Vol.8 No.1 (2025)

Journal of Applied Learning & Teaching

ISSN : **2591-801X**

Content Available at : **http://journals.sfu.ca/jalt/index.php/jalt/index**

Educational justice. Reliability and consistency of large language models for automated essay scoring and its implications

Siti Bealinda Qinthara Rony[A]

Tan Xin Fei[B]

Sasa Arsovski[C]

A *Research Assistant, Faculty of AI and Robotics, Raffles University, Malaysia*

B *Undergraduate Student, Faculty of AI and Robotics, Raffles University, Malaysia*

C *Professor, Dean of Faculty of AI and Robotics, Raffles University, Malaysia*

Keywords

AES;
AI;
analysis;
artificial intelligence;
automated essay scoring;
analysis;
consistency;
generative artificial intelligence;
LLMs;
reliability.

Correspondence

bealindaqinthara@gmail.com [A]

Article Info

Received 29 November 2024
Received in revised form 17 January 2025
Accepted 17 January 2025
Available online 21 January 2025

DOI: https://doi.org/10.37074/jalt.2025.8.1.21

Abstract

Maintaining consistency in automated essay scoring is essential to guarantee fair and dependable assessments. This study investigates consistency and provides a comparative analysis of open-source and proprietary large language models (LLMs) for automated essay scoring (AES). The study utilized student essays, each assessed five times to measure both intrarater (using intraclass coefficient and repeatability coefficient) and interrater (concordance correlation coefficient) reliability across several models: GPT-4, GPT-4o, GPT-4o mini, GPT-3.5 Turbo, Gemini 1.5 Flash, and LLaMa 3.1 70B. Essays and marking criteria are used for prompt construction and sent to each large language model to obtain score outputs. Results indicate that the scores generated by GPT-4o closely align with human assessments, demonstrating fair agreement across repeated measures. Specifically, GPT-4o exhibits slightly higher concordance correlation coefficients (CCC) than GPT-4o mini, indicating superior agreement with human scores. However, qualitatively, it can be observed that all LLM models are not as consistent in terms of their scoring rationale/evaluation. Our study results indicate that the challenges currently faced in automated essay scoring with large language models need to be analyzed not only from a quantitative perspective but also qualitatively. Additionally, we utilize more sophisticated prompting methods and address the inconsistencies observed in the initial measurements. Despite the purported reliability of some models within our study, the selection of LLMs should be considered thoroughly during practical implementations for an AES.

Introduction

In the education sector, new technologies are used to further engage and create more cohesive learning and teaching experiences. Ghosh (2024) investigates the usage of student portals and implementation of online quizzes in classrooms as a part of a competitive game between students. Nevertheless, teachers play a crucial role in the delivery of curriculum and are also the greatest contributor to what a student manages to learn in an educational setting.

Good education should be a necessity, yet with that thought, often there is a great disparity between the ratio of students to teachers, especially in larger institutions. Despite being a necessity, the question of the quality of education and the focus on students' learning experience can often be compromised by the amount of workload a teacher has (Anglia, 2020; Kanwal et al., 2023). Part of what contributes to this overwhelming workload is essay marking (Warschauer & Grimes, 2008).

Automated essay scoring (AES) refers to the usage of machine systems to mark educational assessments and is a possible solution to alleviate the said burden. Such a solution will allow teachers to allocate more time towards teaching, overall maintaining the quality of education to not fall below acceptable standards, and possibly even elevating the quality to higher standards (Hénard & Roseveare, 2012). AES systems are not a new concept, with widely known systems dating back to the 1960s. In recent years, however, techniques such as machine learning and natural language processing algorithms have often been utilized (Wilks, 2005; Ramesh & Sanampudi, 2021).

Our study focuses on the usage of artificial intelligence (AI) powered AES systems, primarily reviewing the performance and consistency of open source and proprietary large language models (Barry, 2023) for the grading/scoring of essays. We compare the consistency across models and analyze consistency criteria for each model's repeated grading runs.

Literature review

AES systems like the Project Essay Grader (PEG), e-rater, and Intelligent Essay Assessor (IEA) were implemented in the 1990s to address the time-intensive nature of essay evaluation. Initial resistance in the 1960s arose from early AES prototypes relying on surface features (e.g., number of propositions, commas, and uncommon words), which overlooked important aspects like content (Hearst, 2000). Although these systems showed high interrater correlation, they failed to assess critical writing skills.

Over the years, the pursuit of better AES systems has put an emphasis on the system's ability to extract and evaluate direct features of an essay. Now, as LLMs' capabilities to process multi-modal inputs (images, text, videos) progress, the possibility that an essay's context and direct linguistic features i.e., semantic features, structural features, etc. can be understood also further expands. Where systematic differences between any system regarding essay assessment often study intrarater and interrater reliability (Kayapinar, 2014), this evaluation method still applies to recent AI-based AES systems. The literature review is further divided into two sections, in which the first section will highlight the challenges and concerns involved in AI-based AES, and the second section talks about recent usage of AI for AES systems, where various works focused on the purported feasibility of such methods.

Challenges in AI-based AES

There are several challenges associated with the creation and feasibility of any AES (Automated Evaluation System) model (Hussein et al., 2019). When analyzing large language models (LLMs) and AI systems that operate as prompt-and-response mechanisms, we must consider not only the robustness of the models but also the inputs they receive, such as the prompts. These challenges are interrelated; one challenge can directly or indirectly impact another.

Marking consistency is a critical limitation in AI grading systems, as similar essays can receive different marks, undermining trust among educators and students (Attali, 2013; Balfour, 2013). Inconsistencies can arise when slight variations in context, sentence structure, or phrasing lead to different scores for identical essays. AI systems also struggle with discrepancies in scoring compared to human experts, even when adhering to predefined rubrics (Perelman, 2014). Furthermore, inconsistencies may occur when grading large volumes of essays simultaneously.

Internal understanding refers to the model's ability to understand grading criteria and generalize. The AI model's inability to grasp nuances, especially in unconventional writing styles, can lead to inaccurate marks, either undervaluing creative work or inflating scores due to its limited context (Li et al., 2021; Awidi, 2024). AI may focus on rubric fulfilment rather than content quality, potentially inflating scores for essays lacking depth (Zhu, 2019).

Recent research in AES

Recent AES research explores deep learning and generative AI models, which don't require traditional feature extraction (Attali & Burstein, 2006; Uto, 2021). Deep learning models, however, show promise when combined with feature-engineered inputs for a hybrid AES system to enhance contextual understanding (Kurniawan et al., 2024; Ortiz-Zambrano et al., 2024; Faseeh et al., 2024). Despite the rise of deep learning, traditional systems still offer value, as evidenced by a model that achieved a high Pearson's correlation using NLP techniques (Adeyanju et al., 2024).

