followed by verification by an independent verifier. The grades were analysed with Krippendorff's Alpha to evaluate the inter-marker reliability and validity. The experimental results demonstrated the grading ability and reliability of the RLMs studied and shed light on the possible roles and values of RLM assistance in marking and grading.

This study concludes that DeepSeek r1 maintains the highest stability in grading among the RLMs studied. To promote effective AI-human co-evaluation in assessments, the development of AI agents for grading, the adaptation of RLM-assisted marking and the deployment of RLMs for grade moderation are recommended.

Keywords: *AI-assisted grading, AI-human co-evaluation, automated essay grading, reasoning language model, inter-marker agreement*

Introduction

Grading essays is a time-consuming task which requires intensive marking of grammatical errors, cohesion and rhetorical structures. These challenges highlight the need for reliable automated essay scoring/grading (AES/AEG) systems. While automated grading is more efficient than human grading, its validity is subject to the quality of machine training and the assessment format, which has an impact on the legitimacy of grades as well as academic fairness for students. Particularly for writing topics that require critical judgement and decision-making, "hallucinated" grades that misrepresent the actual performance of students in assessment would undermine the legitimacy of automated grading.

In this regard, Reasoning Language Models (RLMs) have gained growing significance in the Generative AI (GenAI) market owing to its "thinking" ability in addition to content generation. Comparing to earlier studies about ChatGPT's grading ability and other AES's rubric-aware chain-of-thought (CoT) prompting (Flodén 2025; Xu et al., 2025), RLMs were claimed to possess the ability of reasoning. RLMs breakdown a complex problem and solve it step by step, allowing few-shot learning on tasks for which the model is not specifically trained (Aske et al., 2024). In the context of essay grading, this reasoning ability may enhance RLMs' understanding of a grade descriptor and facilitate grading and feedback.

In this study, the research team aims to compare the grading ability between humans and RLMs by evaluating the validity and inter-marker reliability in a co-evaluation experiment and explore feasible and efficient AI-human co-evaluation approaches including automated grading with AI agents, RLM-assisted marking and RLM grade moderation.

Literature Review

AES is not a novel topic in learning and teaching. AES systems have been developed with machine learning, natural language processing and large language models. There have been discussion and review regarding the process, validity and impacts of AES (Burrows et al., 2015; Conijn et al., 2023; Zupanc & Bosnić, 2015).

Concerning the grading quality of the AES systems, for example, Gabon et al. (2025) compared the performance of their NLP-based AES system against five essay genres. Flodén (2025) evaluated ChatGPT 3.5 against human graders for grading essays with zero-shot prompting, where no rubrics or grading samples were provided. They demonstrated LLMs' overall capability of essay grading with certain challenges about writing genres and the length of answers.

Regarding a comparation of existing LLMs, Kundu (2024) employed several prompting approaches with ChatGPT-3.5-Turbo and Llama 3 for grading essays and compared their grades with human graders'. They reported that the LLMs generally assigned lower scores than human markers and one LLM was comparatively more lenient than the other. In their conclusion, they believed LLMs could not sufficiently replace humans in grading but demonstrated potential in assistance of grading.

Seßler et al. (2025) evaluated open-source and closed-source LLMs for assessing student essays. They observed that closed-source models, especially o1, showed higher reliability and stronger correlations with human assessments compared to open-source models. Wei et al. (2022) claimed the o1 series had a larger parameter set and CoT reasoning capability which helped to better reflect human evaluative practices, particularly in content-related categories.

With regard to co-evaluation, Xiao et al. (2025) conducted human-AI co-grading experiments that resulted in an improvement of the performance and efficiency of different levels of graders. Particularly they showed that human-AI co-evaluation had helped novice graders reach similar accuracy levels to experienced graders.

Methodology

To achieve the objectives of this study, a co-evaluation experiment was carried out from March to July 2025. The experiment was divided into several stages, including RLM generation of sample essays, RLM marking and grading, initial grading by human markers (HMs), moderation of grade differences, grade review by HMs and grade verification by verifier.

