

An Evaluation of Agreement and Uncertainty in LLM-Based Automated Essay Scoring

Yiting Yao, Yanyun Yang, and Huan (Hailey) Kuang

Florida State University

Correspondence: yy21b@fsu.edu

Abstract

This study evaluates score agreement and uncertainty estimation in large language model (LLM)-based automated essay scoring. Results showed that LLMs tended to be overconfident and highly consistent in their assigned scores, and majority voting did not yield stronger agreement with human scores than deterministic scoring across rubric score levels.

1 Introduction

Automated essay scoring (AES) has been used in various educational and assessment settings to support the evaluation of student writing (Williamson et al., 2012). In large-scale, high-stakes assessments, where the demand for timely score reporting makes human scoring impractical, scoring accuracy becomes a central concern in AES research (Dikli, 2006). Traditional AES systems, beginning with the Project Essay Grader (Page, 1966), generally rely on supervised machine learning approaches that use handcrafted linguistic features, such as lexical diversity, syntactic complexity, and discourse markers (Ramesh and Sanampudi, 2022; Voss, 2025). These systems can produce scores that closely align with human scores; however, they face three key challenges: (1) higher-order constructs (e.g., coherence and argument quality) may not be directly measured (Ramesh and Sanampudi, 2022), (2) extending feature pipelines requires substantial linguistic and domain expertise (Alikaniotis et al., 2016), and (3) performance weakens when models are applied across tasks and rubrics (Ridley et al., 2020).

Recent work has applied large language models (LLMs) to AES as they are able to capture semantic and discourse-level structures in an end-to-end manner, reducing the need for handcrafted feature engineering. With carefully constructed prompts and/or light fine-tuning, models such as GPT-4 can produce holistic scores, trait-level

scores (i.e., scores for human-defined dimensions, such as organization and grammar), and rubric-aligned rationales (Lee et al., 2024). Ideally, LLM-generated scores should be consistent with human-assigned scores, such that essays receiving high or low scores from human raters also receive high or low scores from LLMs. However, LLM-generated scores do not consistently demonstrate high agreement with human-assigned scores. Huang and Wilson (2025) analyzed English writing data from students in Grades 3–5 and found that both ChatGPT-4o and Claude Sonnet 4 showed relatively low agreement with human scores under baseline chain-of-thought (CoT) prompting, as measured by quadratic weighted kappa (QWK = 0.46 and 0.30, respectively). They also found that human–LLM score agreement improved when context-enhanced and few-shot CoT prompting were used. Gaggioli et al. (2025) evaluated five LLMs, including Claude Sonnet 3.5 and GPT-4, on Italian-language essays written by college students and found that human–LLM agreement was consistently low and non-significant. A recent synthesis of 65 studies (Li et al., 2026) reported human–LLM agreement ranging from 0.30 to 0.80, with substantial heterogeneity possibly attributable to differences in datasets, model capabilities, prompting strategies, and rubric specificity.

One notable challenge in LLM-based scoring is scoring instability; that is, LLMs may generate different scores across repeated calls using the same prompt. Researchers often control the randomness of model outputs by adjusting the temperature parameter. Specifically, at a temperature of 0, the model’s output is most deterministic and consistent, while at a higher temperature, outputs vary more across repeated runs of the same prompt. Accordingly, two scoring methods can be employed in LLM-based scoring: (1) assigning a single deterministic score, typically with the temperature set to 0, and (2) aggregating scores across multiple runs

using majority voting. A single deterministic score is cost-efficient and easy to implement, whereas multiple stochastic runs produce a predicted score distribution from which a majority vote score can be obtained and used for further analysis (Korthals et al., 2026). Some studies indicated that majority voting improved scoring accuracy and robustness (Korthals et al., 2026), whereas others found that majority voting did not outperform the deterministic scoring method when systematic model-specific bias remained uncorrected (Xue et al., 2026). Furthermore, Mathew et al. (2026) found that LLMs using deterministic scoring tended to assign lower scores to high-scoring essays and higher scores to low-scoring essays. It remains unclear whether majority voting can mitigate this score-compression pattern by introducing sampling variation, particularly for borderline essays in which the model is genuinely uncertain (Xue et al., 2026). The current study therefore evaluates the agreement of LLM-generated deterministic and majority-vote scores with human-assigned scores across rubric score levels.