LLMs, such as GPT, have shown promise in AES due to their linguistic capabilities and reasoning (Mansour et al., 2024; Ouyang et al., 2022). For example, Mizumoto and Eguchi (2023) tested GPT-3 for scoring non-native English essays, finding that GPT marked scores with acceptable adjacent agreement to human raters. They also showed that combining GPT scores with linguistic features improved AES performance. Inspired by Mizumoto and Eguchi's approach, Li and Liu (2024) used the approach of utilizing different prompting techniques (Maclaren, 2024) and demonstrated GPT-4's superiority in scoring non-native Japanese essays compared to several other models when including all linguistic measures in their scoring criteria with multi-shot prompting.

Recent work by Pack et al. (2024) had also expanded the viability testing of essays concerning language proficiency to other mainstream LLMs aside from OpenAI (n.d.)'s, focusing on Google's PaLM 2 (through chatbot Bard), Anthropic's Claude 2, and OpenAI's GPT-3.5 and GPT-4. Using the intraclass correlation coefficient (ICC) reliability score, the researchers were able to analyze the interrater reliability between model scores to humans and intrarater reliability of each model based on measures between two different time gaps. With the exception of GPT-3.5, the other LLMs improved in terms of intrarater reliability over time, in which GPT-4 was the most reliable given repeated measures in separate instances. However, interrater reliability (or the validity of the models' scoring to human scores) decreased for GPT-3.5 and GPT-4 over time.

A notable way to further enforce standardization and increase intrarater abilities of LLMs for AES can be seen in works by Ishida et al. (2024) or Kim and Jo (2024), where they utilize pairwise evaluation or comparative judgement (Pollitt, 2011) with LLMs instead of a rubric-based approach (Ishida et al., 2024; Kim & Jo, 2024). The method generally involves giving a positive point to the essay that scores higher, and giving zero or negative points to the essay that loses in score - repeated in a round-robin fashion. In the study by Kim and Jo (2024), it showed that this approach improved the performance of both GPT-4 and GPT-3, and was statistically significant.

This pairwise comparison approach is likened to the use of Latent Semantic Analysis (Deerwester et al., 1990; Foltz, 1996) for the IEA (Foltz & Landauer, 1999). The IEA, however, would focus on a more semantic comparison where the meaning of a student's essay is extracted and then compared with similar texts of known quality. By looking at the essays within an interconnected setting semantic-wise or rubrics, such approaches may also alleviate the inherent randomness/non-deterministic nature of an LLM (Lee et al., 2022), as proprietary models often get updated - altering its underlying parameters. The problem is that when dealing with a large number of essays, the number of pairwise comparisons that can be made increases exponentially, raising concerns about computational viability.

Despite promising results, challenges thus remain, particularly with the non-deterministic nature of LLMs and potential issues with interrater reliability (Pack et al., 2024). Nevertheless, LLMs' potential as AES tools also extends beyond grading, as they can provide valuable feedback to both students and raters, improving efficiency and consistency (Xiao et al., 2024; Gombert et al., 2024). Despite challenges such as the lack of personalized feedback, it can still serve as a basis or a foundation for students' self-review/study (Meyer et al., 2024). And so, the ability of LLMs to enhance grading consistency positions them as an increasingly viable AES solution.

Dataset and essay marking pipeline

Our dataset consists of 126 essays. We used 26 essays from Diploma in Business students for the "Learning Skills" module for our statistical analysis and 100 essays with additional prompting to confirm study results. The dataset for statistical analysis includes the marked essays with grades (human scores), the original ungraded submissions, and marking criteria (Appendix B). These are processed through a general essay marking pipeline, as shown in Figure 1, where each essay is paired with the marking criteria for zero-shot prompting. The resulting scores are compiled into a table (Appendix A), and sample outputs are presented in Appendices C and D.

The study uses six models: five proprietary models (GPT-3.5 Turbo, GPT-4, GPT-4o, GPT-4o mini, and Gemini 1.5 Flash) and Meta (n. d.)'s LLaMa 3.1 70B, which is a more democratized LLM (in freely providing access to trained weights for implementation, and the general architecture). A brief overview of the models is provided below:

- **GPT-3.5 Turbo** (OpenAI): A fast and efficient version of GPT-3.5, ideal for time-sensitive tasks and general-purpose text generation.
- **GPT-4o** (OpenAI): A versatile version of GPT-4 designed for various applications, offering improved efficiency and adaptability.
- **GPT-4o mini** (OpenAI): A lightweight variant of GPT-4, optimized for small-scale applications requiring efficiency and compact form.
- **GPT-4** (OpenAI): A multimodal model with advanced reasoning and higher accuracy, capable of processing both text and images.
- **Gemini 1.5 Flash** (DeepMind): A fast and contextually accurate model, excelling in multi-modal processing and natural language understanding.
- **LLaMa 3.1 70B** (Meta): A powerful language model with 70 billion parameters, offering advanced text generation, complex reasoning, and translation capabilities.

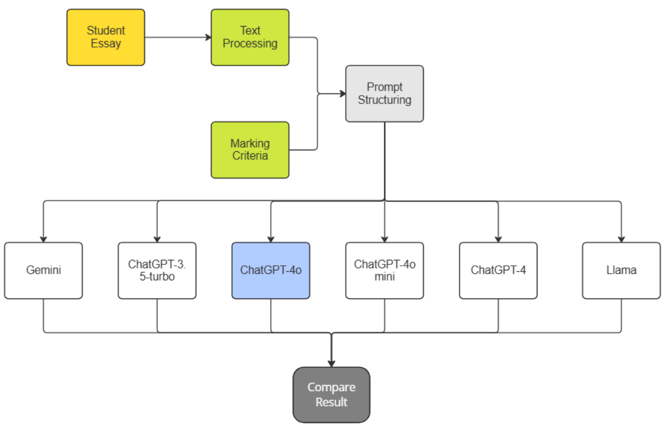

Figure 1. Essay marking pipeline.

Analysis of results

This section presents the results of five repeated measures of each student's essay for each model, focusing on consistency. The analysis includes both score marks and qualitative assessments of the text outputs. Reliability refers to the consistency and reproducibility of measurements, with two types (Baumgartner, 1989): relative reliability, which evaluates consistency across repeated trials (Bruton et al., 2000), and absolute reliability, which assesses measurement error, such as the Standard Error of the Mean (Stratford & Goldsmith, 1997).

To compare the marks given by LLMs and humans, we analyze both intrarater (the model's internal consistency) and interrater reliability (comparison between the model and human raters). Subjectively, we examine how well the model adheres to the baseline rubric across repeated measurements and students. The focus is on consistency, with a brief discussion of the reasonableness of generated feedback. Statistical methods used do not account for subjective score variations based on the extracted criteria.