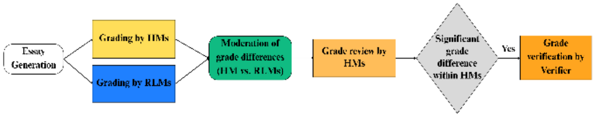

Essay Generation
Grading by HMs
Grading by RLMs
Moderation of grade differences (HM vs. RLMs)
Grade review by HMs
Significant grade difference within HMs
Yes
Grade verification by Verifier

Figure 1. Procedures of experiment

An essay writing question about the pros and cons of workplace monitoring was adapted from the 2023 Hong Kong Diploma of Secondary Education examination (HKDSE) as the essay topic for the experiment. Four RLMs, namely DeepSeek r1, Grok 3, OpenAI o3-mini and Claude 3.7 Sonnet, were employed. Five English and Communication teachers were recruited as markers and the verifier.

A grade descriptor was adapted from an English writing module of the Hong Kong Institute of Vocational Education. In this co-evaluation experiment, markers were required to assign each essay one grade of A, B, C, D or F, where D was passing and F failing. There were no sub-grades such as B+ or C- nor zero marks.

1. **RLM Generation of Sample Essays:** Each RLM received meticulously designed prompts to tackle the essay writing question and generate five essay responses. A total of twenty essays were collected. For the purpose of experiment, two sets of essays contained spelling mistakes in keywords or irrelevant content.
2. **RLM Marking and Grading:** The essay collection was graded by RLMs according to the rubrics. To gauge the impact of the number of essays on grading, two marking procedures were adopted. In Round One, a new chat window was opened for grading every essay (hereby referred to as 1 essay/prompt grading). In Round Two, twenty essays were graded simultaneously with one prompt in one chat window (hereby referred to as 20 essays/prompt grading). Grades and comments were returned.
3. **Initial Grading by Human Markers:** The HMs were instructed to conduct marking and grading according to the rubrics without the help of any others or any GenAI tools. The RLM authorship was disclosed but not identified. Four marking files with marking, grades and remarks were collected.
4. **Moderation of Grade Differences:** All the grades assigned by RLMs and HMs were compared. Three levels of grade differences were categorised, namely Major Grade Difference where only one or no RLM had assigned the same grade with an HM, Minor Grade Difference where two RLMs had assigned the same grade with an HM and Mild or No Grade Difference where three or all the RLMs assigned the same grade with an HM.
5. **Grade Review by Human Markers:** The grade results and remarks of the RLMs as well as the grade differences were shared with the HMs. The RLMs were also identified. The HMs reviewed all their grades, particularly those with minor and major grade differences. Reviewed grades and additional remarks were collected.
6. **Grade Verification by Verifier:** Subsequent to grade review, an independent experienced marker was invited to be the verifier for verifying the reviewed grades that contained major (three or four different grades, e.g. ABBD/BCDF) and minor (two dominant grades, e.g. AABB/ CCFF) grade differences among the HMs. Access to the grades and marking of both HMs and RLMs were provided to assist in verification. Upon completion of verification, an interview was conducted with the verifier concerning the following questions:
   Q1: How relevant and valid are the marking and grades of RLMs?
   Q2: What roles can RLMs perform in the co-evaluation process?
   Q3: Are there any other observations of the whole RLM-assisted verification process?

To measure the reliability of RLM markers, Krippendorff's Alpha (K-Alpha) was calculated for the inter-marker reliability within and between the two groups of markers (excluding the verifier) using an open-source Python program (Santiago Castro, 2017). As illustrated in Table 1, K-Alpha indicates the level of inter-marker agreement.

Table 1. Overview of the Krippendorff's Alpha values (Marzi et al. 2024, p. 3)