In addition to scoring stability, we address a related question: Can LLMs’ uncertainty (lack of certainty) estimates indicate when their scores are likely to disagree with human ratings? Uncertainty estimates can provide insight into LLM decision-making, offer evidence regarding the reliability of LLM-generated scores, and make the scoring process more transparent and interpretable. This uncertainty-based framework also enables selective automation, in which LLMs automatically score essays with high certainty while sending uncertain cases to human raters (Ferrer et al., 2026). Such a human–AI collaborative workflow may help preserve scoring quality for difficult-to-score essays while substantially reducing human raters’ workload on routine cases.

The current study examines two uncertainty estimation methods. The first method is verbalized confidence, also called self-reported confidence, which asks the model to report a confidence value (e.g., 0–100) for its assigned score. This approach is straightforward and easy to implement. Xiong et al. (2024) found that LLMs tended to be overconfident when verbalizing their confidence. Yet this approach may still provide useful information, as Yang et al. (2024) reported that verbalized confidence can serve as an effective signal of human–LLM scoring disagreement when appropriate prompts are used.

The second approach is to use score consistency to quantify uncertainty in LLM scoring. LLMs score the same essay multiple times using the same prompt, and the standard deviation of the generated scores is used as a natural estimate of uncertainty (Ferrer et al., 2026; Korthals et al., 2026). Reporting score consistency is intuitive, as for genuinely difficult essays, such as essays that are vague, rhetorically complex, or near a score boundary, LLMs are likely to produce unstable scores, whereas scores for high-confidence cases should show stronger agreement across runs. Korthals et al. (2026) found that the score-consistency method improved human–LLM scoring agreement; however, Ferrer et al. (2026) indicated that score consistency did not show significant advantages over verbalized confidence while also requiring more LLM calls.

2 Study Purpose

Prior studies on LLM uncertainty estimation have focused predominantly on binary outcomes, such as correct versus incorrect answers, whereas less attention has been given to polytomous scoring contexts, such as ordinal scores aligned with a scoring rubric. In addition, it remains unclear whether uncertainty estimates function similarly across different LLMs. The current study fills these gaps by applying two state-of-the-art LLMs, GPT-5.1 and Claude Sonnet 4.6, to score essays from a corpus using a six-level holistic rubric. GPT-5.1 and Claude Sonnet 4.6 were selected because they represent two current frontier LLMs from major providers, allowing the study to evaluate LLM-based AES with contemporary model capabilities. This study addresses two research questions to evaluate scoring agreement and uncertainty estimation in LLM-based AES:

RQ1: Does human–LLM score agreement differ between deterministic scoring and majority voting across rubric score levels?

RQ2: Do verbalized confidence and score consistency methods flag essays for which LLM scores disagree with human-assigned scores?

3 Method

3.1 Data

This study uses data from the Automated Student Assessment Prize (ASAP) 2.0 corpus (Crossley et al., 2025), which includes approximately 25,000 source-based argumentative essays for seven essay

stems written by U.S. secondary students. Each essay was scored holistically by human raters on a six-point rubric, with 1 indicating the least proficient response and 6 the most proficient. We selected one Grade 10 essay stem from the dataset, namely, *Facial Action Coding System*. In this prompt, students read an informational article and wrote an essay arguing whether using emotion-reading technology in classrooms is valuable. To obtain a balanced design across the score scale, we drew a sample of 300 essays, with 50 essays randomly selected at each of the six score levels. An additional twelve exemplars (two per score level) were randomly drawn as few-shot examples. Each model prompt presented the full scoring rubric and elicited a brief CoT rationale before asking the model to assign a score.

3.2 Models

We evaluated two state-of-the-art LLMs: GPT-5.1 (OpenAI, 2025) and Claude Sonnet 4.6 (Anthropic, 2026). The temperature parameter was manipulated for both models to control randomness in model output. For the verbalized confidence run, the temperature was set to 0 to produce a deterministic score and confidence value per essay. For the score consistency run, we set the temperature to 0.7 to ask the model to score the same essay five times (Wang et al., 2023).