Intrarater reliability

Intrarater reliability can be considered a type of relative reliability study that refers to how reliable *one* instrument/rater is across repeated measures (Gwet, 2008). Our study examines the consistency of a single model's scoring across repeated measures for multiple essays. We use the ICC to assess 'relative' reliability and measure internal scoring consistency. The ICC allows for the estimation of two types of consistency: absolute agreement and consistency (Nichols, 1998).

The ICC measure for absolute agreement assesses whether a model assigns identical scores to essays across repeated measurements. The ICC for consistency evaluates whether a model maintains the same relative ranking of students across measurements, regardless of score values. For example, if Student 1 ranks highest in the first measurement, consistency checks if they retain the top rank in subsequent measurements. Note that this ICC measure focuses on relative ranking, not absolute score consistency.

Based on a comprehensive research article on ICC by Liljequist et al. (2019), in our intrarater reliability study, we opted to analyze both ICC measures. The ICC measure of consistency excludes variance caused by bias in repeated measurements, unlike the absolute agreement measure, which accounts for it. While the choice of model (one-way random, two-way random, or two-way mixed) defines the scope and assumptions about bias, the calculation for consistency and absolute agreement remains the same for two-way random and mixed effects models. Hence, looking from an abstract level, the difference between the formula used between the two ICC measures for both two-way random and two-way mixed effect models can be said to lie in the inclusion of σ_c^2 (bias variance) for each repeated measurement with the same rater.

However, for our study, we will frame the ICC measures to use the two-way fixed effects model for a single measurement type, as declared below:

- A two-way mixed effects model: Our results only represent the specific reliability of each LLMs involved, where the LLMs determined as raters are not considered as a sample of a bigger population (Shrout & Fleiss, 1979). In practice, it means that the same rater is utilized for subsequent measurements considering a fixed bias (Koo & Li, 2016; Liljequist et al., 2019), such that the calculation of ICC cannot be generalized to other LLMs with similar characteristics;

- Single measurement type: In our study, we are not interested in the k average of raters, and are calculating ICC based only on a singular rater (model) for repeated measurements.

We use both ICC measures of consistency and absolute agreement to assess each model's internal consistency. ICC (A, 1) represents an absolute agreement, while ICC (C, 1) reflects consistency in our analysis. The primary objective of this study is to evaluate consistency across repeated measures for each model, with repeatability also considered.

Repeatability refers to the variation between two measurements observed under identical conditions (Bartlett & Frost, 2008). To better understand how consistent a model scores the same essay across repeated measurements, we are also thus concerned with the difference in values in successive measurements, where less variation should indicate better repeatability and better internal consistency. The repeatability coefficient use will quantify the maximum expected difference and is calculated as seen in Equation (1). The multiplier 2.77 arises from the properties of the normal distribution and accounts for the variability between successive measurements within 95% confidence interval (CI) (Vaz et al., 2013). Our RC calculation utilizes the mean pooled standard deviation to account for within-subject variances and between-subject variances, as seen in Equation (2).

$$RC = 2.77\sigma_w \quad (1)$$

$$\sigma_w = \sqrt{\frac{\sum_{i=1}^{5}(n_i - 1).Var_i}{\sum_{i=1}^{5}(n_i - 1)}} \quad (2)$$

All calculations of the ICC and repeatability consistency is calculated for each model, where the "irr" R statistical package, and Python are used for calculations. Intrarater Reliability Statistical Analysis given in Table 1.

We assess the relative reliability of the models using repeated measures under identical conditions, with values < 0.5 indicating poor reliability, 0.5-0.75 moderate, 0.75-0.9 good, and > 0.9 excellent (Koo & Li, 2016). Based on Table 1., GPT-4o mini has the highest ICC values for both absolute agreement and consistency, followed by GPT-4o with ICC (C, 1) values of 0.7819 and 0.7484, and ICC (A, 1) values of 0.787 and 0.749. Only these two models achieved ICC values

Table 1. Intrarater reliability statistical analysis of tested large language models.

	GPT-3.5 Turbo	GPT-4	GPT-4o	GPT-4o mini	Gemini 1.5 Flash	LLaMa 3.1 70B
ICC (C,1)						
Value	0.569*	0.509*	0.748*	0.782*	0.187**	0.285*
CI95%	[0.40,0.74]	[0.33,0.69]	[0.61,0.86]	[0.66,0.88]	[0.041,0.397]	[0.122,0.499]
ICC (A,1)						
Value	0.571*	0.516*	0.749*	0.787*	0.0815***	0.266*
CI95%	[0.4,0.74]	[0.339,0.7]	[0.62,0.86]	[0.67,0.883]	[0.001,0.221]	[0.111,0.475]
RC	20.4832	15.1534	12.6063	8.4700	24.2112	24.2988
Mean SD	7.3946	5.4706	4.5510	3.0578	8.7405	8.7721

**p<0.001, **p=0.004, ***p=0.0226. df1 for each model is 25 and df2 is 104.*

above 0.7, while older GPT versions (GPT-3.5 Turbo and GPT-4) had ICC values around 0.5.

Non-OpenAI models, Gemini 1.5 Flash and LLaMa 3.1 70B, show the lowest ICC values (<0.5) for both measures, with Gemini 1.5 Flash performing particularly poorly. Its absolute agreement ICC is below 0.1, indicating the model struggles to maintain consistent scores and relative rankings across repeated measures. Notably, GPT-4o and GPT-3.5 Turbo are the only models where the absolute agreement ICC slightly exceeds the consistency ICC, which could reflect data artifacts or bias in score variability.

The repeatability coefficient (RC) generally aligns with ICC values, with GPT-4o mini showing the smallest range of 8.47, indicating the highest consistency for successive measures. GPT-4, while achieving moderate reliability, has a higher RC of 12.61, but still ranks second in reliability. GPT-3.5 Turbo, despite its higher ICC values, has a larger RC, indicating more variability in successive measures.

Our findings suggest that GPT-4o mini and GPT-4o demonstrate the best internal consistency and reliability, particularly in maintaining relative rankings across repeated measures. GPT-4o mini, based on intrarater reliability, emerges as the most consistent model for essay scoring, ensuring reliable scores across repeated and successive measures.

Interrater reliability

Interrater reliability measures the agreement between raters (Lange, 2011). In our study, we assess the reliability of LLM-generated scores against a human rater. The analysis compares the scores from each model with the human rater using the concordance correlation coefficient (CCC), calculated pairwise for each essay. CCC was chosen over Pearson's r to quantify both correlation and agreement (Lin, 1989). The calculation treats the dataset as pairwise data, with repeated measurements forming new data points for each essay.