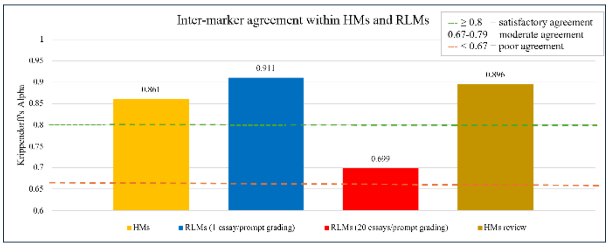
K-Alpha	Interpretation
1	Indicates *perfect agreement* among raters. It is the scenario where all raters have provided the exact same ratings for each item evaluated.
≥ 0.80	This value is generally considered a *satisfactory level of agreement*, indicating a reliable rating. In many research contexts, a Krippendorff's Alpha equal to or above 0.80 is acceptable for drawing triangulated conclusions based on the rated data.
0.67-0.79	This range is often considered the lower bound for tentative conclusions. A Krippendorff's Alpha in this range suggests *moderate agreement*; thus, outcomes should be interpreted with concern, questioning the roots of such diverging ratings.
< 0.67	This is indicative of *poor agreement* among raters. Data with a Krippendorff's Alpha below this threshold are often deemed unreliable for drawing triangulated conclusions. It suggests that the raters are not applying the coding scheme consistently or that the scheme itself may be flawed.
0	Indicates *no agreement* among raters other than what would be expected by chance. It is similar to a random rating pattern.
< 0	A negative value of Krippendorff's Alpha indicates *systematic disagreement* among raters. This situation might arise in cases where raters are systematically inclined in opposite rating directions.

Since the grades of 1 essay/prompt grading obtained the higher K-Alpha (0.874), the grade set was selected as reference for conducting grade review. For this reason, unless otherwise specified, the analysis of grades assigned by RLMs refers to the grades of 1 essay/prompt grading.

Results and Discussion

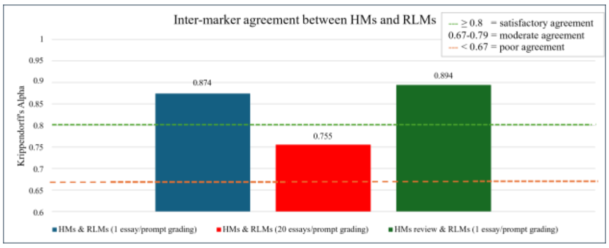
1. Inter-marker Agreement on Grades

In 1 essay/prompt grading, K-Alpha between the RLMs was 0.911, suggesting strong agreement of grades within the RLMs (Graph 1). However, in 20 essays/prompt grading, K-Alpha between the RLMs dropped to 0.699, implying only moderate agreement, significantly weaker. OpenAI o3-mini likely attributed to the decrease since it only obtained moderate agreement with HMs' (K-Alpha = 0.689), while all the other RLMs surpassed 0.8 (Graph 3).



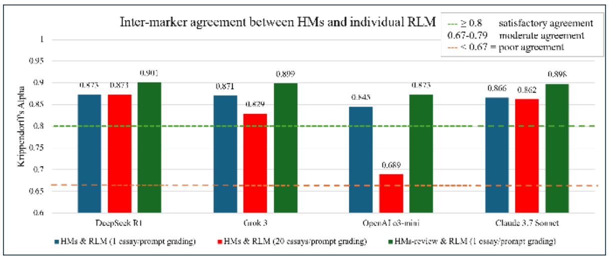
Graph 1. Inter-marker agreement within HMs and RLMs

The grades assigned by RLMs were generally comparable with HMs', in other words, highly reliable. For 1 essay/prompt grading, K-Alpha between HMs and RLMs were 0.874, suggesting satisfactory agreement between the two marker groups (Graph 2). If the grades assigned by HMs were assumed to be valid, the satisfactory agreement had affirmed the grading validity of RLMs. It is reasonable to believe, as suggested in studies conducted with LLMs (Gabon et al., 2025; Flodén, 2025), RLMs are capable of grading essays with satisfactory alignment with marking requirements.



Graph 2. Inter-marker agreement between HMs and RLMs

Among the RLMs, DeepSeek r1 outperformed the other three RLMs by a tiny K-Alpha margin between 0.002 and 0.007 (Graph 3). In fact, DeepSeek r1 was the only RLM that maintained the same K-Alpha with HMs in both 1 and 20 essays/prompt grading, demonstrating a highly consistent grading stability.



Graph 3. Inter-marker agreement between HMs and individual RLM

2. Identification of Failing Basic Requirements

The ability to identify content that fails to meet the basic requirement is also important since the percentage of failure indicates the performance and validity of an assessment. For this reason, two "tricky" essay types were created. Essays E3 featured repetitive paragraphs, and essays E4 contained grammatically correct but irrelevant content, both simulating scripts that failed the basic requirements.