Both models received the same prompt with a fixed structure: (a) a role assignment instructing the model to act as an experienced Grade 10 English teacher, (b) the source article students read for the writing task, (c) the six-level scoring rubric, (d) twelve exemplars, each shown as an essay paired with its human-assigned score, and (e) the target essay to be scored. The prompt instructed the model to consider every element in the rubric before assigning a score and to return its response as a structured JSON object with a rationale, the score, and a confidence value (see Appendix). This prompt design combines the potential benefits of few-shot and CoT prompting by illustrating each rubric level with two concrete scored examples and directing the model to analyze the rubric before assigning a score.

3.3 Uncertainty Estimation

Verbalized Confidence. The model was prompted to report its confidence in the assigned score on a 0–100 scale. Each application programming interface call returned a rationale, a score, and the associated

confidence value as a single structured output. The computational cost was one LLM call per essay per model.

Score Consistency. The model scored the same essay five times (Xiong et al., 2024; Xue et al., 2026). The majority vote across the five samples was taken as the final score, and ties were resolved by selecting the higher tied score (e.g., selecting 3 from scores of 3–3–2–2–1), given that prior research suggests that LLMs tend to underrate top-performing essays (Korthals et al., 2026). The standard deviation across the five samples was used to quantify score consistency. The computational cost was five LLM calls per essay per model.

3.4 Evaluation of Model Performance

For RQ1, we evaluated human–LLM agreement in each model using two scoring methods (deterministic score vs. majority voting) with three metrics: QWK, exact agreement, and adjacent agreement. QWK measures the extent of agreement between two sets of ratings using quadratic weights that penalizes large score differences more heavily than near misses. A QWK value of 0 indicates agreement due to chance, whereas a value of 1 indicates perfect agreement (Williamson et al., 2012). A threshold of .70 is often used as a rule of thumb. QWK values were calculated to evaluate the overall agreement and agreement by rubric score level for both models. Exact agreement is the proportion of essays on which the LLM score matches the human score. Adjacent agreement is the proportion of essays for which the LLM score is within 1 point of the human score. A count-weighted scatter plot of human–LLM scores was presented to visualize any systematic underrating or overrating by LLMs.

For RQ2, we reported descriptive statistics for the two uncertainty estimates in each model: verbalized confidence and score consistency. Bar charts were also used to examine whether uncertainty estimates distinguished essays with matched human–LLM scores from those with mismatched scores.

4 Results

4.1 RQ1: Human–LLM Agreement

To answer RQ1, Table 1 reports QWK values and exact/adjacent agreement rates for each scoring method and model. GPT-5.1 achieved strong agreement with human raters (QWK = 0.80 and 0.79 for deterministic scores and majority voting, re-

Metric	Scoring Method	GPT-5.1	Claude 4.6
QWK	Deterministic	0.80	0.65
	Majority Vote	0.79	0.66
Exact Agreement	Deterministic	42.00%	34.33%
	Majority Vote	41.33%	32.00%
Adjacent Agreement	Deterministic	91.00%	71.33%
	Majority Vote	89.33%	71.67%

Table 1: Human–LLM Agreement by Scoring Method and Models.

Score Level	Scoring Method	GPT-5.1	Claude 4.6
Low (1–2)	Deterministic	0.45	0.68
	Majority Vote	0.44	0.64
Medium (3–4)	Deterministic	0.44	0.22
	Majority Vote	0.38	0.23
High (5–6)	Deterministic	0.11	0.06
	Majority Vote	0.10	0.07

Table 2: Human–LLM Agreement (QWK) by Rubric Score Level and Scoring Method.

spectively) and outperformed Claude Sonnet 4.6 (QWK = 0.65 and 0.66 for deterministic scores and majority voting, respectively). Within each model, there was no noticeable difference in QWK when comparing majority voting and deterministic scoring.