We interpret CCC values using Landis and Koch's scale: <0 (no agreement), 0–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), and 0.81–1 (almost perfect). The table includes the CCC estimate and 95% CI for each model. Unlike intrarater reliability, which measures consistency within a single rater, interrater reliability in our study evaluates the agreement between model outputs and human scores over multiple repetitions. Consistency is determined by how well the repeated model outputs align with a given human score.

A Bland-Altman analysis (also included in Table 2) was also conducted, though its assumptions - such as equivalent precision and constant bias - may not fully apply here (Taffé, 2021; Silveira et al., 2024). Since tolerance limits for acceptable score differences were not defined, the analysis serves as a general illustration rather than a definitive assessment of agreement (Indrayan, 2022). The calculations for CCC and Bland-Altman bias were performed using the "SimplyAgree" R package.

Table 2. Interrater reliability statistical analysis of tested large language models.

	GPT-3.5 Turbo	GPT-4	GPT-4o	GPT-4o mini	Gemini 1.5 Flash	LLaMa 3.1 70B
CCC						
Estimate	0.2860	0.0158	0.3954	0.3540	0.1716	0.2690
CI95%	[0.14, 0.42]	[-0.05,0.08]	[0.24,0.53]	[0.21,0.48]	[0.01,0.32]	[0.15,0.38]
Bland-Altman						
Mean Bias	5.1923	15.2615	2.0846	-1.9923	2.9154	8.7923
Lower LoA	-18.7316	-8.6804	-17.9151	-20.2980	-20.1026	-12.9478
Upper LoA	29.1162	39.2034	22.0844	16.3134	25.9333	30.5324

LoA = Limits of Agreement

Our interrater reliability study (Table 2.) shows that GPT-4o and GPT-4o mini achieve the highest CCC values, indicating fair agreement with human scores across repeated measures. This suggests both models produce consistent and reasonably accurate score outputs. However, GPT-4o demonstrates slightly better agreement than GPT-4o mini. In contrast, GPT-4 has the lowest CCC estimate, with true values close to 0, indicating minimal to no agreement with human scores.

The Bland-Altman analysis reveals that GPT-4o and GPT-4o mini exhibit similar bias (around 2 points), with GPT-4o mini underestimating and GPT-4o overestimating scores. Overall, GPT-4o and GPT-4o mini show the highest interrater reliability and are better suited for tasks requiring alignment with human evaluation.

Qualitative analysis

A qualitative review of the LLM's text outputs reveals a degree of randomness in their marking criteria. Shown in Appendix C are excerpts of GPT-4o and Gemini 1.5 Flash's first and repeated outputs, which we primarily investigate to give our qualitative analysis. Research data of the full-text outputs for all repeated measurements on each model will be made available upon reasonable request. Yet it is clear that qualitatively, repeated measures show notable differences in rationale which often correlate with variations in numerical scores. Some models also fail to correctly sum scores from subcriteria, leading to inconsistencies.

For instance, the GPT-4o model may change the naming and maximum score of subcriteria across repeated measures, such as changing "Referencing and Citation" from 10 marks to "Use of Sources and Referencing" with 15 marks. Additionally, some models would hallucinate new subcriteria or omit specified ones. Inconsistencies can also occur in how subcriteria are evaluated, such as Gemini 1.5 Flash initially breaking down "Content" into smaller aspects, but later assessing it based on identified strengths and weaknesses. Hence, while numerical scores may suggest some reliability, the rationale behind these scores reveals that LLMs' consistency in essay marking is less dependable than statistical analyses alone might imply.

Discussion

A key concern in using AI models as Automated Essay Scoring (AES) systems is whether they produce accurate scores compared to human raters. This issue arises not only from the subjectivity or objectivity of the scoring models but also from the inherent variability of AI generative models. The practical deployment of AI in AES requires trust in its scoring analysis, as the quality of AI models as replacements for human teachers remains uncertain (Barshay, 2024). In our study, we assess the consistency and reliability of selected AI models, helping readers to evaluate their suitability for practical applications.

Our findings suggest that GPT-4o and GPT-4o mini are internally reliable models. However, if both reliability and agreement with human scores are priorities, GPT-4o should be preferred, as it shows slightly higher agreement with human scores (higher CCC value). While GPT-4 demonstrates internal reliability (ICC > 0.5), it has the lowest CCC value when compared to human scores, showing minimal agreement and making it less suitable for practical use. Non-OpenAI proprietary models generally exhibit low consistency and agreement with human scores, supporting the preference for OpenAI models in AES applications.

Despite good internal consistency in models like GPT-4o and GPT-4o mini, achieving sufficient human-model agreement remains challenging, even with repeated measures showing fair agreement. Therefore, selecting the appropriate LLM for AES requires careful consideration. Qualitative inconsistencies are also evident across all models, including:

- Hallucination or omission of subcriteria
- Inconsistent definition of maximum marks or awarding of marks across repeated measurements
- Inconsistent evaluation of aspects within similar subcriteria
- Errors in summing total marks from subcriteria

While early research into the reliability of models like the GPT-3.5 (Khademi, 2023) showed low inter-reliability with human scores, more recent research involving GPT-4 (Pack et al., 2024; Tate et al., 2024), has shown it has become the most reliable for essay scoring in comparison to earlier GPT models and other non-Open AI models. Our study now suggests that GPT-4o and GPT-4o mini offer better reliability and human score agreement. This trend likely indicates that future versions of OpenAI's LLMs will become more reliable for AES applications.

Our study acknowledges limitations, such as the evolving nature of proprietary AI models and the impact of prompt engineering (Stahl et al., 2024). Since we used a simple rubric for zero-shot prompting, alternative methods may improve consistency. Nevertheless, our approach is suitable for measuring both internal consistency and human-model agreement for repeated measures. We tested further using additional prompting to mark 100 student essays using the GPT-4o mini. The results of further testing corroborated our study results, where the model achieves similar internal consistency and fair agreement.

The change in interrater reliability or intrarater reliability (if any) in such models is thus assumed to be not just due to an LLM's non-deterministic nature (Lee et al., 2022), but also on how proprietary models are constantly updated, thereby frequently altering its parameters. This, alongside a neural network's black-box nature, is a point to consider when trying to employ an LLM as an AES. Clear instructions through prompts thus cannot be overstated to standardize scoring and avoid cases of hallucination, as LLMs will generate different scoring criteria for each input when no criteria are specified (Ishida et al., 2024; Xiao et al., 2024). Future research thus could explore alternative prompting methods (Li, 2024; Kim & Jo, 2024) to further refine consistency analysis across different AI models, and also as a means to combat hallucinations – leading to more reliable feedback (Rudolph et al., 2024).