Grading Essays with Meaningless Repetitions (essays E3)

Each essay E3 consisted of a relevant heading followed by five identical paragraphs. Although such repetition is unlikely to be present in a synchronous assessment setting, the research team simulated this scenario to assess whether RLMs could identify the repetition and assign a failure grade appropriately. This is critical since students might exploit this loophole in assignments or take-home examinations to achieve high grades with minimal efforts.

Essays with repetitive paragraphs should unarguably receive a failing grade. As assumed, all the HMs assigned F to all essays E3. In both 1 and 20 essays/prompt grading, both DeepSeek r1 and Claude 3.7 Sonnet graded essays E3 with F consistently. However Grok 3 graded D to one essay E3 (essay no.15) despite the acknowledgment that it "reus[es] the same paragraph five times—severely limits its organization, development, and cohesion, preventing it from meeting the 500-word requirement with meaningful content." This highlighted Grok 3's awareness of repetitive content but also underscored its misalignment between observations (marking) and judgement (grading). Accuracy in error pinpoints did not guarantee corresponding grading decisions. More specific grading instructions on certain foreseeable essay performance via pre-prompting might improve the result.

Grading Essays with Irrelevant Content (essays E4)

The markers were also experimented on grading apparently irrelevant essays. Essays E4 were designed to elaborate on relevant themes such as "computer software and workplace" and "cameras and productivity" and avoid references related to surveillance, monitoring and privacy. This simulation replicated circumstances where students with commendable writing skills demonstrate partial or insufficient comprehension to the question task. Their writing might read logically coherent and linguistically proficient yet discursively superficial or substantially deviate from the core of the question, which result in a deficiency that should be reflected in grading.

The average grades assigned to essays E4 by HMs and RLMs showed satisfactory agreement. Two HMs assigned F to all essays E4 while DeepSeek r1 and Grok 3 assigned F to 75% of them. Notably RLMs assigned C or above[1] to 25% of essays E4, 15% higher than HMs. In 20 essays/prompt grading, the passing percentage for essays E4 soared as the number of F grade dropped from 9 to 4, which all came from DeepSeek r1. This contrast challenges the grading reliability of grading multiple essays per prompt for RLMs.

In fact, remarks given by the RLMs including "limited discussion" of the essential components, "partial fulfilment of the task" and even "misalignment with the core task requirements" demonstrated their awareness of absent yet essential information in essays E4. Some other remarks directly pointed out "off-topic" content. However, Grok 3 mentioned a critical knowledge gap—"the rubrics...do not specify *how to weigh or combine the grades*" (remarks for essay E4 no.3). The RLMs might have relied on their own logical reasoning to weigh the grades of different performance criteria, resulting in assigning passing grades to perhaps eloquent but off-topic essays.

To align RLMs' observations with their subsequent grading judgements, more training about the use of grade descriptors should be arranged. Prompts should also be refined with more specific mentions such as "irrelevant content" and "deviations from the theme" to assist RLMs in categorising related performance into failure.

3. Grade Verification and Verifier's Observations

The verified grades were not analysed for agreement between HMs and RLMs due to partial representation. However, it is obviously that the verified grades further improved the inter-marker agreement between HMs and RLMs (Table 2).

Table 2. Verification results of reviewed grades with significant differences among HMs

Essay	HM1 rev	HM2 rev	HM3 rev	HM4 rev	Deep Seek r1	Grok 3	OpenAI o3-mini	Claude 3.7 Sonnet	Verifier
E4 no.2	C	F	F	D	B	C	B	C	**C**
E4 no.3	D	F	F	D	F	F	D	F	**F**
E4 no.4	D	F	F	D	F	F	D	C	**F**
E2 no.10	B	B	C	A	C	B	B	B	**B**
E4 no.18	C	F	F	D	F	F	F	F	**F**

The verifier stated that when a student fails to meet the basic requirements in *any* criterion in the grading descriptor, no passing grade should be assigned. Yet, he assigned F to all essays E4 except essay no.2, citing "relevant ideas [were] shown". It is difficult to determine whether the grades and remarks of RLMs had directly contributed to the verified pass for essay no.2 as its content might be minimally relevant to the writing question, which the research team does not intend to discuss. Nevertheless, if RLMs were employed for automated grade moderation, the verification of their grading accuracy and validity would be crucial.