To examine whether score agreement differed across rubric score levels, we computed QWK within three rubric score levels (shown in Table 2): low (human scores 1–2), medium (3–4), and high (5–6). For both models and both scoring methods, QWK declined monotonically from low-scoring to high-scoring essays, though to differing extents. Under deterministic scoring, GPT-5.1’s QWK dropped from 0.45 for low-scoring essays to 0.11 for high-scoring essays, while Claude Sonnet 4.6’s QWK dropped from 0.68 to 0.06. The majority voting produced near-identical declines (GPT-5.1: 0.44 to 0.10; Claude Sonnet 4.6: 0.64 to 0.07). Within each scoring level, the two scoring methods yielded QWKs that differed by no more than 0.06, indicating that majority voting over five runs did not improve score agreement over deterministic scoring at any rubric score level. Larger QWK differences were observed between models. Claude Sonnet 4.6 showed stronger agreement with human raters for low-scoring essays, whereas GPT-5.1 showed stronger agreement for medium-scoring essays. Neither model performed well on high-scoring essays.

Method	Statistic	GPT-5.1	Claude 4.6
Verbalized Confidence	Mean	87.45	78.08
	SD	2.83	6.74
	Min	78.00	55.00
	Max	98.00	99.00
Score Consistency	Mean	0.20	0.11
	SD	0.22	0.19
	Min	0.00	0.00
	Max	0.63	0.49

Table 3: Descriptive Statistics for Uncertainty Estimation Methods.

The count-weighted scatterplot in Figure 1 further illustrates this pattern. For low-scoring essays, Claude Sonnet 4.6 produced 42 exact matches whereas GPT-5.1 had 22–24 exact matches for essays with a human score of 1; both models produced similar number of exact matches (40–43) for essays with a human score of 2. Among the mismatched cases, GPT-5.1 tended to overrate low-scoring essays. For medium-scoring essays, GPT-5.1 had 22–28 exact matches, and most mismatched cases reflected underrating. In contrast, Claude Sonnet 4.6 produced only 5–11 exact matches, and all mismatched cases were underrated. Underrating was more severe for high-scoring essays, with GPT-5.1 performing slightly better than Claude Sonnet 4.6. Most mismatched GPT-5.1 scores remained within the adjacent-agreement range, whereas Claude Sonnet 4.6 often underrated essays by two score levels. This pattern is also reflected in the exact and adjacent agreement rates reported in Table 1. GPT-5.1 consistently outperformed Claude Sonnet 4.6 on both measures, with the difference being particularly pronounced for adjacent agreement. Specifically, GPT-5.1 achieved adjacent agreement rates of around 90% or higher, whereas Claude Sonnet 4.6 achieved only approximately 71%. The lower adjacent agreement for Claude Sonnet 4.6 was primarily attributable to its tendency to underrate high-scoring essays.

4.2 RQ2: Uncertainty Estimates as Signals of Score Disagreement

To answer RQ2, Table 3 summarizes the descriptive statistics for verbalized confidence and scoring consistency across 300 essays for each model. GPT-5.1 reported higher confidence in its assigned scores (mean = 87.45, range = 78–98) than Claude Sonnet 4.6 (mean = 78.08, range = 55–99). However, GPT-5.1 produced less consistent scores (mean = 0.20) than Claude Sonnet 4.6 (mean = 0.11).