Regardless, even in cases with clear scoring outlines, the chances of encountering the models' quirks are never zero. Hence, as of current time, we once again note that both caution and deliberation should always be included in the usage of AI-based AES. Popenici (2022) noted even more reasons why this degree of caution should be exercised. AI is prone to algorithmic bias, and cases where bias stemming from technology affecting people is increasingly real (Popenici et al., 2023; Rudolph, 2023). Education, implemented justly is one of the fields where people of all kinds can truly share their experiences and knowledge, enriching one another through meaningful interaction on an even playing field. So, what happens if AI-based AES systems were implemented on a wide scale while still being riddled with the various concerns of reliability and fairness? The focus on consistency studies, or the pursuit of a more consistent AI model, is thereby highly imperative in education – as it is impossible to completely avoid rapid advancements in technology. Instead, its usages should be as tools by both teachers and students alike to enrich the learning experience, assisting and elevating current educational foundations and not as a means that redefines said foundations.

Conclusion

This study reinforces the critical role of consistency metrics in assessing LLMs for automated essay scoring, both quantitatively and qualitatively. By focusing on intrarater reliability to measure internal consistency and interrater reliability to see the model's consistency and agreement with human raters, we identified GPT-4o as a strong candidate for practical implementations. However, achieving human-level alignment without any possible aspect of randomness remains to be a challenge. While our findings provide a foundational approach to reliability testing in AES, future research must adapt to the evolving nature of LLMs, ensuring that the usage of these tools in the field meets the nuanced demands involved in educational assessment. Our study thus lays the groundwork for selecting and deploying more robust and trustworthy AI-based AES systems to further elevate current educational foundations.

References

Adeyanju, I. A., Rachael, O. K., Titilayo, A. O., Ajoke, G. O., Oyeladun, M. B., & Samuel, F. A. (2024). Artificial intelligence based essay grading system. *Engineering and Technology Journal, 9*(7). https://doi.org/10.47191/etj/v9i07.06

Anglia, N. (2020, August 19). *Does class size matter? The educational impact of teacher-student ratios*. Nord Anglia Education. https://www.nordangliaeducation.com/news/2020/08/19/does-class-size-matter-the-educational-impact-of-teacherstudent-ratios

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (2nd ed., pp. 181-198). Routledge. https://www.researchgate.net/profile/Yigal-Attali/publication/292810655_Validity_and_reliability_of_automated_essay_scoring/links/5bfbfe31299bf10737f8b7cf/Validity-and-reliability-of-automated-essay-scoring.pdf

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning, and Assessment, 4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Awidi, I. T. (2024). Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence, 6,* 100226. https://doi.org/10.1016/j.caeai.2024.100226

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated essay scoring and calibrated peer review. *Research & Practice in Assessment, 8*, 40-48. https://eric.ed.gov/?id=EJ1062843.

Barry, D. (2023, September 15). *A look at the large language model landscape*. Reworked. https://www.reworked.co/information-management/a-look-at-the-large-language-model-landscape/

Barshay, J. (2024, May 20). *Proof points: AI essay grading is already as 'good as an overburdened' teacher, but researchers say it needs more work*. The Hechinger Report. https://hechingerreport.org/proof-points-ai-essay-grading/

Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. *Ultrasound in Obstetrics and Gynecology, 31*(4), 466–475. https://doi.org/10.1002/uog.5256

Baumgartner, T. A. (1989). Norm-referenced measurement: Reliability. In M. J. Safrit & T. M. Wood (Eds.), *Measurement concepts in physical education and exercise science* (pp.45-72). Champaign: Human Kinetics Books.

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy, 86*(2), 94–99. https://doi.org/10.1016/s0031-9406(05)61211-4

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y. (2024). Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. *Mathematics, 12*(21), 3416. https://doi.org/10.3390/math12213416

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods Instruments & Amp Computers, 28*(2), 197–202. https://doi.org/10.3758/bf03204765

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer – Enhanced Learning, 1*(2). https://www.researchgate.net/publication/243770899_The_intelligent_essay_assessor_Applications_to_educational_technology

Ghosh, A. (2024, February 27). *Student portals: Fostering connectivity and educational growth*. Buddy4Study. https://www.buddy4study.com/article/student-portals

Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., & Drachsler, H. (2024). From the automated assessment of student essay content to highly informative feedback: A case study. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-023-00387-6

Google DeepMind. (n.d.). *Gemini 1.5 Flash*. Google Deep Mind. https://deepmind.google/technologies/gemini/flash/

Gwet, K. L. (2008). Intrarater reliability. *Wiley Encyclopedia of Clinical Trials*, 1–13. https://doi.org/10.1002/9780471462422.eoct631

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and Their Applications, 15*(5), 22–37. https://doi.org/10.1109/5254.889104

Hénard, F., & Roseveare, D. (2012). Fostering quality teaching in higher education: Policies and practices. *An IMHE Guide for Higher Education Institutions, 1*(1), 7-11. http://dx.doi.org/10.5901/mjss.2014.v5n25p272

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5,* Article e208. https://doi.org/10.7717/

peerj-cs.208

Indrayan, A. (2022). *Direct use of clinical tolerance limits for assessing agreement: A robust nonparametric approach.* SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4189799

Ishida, T., Liu, T., Wang, H., & Cheung, W. K. (2024, May 28). *Large language models as partners in student essay evaluation.* arXiv.org. https://arxiv.org/abs/2405.18632

Kanwal, A., Rafiq, S., & Afzal, A. (2023). Impact of workload on teachers' efficiency and their students' academic achievement at the university level. *Gomal University Journal of Research, 39*(2), 131–146. https://doi.org/10.51380/gujr-39-02-02

Kayapinar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research, 14*(57). https://doi.org/10.14689/ejer.2014.57.2

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning & Teaching, 6*(1), 75-80. https://doi.org/10.37074/jalt.2023.6.1.28

Kim, S., & Jo, M. (2024, July 8). *Is GPT-4 alone sufficient for automated essay scoring?: A comparative judgment approach based on rater cognition.* arXiv.org. https://arxiv.org/html/2407.05733v1

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kurniawan, W., Riantoni, C., Lestari, N., & Ropawandi, D. (2024). A hybrid automatic scoring system: Artificial intelligence-based evaluation of physics concept comprehension essay test. *International Journal of Information and Education Technology, 14*(6), 876–882. https://doi.org/10.18178/ijiet.2024.14.6.2113

Lange, R. T. (2011). Inter-rater reliability. In J. S. Kreutzer, J. DeLuca, B. (Eds.), *Encyclopedia of clinical neuropsychology.* Springer. https://doi.org/10.1007/978-0-387-79948-3_1203

Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. *CHI Conference on Human Factors in Computing Systems,* (388), 1-19. https://doi.org/10.1145/3491102.3502030

Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social Sciences Communications, 11*(1). https://doi.org/10.1057/s41599-024-03209-9

Li, Z., Zhang, J., Fei, Z., Feng, Y., & Zhou, J. (2021, June 4). *Addressing inquiries about history: An efficient and practical framework for evaluating open-domain chatbot consistency.* arXiv.org. https://arxiv.org/abs/2106.02228

Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation – a discussion and demonstration of basic features. *PLoS ONE, 14*(7), Article e0219854. https://doi.org/10.1371/journal.pone.0219854

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255. https://doi.org/10.2307/2532051

Maclaren, U. (2024). *Do you know when to use 0-shot, 1-shot, or multi-shot prompts (e.g. give it 1 or more examples)?* SSW Rules. https://www.ssw.com.au/rules/shot-prompts/

Mansour, W., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024, April 26). *Can large language models automatically score proficiency of written essays?.* arXiv.org. https://arxiv.org/pdf/2403.06149

Meta. (n. d.). *Llama 3.1 70B.* Hugging face. https://huggingface.co/meta-llama/Llama-3.1-70B

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*(1), 100199. http://dx.doi.org/10.1016/j.caeai.2023.100199.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050.

Nichols, D. P. (1998). *SPSS Library: Choosing an intraclass correlation coefficient.* OARC Stats. https://stats.oarc.ucla.edu/spss/library/spss-library-choosing-an-intraclass-correlation-coefficient/

OpenAI. (n. d.). *Models*. OpenAI platform. https://platform.openai.com/docs/models

Ortiz-Zambrano, J. A., Espín-Riofrío, C. H., & Montejo-Ráez, A. (2024). Deep encodings vs. linguistic features in lexical complexity prediction. *Neural Computing and Applications.* https://doi.org/10.1007/s00521-024-10662-9

Ouyang, L., Wu, J., Jiang, X., Almeida, D., L. Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, Al., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback (*36th Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems, Vol. 35*). https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Pack, A., Barrett, A., & Escalante, J. (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. https://doi.org/10.1016/j.caeai.2024.100234

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104–111. https://doi.org/10.1016/j.asw.2014.05.001

Pollitt, A. (2011). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Popenici, S. (2022). *Artificial intelligence and learning futures*. https://doi.org/10.4324/9781003266563

Popenici, S., Rudolph, J., Tan, S., & Tan, S. (2023). A critical perspective on generative AI and learning futures. An interview with Stefan Popenici. *Journal of Applied Learning & Teaching, 6*(2), 311-331. https://doi.org/10.37074/jalt.2023.6.2.5

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Rudolph, J. (2023). Book review. Popenici, Stefan (2023). Artificial intelligence and learning futures. Critical narratives of technology and imagination in higher education. Routledge. *Journal of Applied Learning & Teaching, 6*(2), 420-425. https://doi.org/10.37074/jalt.2023.6.2.27

Rudolph, J., Ismail, F., & Popenici, S. (2024). Higher education's generative artificial intelligence paradox: The meaning of chatbot mania. *Journal of University Teaching and Learning Practice, 21*(6). https://doi.org/10.53761/54fs5e77

Silveira, P. S. P., Vieira, J. E., & De Oliveira Siqueira, J. (2024). Is the Bland-Altman plot method useful without inferences for accuracy, precision, and agreement? *Revista De Saúde Pública, 58*(1). https://doi.org/10.11606/s1518-8787.2024058005430

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024, April 24). *Exploring LLM prompting strategies for joint essay scoring and feedback generation*. arXiv.org. https://arxiv.org/abs/2404.15845

Stratford, P. W., & Goldsmith, C. H. (1997). Use of the standard error as a reliability index of interest: An applied example using elbow flexor strength data. *Physical Therapy, 77*(7), 745–750. https://doi.org/10.1093/ptj/77.7.745.

Taffé, P. (2021). When can the Bland & Altman limits of agreement method be used and when it should not be used. *Journal of Clinical Epidemiology, 137*, 176–181. https://doi.org/10.1016/j.jclinepi.2021.04.004

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education Artificial Intelligence, 7*, 100255. https://doi.org/10.1016/j.caeai.2024.100255.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika, 48*(2), 459–484. https://doi.org/10.1007/s41237-021-00142-y

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test–retest reliability. *PLoS ONE, 8*(9), Article e73990. https://doi.org/10.1371/journal.pone.0073990

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies an International Journal, 3*(1), 22–36. https://doi.org/10.1080/15544800701771580.

Wilks, Y. (2005). *The history of natural language processing and machine translation. Encyclopedia of language and linguistics* (Vol. 9).

Xiao, C., Ma, W., Song, Q., Xu, S. X., Zhang, K., Wang, Y., & Fu, Q. (2024, January 12). *Human-AI collaborative essay scoring: A dual-process framework with LLMs*. arXiv.org. https://arxiv.org/abs/2401.06431

Zhu, W. (2019). A study on the application of automated essay scoring in college English writing based on PigAi. *Proceedings of the 7th International Conference on Social Science and Higher Education (ICSSHE 2021)*. https://doi.org/10.2991/icsshe-19.2019.188

Appendices

Appendix A: Score output for repeated measurements.

Human Scoring	GPT-3.5 Turbo				
	Measure 1	Measure 2	Measure 3	Measure 4	Measure 5
76	100	94	93	100	83
57	90	90	83	90	92
59	75	81	75	84	89
72	84	71	78	88	85
79	90	93	85	91	83
83	72	100	93	70	72
76	82	77	83	82	90
71	68	68	68	96	69
84	72	76	76	76	75
90	83	83	83	83	92
76	85	87	77	82	82
65	100	87	82	74	81
63	70	70	68	72	72
89	82	77	84	78	80
69	75	77	87	90	76
59	71	60	74	70	42
74	62	67	65	67	52
86	93	69	93	84	92
63	70	74	77	73	85
74	74	86	79	72	68
74	77	77	72	78	77
85	91	100	92	85	83
67	86	59	59	80	81
89	84	90	93	93	86
81	83	90	86	86	80
67	49	57	55	67	49