Apart from reading the essay scripts, the verifier noted that he had compared his own marking with the remarks given by HMs and RLMs. He commented that the remarks by RLMs were useful in informing the overall performance of an essay. For instance, RLMs could not only identify grammar mistakes and language structures but also assess an essay's relevance to the topic, which enabled a convenient comprehensive evaluation. He thus believed RLMs' analysis of the content, structure and errors of an essay could accelerate human's marking and grading. In short, RLM-assisted marking is beneficial for enhancing the efficiency of human marking.

The RLMs' ignorance of the grading mechanism behind the grade descriptors was also observed. Although

[1] Essay no.2 was graded B by DeepSeek r1 and OpenAI o3-mini, and C by Grok 3 and Claude 3.7 Sonnet; essay no.4 was graded C by Claude 3.7 Sonnet.

the RLMs were given a grade descriptor to differentiate the level of performance for each grade, the RLMs were left independent in interpreting the grading approach. The RLMs failed to execute a holistic evaluation approach. For over 70% of the essays, for instance, the RLMs assigned grades to individual grade criterion outlined in the grade descriptor, an unnecessary practice which posted a contrast to the remarks of the HMs. This finding is coherent with the grade weighting issue discussed earlier for grading irrelevant content.

Recommendations

1. Developing AI Agents for Co-evaluation

The development of an AI agent for grading can facilitate alignment of grading between HMs and RLMs. Developing an AI agent allows an educator to train a marking and grading-specific RLM more comprehensively by establishing a customised knowledge base and finetuning the RLM with co-evaluation specifications. In addition to the provision of rubrics, an educator can introduce the marking procedures and explain the rubrics and expectations explicitly in pre-prompting. Grading samples can also be uploaded to familiarise the RLM with the grading mechanism and approaches. As a result, the marking approaches, grade weighting methods and other important considerations for grading can be communicated to optimise RLM assistance.

Furthermore in writing assessments that require students to extract information from a data file, the educator can consider limiting or terminating the use of pre-trained data of RLM so that the writing content will be evaluated entirely according to the assessment materials and criteria uploaded in the knowledge base to avoid inaccurate judgement arising from RLM's general knowledge. Rigorous testing and verification are nevertheless warranted to ensure validity and stability.

2. Adopting RLM-assisted Marking

It is recommended that HMs make close reference to the marking of RLMs to enhance the effectiveness and efficiency of grading. Apart from grading, RLMs can point out and categorise grammar mistakes, cohesive devices and rhetorical structures quickly, accurately and systematically. RLMs can also analyse the diction, register and cohesion to evaluate the fluency of an essay. By reviewing RLM s' marking, HMs can identify both the strengths and weaknesses of an essay in shorter time. It not only reduces the chance of oversight and human marking errors but also facilitates HMs' devotion of time and efforts for grading. As a result, the benefits of the assessed can be upheld.

3. Deploying RLMs for Grade Moderation

In human marking without AI's assistance, RLMs can assume the responsibility of a grade moderator. RLM grades can be compared with the HMs' to find out essays with significant grade differences. Enquiries should be made with the RLM to ascertain the reasons behind the grade difference. For a more in-depth investigation, if possible, the marking of the HM should also be shared with the RLM.

As demonstrated in the co-evaluation experiment, an experienced teacher can review the grade difference as well as the comparison of HM and RLM marking to determine the final grade.

The proposed moderation procedure can upgrade the validity of grades especially for existing evaluation procedures with only one HM or multiple novice HMs. It addresses the limitations posed by time and human resources which are necessary for review and verification in an assessment. In other words, as a grade moderator, RLM can facilitate review and assure validity of an assessment.

Conclusions

This study has explored the potential of RLMs in AES for AI-human co-evaluation. DeepSeek r1 has been shown as the most reliable and stable RLM with consistently high inter-marker agreement across different experimented grading approaches. While RLMs have exhibited strong grading reliability, further development and training are still necessary to improve the grading consistency and stability. Therefore this study recommends three strategies, a) developing AI agents for co-evaluation, b) adopting RLM-assisted marking and c) deploying RLMs for grade moderation, for implementing AI-human co-evaluation in assessment.