Human-LLM Agreement by Scoring Method

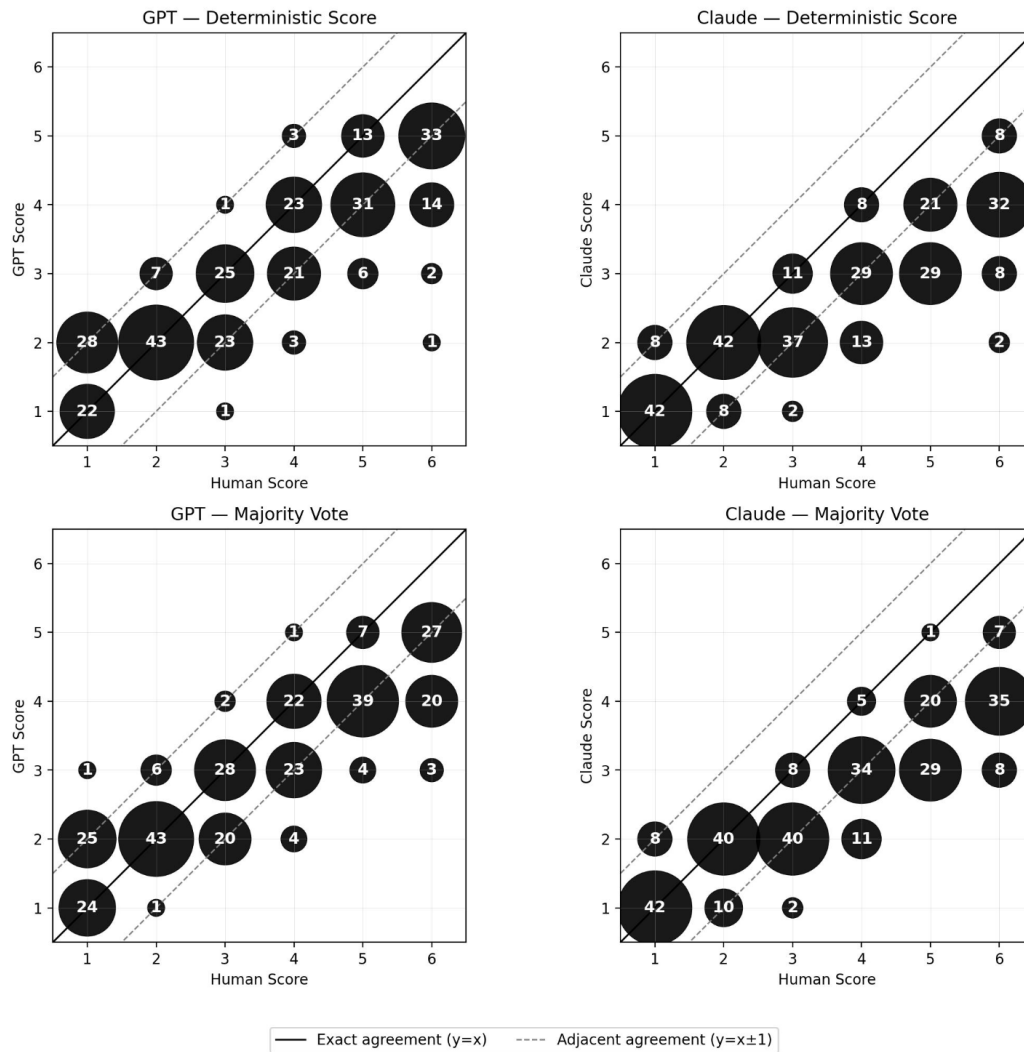


Figure 1: Weighted Scatter Plot for Human-LLM Agreement across Scoring Methods.

To examine whether uncertainty estimates varied in their ability to flag disagreement across rubric score levels, Figures 2–3 demonstrate the distribution of uncertainty estimates for each of the six rubric score levels, along with the human-LLM matched cases and mismatched cases. Matched cases were defined by the deterministic score for the verbalized confidence method and the majority vote for the score consistency method. Both models tended to be confident and consistent in their assigned scores as verbalized confidence centered on the upper scale, and most essays across five runs had standard deviations of zero. For low-scoring essays, both models produced exact matches for most essays. However, matched and mismatched cases were both associated with high verbalized confidence (above 80) and little variation across samples

(standard deviation around zero). For high-scoring essays, most essay scores were mismatched, yet these cases showed uncertainty levels comparable to those of matched cases at lower score levels. For Claude Sonnet 4.6, among the medium- and high-scoring essays, most mismatched cases had standard deviations smaller than 0.10, indicating that the model produced scores that consistently disagreed with human scores rather than showing scoring uncertainty. The same pattern held for GPT-5.1 for high-scoring essays. Therefore, neither verbalized confidence nor score consistency could flag essays on which the LLM disagreed with the human rater.