Human Scoring	GPT-4				
	Measure 1	Measure 2	Measure 3	Measure 4	Measure 5
76	100	92	90	100	100
57	100	100	94	88	100
59	95	87	90	100	85
72	100	95	100	95	100
79	94	94	92	94	100
83	92	100	100	100	92
76	84	75	82	70	80
71	100	87	98	95	100
84	75	90	90	88	95
90	92	94	100	93	96
76	84	84	85	85	89
65	92	90	91	88	92
63	83	80	85	80	84
89	80	80	90	100	91
69	92	95	90	88	90
59	100	90	80	100	100
74	75	81	90	81	75
86	95	100	91	80	80
63	75	100	75	81	80
74	89	92	85	90	86
74	72	85	77	88	77
85	87	94	92	93	96
67	83	90	80	82	90
89	85	75	83	90	83
81	100	95	100	100	95
67	85	83	80	86	86

Human Scoring	GPT-4o				
	Measure 1	Measure 2	Measure 3	Measure 4	Measure 5
76	78	81	87	84	83
57	76	80	75	75	80
59	75	74	74	73	97
72	72	73	72	73	67
79	86	83	92	92	75
83	83	83	84	83	83
76	64	67	67	67	68
71	85	85	83	85	85
84	74	73	78	81	75
90	85	85	85	90	89
76	73	73	61	61	56
65	80	85	81	81	73
63	74	74	69	69	65
89	86	81	77	77	85
69	66	69	69	69	72
59	78	75	69	81	83
74	63	61	61	61	57
86	84	81	84	84	91
63	65	65	67	71	69
74	74	81	79	75	95
74	75	69	78	67	72
85	83	83	83	81	89
67	68	71	69	71	78
89	80	80	81	88	95
81	82	82	82	82	82
67	61	61	61	56	57

Human Scoring	GPT-4o mini				
	Measure 1	Measure 2	Measure 3	Measure 4	Measure 5
76	87	81	81	80	80
57	73	71	71	76	77
59	69	66	71	65	66
72	76	71	75	71	76
79	78	79	82	81	76
83	83	77	85	72	78
76	64	64	62	64	65
71	66	76	78	78	78
84	69	69	68	75	70
90	87	87	83	77	90
76	65	66	65	65	64
65	72	72	72	72	72
63	60	59	66	66	63
89	71	71	77	77	72
69	68	68	68	68	68
59	71	80	73	80	73
74	63	73	59	67	66
86	75	76	75	78	78
63	69	69	67	67	69
74	75	74	76	72	69
74	71	71	71	71	72
85	80	79	80	80	80
67	70	83	70	70	69
89	70	71	71	71	71
81	71	71	75	76	71
67	61	60	61	59	60

Human Scoring	LLaMa 3.1 70B				
	Measure 1	Measure 2	Measure 3	Measure 4	Measure 5
76	93	92	86	93	95
57	88	74	70	78	66
59	81	78	92	70	71
72	84	84	92	83	80
79	75	84	91	92	91
83	84	90	94	98	90
76	75	75	77	93	82
71	91	91	92	95	95
84	93	95	97	91	95
90	93	94	94	94	92
76	75	81	79	91	83
65	71	78	84	91	65
63	71	74	79	91	76
89	86	90	92	68	74
69	67	80	64	60	66
59	84	88	84	60	83
74	95	96	81	85	83
86	95	95	84	74	74
63	94	92	82	47	74
74	94	97	76	78	80
74	83	94	73	62	73
85	94	92	83	80	84
67	94	83	81	56	74
89	92	86	86	69	84
81	77	83	93	83	80
67	74	75	93	72	76

Appendix B: Scoring criteria.

FINAL ESSAY ASSESSMENT (100 MARKS)					
CLO	CRITERIA	0-4 marks	5-9 marks	10-17 marks	18-20 marks
3	*Organisation of ideas (coherence and logical structure)*	Poor coherence and/or logical structure.	Average coherence and/or logical structure.	Good coherence and/or logical structure.	Excellent coherence and/or logical structure.
2 & 3	*Critical analysis, evaluation of thesis*	Poor analysis, evaluation and/or reflection. Mostly own opinion with no evidence presented.	Average analysis, evaluation and/or reflection. Lack of reasonable evidence given.	Good analysis, evaluation and/or reflection. Average evidence presented to support analysis.	Excellent analysis, evaluation and/or reflection. Concrete evidence presented to support critical analysis.
2 & 3	*Comprehension of the topic and related issues*	Poor understanding of the topic and/or related issues.	Average understanding of the topic and/or related issues.	Good understanding of the topic and/or related issues.	Excellent understanding of the topic and/or related issues.
		1-2 marks	3-5 marks	6-8 marks	9-10 marks
2 & 3	*conclusion*	Poor evidence of preceding arguments and/or overall conclusion.	Average evidence of preceding arguments and/or overall conclusion.	Good evidence of preceding arguments and/or overall conclusion.	Excellent evidence of preceding arguments and/or overall conclusion.
3	*criteria adherence, formatting and use of language*	Poor criteria adherence, formatting and/or use of language.	Average criteria adherence, formatting and/or use of language.	Good criteria adherence, formatting and/or use of language.	Excellent criteria adherence, formatting and/or use of language.
3	*wide reading (citation and referencing)*	Poor use of citation and/or referencing.	Average use of citation and/or referencing.	Good use of citation and/or referencing.	Excellent use of citation and/or referencing.
		1 mark	2 marks	3-4 marks	5 marks
3	*Time management (Draft)*	Late submission (delay 3 days)	Late submission (delay 2 days)	Late submission (delay 1 day) with incomplete work submitted	Submit work on time and complete (both map & questions)
3	*Time management (Final)*	Late submission (delay 3 days)	Late submission (delay 2 days)	Late submission (delay 1 day) with incomplete work submitted	Submit work on time and complete (both map & questions)

Appendix C: Excerpt text outputs.

GPT-4o (prompt and output)

Prompt

You are a business course lecturer. Here are the criteria for marking: *attach pdf file*.
Please mark the essay according to the provided criteria, detailing the marking of each part. The total marks should be out of 100.
Here is the essay submitted by the student using docs: *attach docs file*.

Output: first measure

Marking Criteria Breakdown

1. **Introduction and Thesis Statement (10 marks)**
 - **Marks Awarded: 8/10**
 - **Comments:** The introduction effectively sets the stage for the essay by highlighting the urgency of climate change and the need for renewable energy. The thesis statement is clear and concise, stating that renewable energy is essential to human life because it affects food, water supply, and health. However, it could be more engaging to capture the reader's attention better.

2. **Understanding and Use of Sources (20 marks)**
 - **Marks Awarded: 18/20**
 - **Comments:** The essay demonstrates a strong understanding of the sources cited. The references are relevant and well-integrated into the argument. The student uses a variety of credible sources, including reports from the National Oceanic and Atmospheric Administration and the Union of Concerned Scientists. However, there is a slight over-reliance on the National Oceanic and Atmospheric Administration, which could have been balanced with more diverse sources.