Some limitations have been identified in this pilot study. Although this study aimed to evaluate zero-shot RLM grading, the absence of sample graded scripts may have placed RLMs at a disadvantage compared to HMs with experience in teaching and assessment.

Furthermore, the significance of this study is also limited by the availability of authentic essay samples. RLM-generated essays are different from real-time student-written essays. Observations of certain writing patterns and RLM-made spelling mistakes and errors might have affected the grading behaviours of HMs. On the other hand, the lack of an HM control group with RLM-assistance at the initial grading stage could have challenged the observations and recommendations made with regard to RLM-assisted marking.

In response to the limitations, considerations should be made in future research to test and refine the co-evaluation strategies for a more in-depth analysis of the change and differences of marking and grading with and without RLM assistance. It is also valuable to implement the strategies in the evaluation of authentic assessments to identify practical challenges.

Appendix

Other details of this pilot study such as the prompts for essay generation and grading, marking files of the HMs and the grading responses of the RLMs have been documented in the appendix, which can be retrieved from https://drive.google.com/drive/folders/1bkr5AXCRdtI1p12JdJYvNufz9SSuhMsh?usp=sharing.

References

Anthropic (2025). Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, 10(1), 37–53. https://doi.org/10.18608/jla.2023.7801

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, 51, 201-224. https://doi.org/10.1002/berj.4069

Fragiadakis, G., Diou, C., Kousiouris, G., & Nikolaidou, M. (2025). Evaluating Human-AI Collaboration: A Review and Methodological Framework. https://arxiv.org/pdf/2407.19098v2

Gabon, D. C., Vinluan, A. A., & Carpio, J. T. (2025). Automated Grading of Essay Using Natural Language Processing: A Comparative Analysis with Human Raters Across Multiple Essay Types. *Journal of Information Systems Engineering and Management*, 10(6s), 65-72. 10.52783/jisem.v10i6s.700

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., & Gao, Z. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. https://arxiv.org/pdf/2501.12948

Jason, W., Xuezhi, W., Dale, S., Maarten, B., Brian, I., Fei, X., Ed, C., Quoc, L., Denny, Z. (2022). Chain-of Thought Prompting Elicits Reasoning in Large Language Models. https://doi.org/10.48550/arXiv.2201.11903

Krippendorff, K. (2019). *Content Analysis: An Introduction to Its Methodology* (4th Ed.). SAGE Publications. https://doi.org/10.4135/9781071878781

Kundu, A. (2024). Are Large Language Models Good Essay Graders? [Master's thesis, University of Alberta]

Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator — Krippendorff's Alpha Calculator: A User-Friendly Tool for Computing Krippendorff's Alpha Inter-Rater Reliability Coefficient. *MethodsX*, 12, 102545. https://doi.org/10.1016/j.mex.2023.102545

OpenAI (2025). OpenAI o3-mini Pushing the frontier of cost-effective reasoning. https://openai.com/index/openai-o3-mini/

Plaat, A., Wong, A., Verberne, S., Broekens, J., van Stein, N., & Bäck, T. (2024). Multi-Step Reasoning with Large Language Models, a Survey. https://doi.org/10.48550/arXiv.2407.11511

Santiago, C. (2017). Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://pypi.org/project/krippendorff/

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. https://doi.org/10.1145/3706468.3706527

xAI (2025). Grok 3 Beta — The Age of Reasoning Agents. https://x.ai/news/grok-3

Xiao, C., Ma, W., Song, Q., Xu, S., Zhang, K., Fu, Q., & Wang, Y. (2025). Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. https://dl.acm.org/doi/pdf/10.1145/3706468.3706507

Xu, W., Shahreeza, M., Hoo, W. L., & Yang, W. (2025). Explainable AI for Education: Enhancing Essay Scoring via Rubric-Aligned Chain-of-Thought Prompting. https://www.preprints.org/manuscript/202504.2338/v1

Zupanc, K., & Bosnić, Z. (2015). Advances in the field of automated essay evaluation. Informatica, 39(4), 383–395.