3. **Argument Development and Cohesion (20 marks)**
 - **Marks Awarded: 16/20**
 - **Comments:** The essay is well-structured, with each paragraph focusing on a specific aspect of the argument. The points are logically developed and supported by evidence. However, some transitions between paragraphs could be smoother to enhance the overall flow of the essay.

4. **Critical Analysis and Original Thought (20 marks)**
 - **Marks Awarded: 15/20**
 - **Comments:** The essay provides a good critical analysis of the impact of climate change on food, water, and health. The student makes a compelling case for the necessity of renewable energy. However, the analysis could be deepened by exploring counterarguments or potential challenges in implementing renewable energy solutions.

5. **Conclusion (10 marks)**
 - **Marks Awarded: 8/10**
 - **Comments:** The conclusion effectively summarizes the main points and reinforces the thesis statement. It also provides a call to action for government initiatives, which adds a practical dimension to the argument. However, it could be more impactful by reiterating the urgency and importance of the issue.

6. **Grammar, Spelling, and Formatting (10 marks)**
 - **Marks Awarded: 9/10**
 - **Comments:** The essay is well-written with minimal grammatical and spelling errors. The formatting is consistent, and the references are correctly cited. There are a few minor errors, such as the lowercase "the" at the beginning of a sentence in the third paragraph.

7. **Referencing and Citation (10 marks)**
 - **Marks Awarded: 9/10**
 - **Comments:** The essay includes a comprehensive list of references, which are correctly formatted according to the required citation style. The in-text citations are appropriately placed. However, the reference list could be more alphabetically organized.

Total Marks: 83/100

Summary
The essay is well-researched and effectively argues the importance of renewable energy in mitigating the impacts of climate change on food, water, and health. It demonstrates a good understanding of the topic and uses credible sources to support the argument. The structure is logical, and the writing is clear and concise. However, there is room for improvement in the depth of critical analysis and the smoothness of transitions between paragraphs. Overall, it is a strong essay that meets most of the marking criteria effectively.

Output: repeated measure

Marking Criteria Breakdown

1. **Introduction and Thesis Statement (10 marks)**
2. **Argument Development and Evidence (30 marks)**
3. **Organization and Structure (20 marks)**
4. **Clarity and Style (15 marks)**
5. **Use of Sources and Referencing (15 marks)**
6. **Conclusion (10 marks)**

Detailed Marking

1. Introduction and Thesis Statement (10 marks)
- **Score: 8/10**
- **Comments:** The introduction effectively sets the stage for the essay by highlighting the urgency of climate change and the need for renewable energy. The thesis statement is clear and outlines the main points that will be discussed: food and water supply, and health. However, it could be more concise.

2. Argument Development and Evidence (30 marks)
- **Score: 25/30**
- **Comments:** The essay provides a well-rounded argument supported by credible sources. Each point is backed by evidence, primarily from reputable organizations like the National Oceanic and Atmospheric Administration and the Union of Concerned Scientists. However, some sections are repetitive, and the argument could be more tightly woven.

3. Organization and Structure (20 marks)
- **Score: 18/20**
- **Comments:** The essay is well-organized, with each paragraph focusing on a specific aspect of the argument. The transitions between paragraphs are smooth, and the structure is logical. However, the essay could benefit from more subheadings to guide the reader through the different sections.

4. Clarity and Style (15 marks)
- **Score: 12/15**
- **Comments:** The writing is generally clear and concise, with a formal academic tone. However, there are a few grammatical errors and awkward phrasings that detract from the overall readability. For example, "the farmers and ranchers that provide the food on our tables" could be more succinctly phrased.

5. Use of Sources and Referencing (15 marks)
- **Score: 13/15**
- **Comments:** The essay makes excellent use of a variety of sources, all of which are properly cited in APA format. The references are relevant and add credibility to the argument. However, the essay could benefit from integrating the sources more seamlessly into the narrative.

6. Conclusion (10 marks)
- **Score: 8/10**
- **Comments:** The conclusion effectively summarizes the main points and reiterates the importance of renewable energy. It also provides a call to action for government involvement, which adds a persuasive element. However, it could be more concise and impactful.

Total Score: 84/100

General Comments
The essay is well-researched and presents a compelling argument for the importance of renewable energy. It is well-organized and makes good use of credible sources. However, there are areas for improvement in terms of clarity, conciseness, and integration of sources. Overall, it is a strong essay that meets the criteria well.

Gemini 1.5 Flash (prompt and output)

Prompt

You are a business course lecturer. Here are the criteria for marking image: *attach image*.
Here is the essay submitted by the student.
[Student Essay in plain text format]
Please mark the essay according to the provided criteria, detailing the marking of each part. The total marks should be out of 100.

Output: first measure

Essay Marking: Why Renewable Energy is Essential to Human Life

Total Marks: 85

Content (40 Marks):

- **Relevance to Topic (10 Marks):** The essay directly addresses the importance of renewable energy. (10/10)
- **Depth of Knowledge (15 Marks):** The essay outlines the negative impacts of climate change on food, water, and health, and how renewable energy can help mitigate these issues. However, the discussion on economic and social aspects of renewable energy is missing. (10/15)
- **Accuracy of Information (15 Marks):** The essay uses credible sources to support its claims. Citations are well-formatted. (15/15)

Output: repeated measure

Essay Marking: Why Renewable Energy is Essential to Human Life

Total Marks: 85

Content (40 marks)

- Strength:
  - Clearly establishes the importance of renewable energy by highlighting the detrimental effects of climate change on food, water, and health.
  - Uses credible sources to support arguments.
  - Mentions the advantages of renewable energy (low emissions, sustainability).
- Weakness:
  - Lacks a strong focus on the business aspects of renewable energy (e.g.,

I will mark the essay according to the provided criteria.

CLO 3: Organisation of Ideas (Coherence and Logical Structure): 18/20 The essay has a clear and logical structure, with a well-defined introduction, body, and conclusion. The writer presents a clear thesis statement and supports it with relevant evidence throughout the essay. However, there are some minor issues with transitions between paragraphs, which could be improved to enhance the overall coherence.

CLOs 2 & 3: Critical Analysis, Evaluation of Thesis: 19/20 The writer demonstrates excellent critical analysis and evaluation skills, presenting a clear and well-supported argument for the importance of renewable energy. The writer provides concrete evidence from credible sources to support their claims, and effectively evaluates the impact of climate change on human life. However, there could be more nuance in the analysis, and some of the points could be more deeply explored.

CLOs 2 & 3: Comprehension of the Topic and Related Issues: 19/20 The writer demonstrates an excellent understanding of the topic and related issues, including the impact of climate change on human life, the ...ance of renewable energy, and the role of government initiatives. The writer provides a comprehensive

Tools (0)