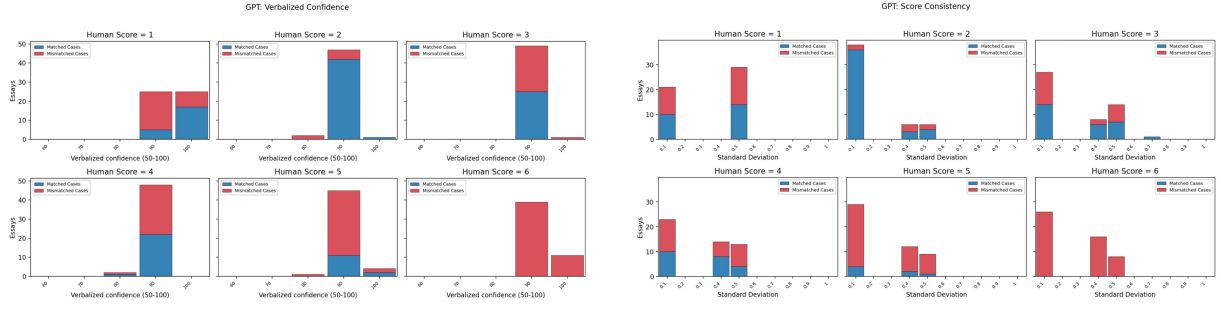


Figure 2: GPT Verbalized Confidence and Score Consistency Distribution Across Human Scores.

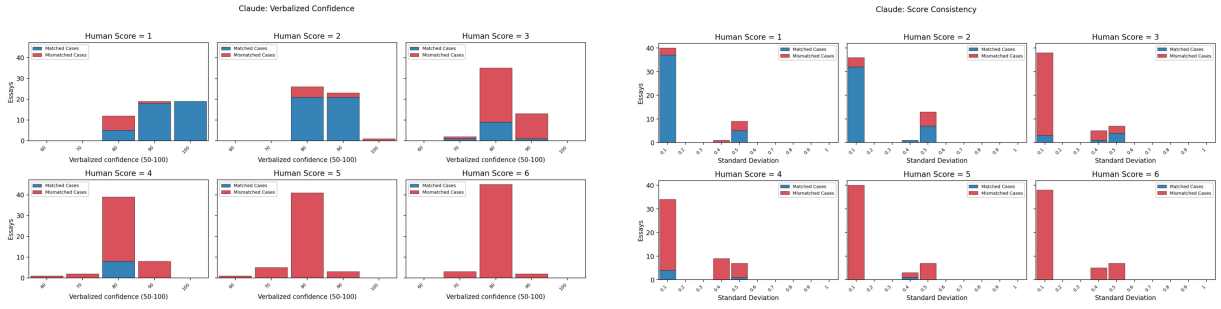


Figure 3: Claude Sonnet 4.6 Verbalized Confidence and Score Consistency Distribution Across Human Scores.

5 Discussion

Inter-rater inconsistency in human scoring has long been a focus of educational measurement. Large language model (LLM)-based automated essay scoring (AES) shares this concern in two ways: (1) an LLM rater can disagree with human raters, and (2) an LLM rater can be inconsistent with itself across repeated runs of the same essay. In this study, we examined human-LLM scoring agreement and uncertainty estimates in two models and investigated whether these patterns differed across rubric score levels.

Our results show that model selection matters for human-LLM scoring agreement. GPT-5.1 achieved satisfactory overall agreement with human raters, whereas Claude Sonnet 4.6 exhibited only moderate agreement. Moreover, score agreement varied by rubric score levels. Both models showed a declining trend in score agreement from low-scoring to high-scoring essays, indicating that LLMs might be more reliable in identifying low-scoring essays. As Mathew et al. (2026) reported that LLMs tended to overrate low-scoring essays, we found this pattern only for GPT-5.1, not for Claude Sonnet 4.6. However, both models tended to underrate high-scoring essays, consistent with previous studies (Korthals et al., 2026; Mathew et al., 2026). Extending prior research, we found that majority voting did not lead to substantial im-

provement in score agreement across the six rubric score levels compared with deterministic scoring. This finding aligns with Xue et al. (2026), suggesting that variation across repeated runs may largely reflect random noise rather than disagreement, thus contributing negligibly to accuracy.

In terms of uncertainty estimates, neither verbalized confidence nor score consistency served as a reliable signal of human-LLM mismatched essays. As the majority of LLM-generated scores were associated with high confidence and high consistency, the results imply that both models may be overconfident and may produce consistently mismatched scores, especially for high-scoring essays.

Limitations and Future Research

While this study examined score agreement and uncertainty estimates across six rubric score levels in two state-of-the-art LLMs, several limitations should be acknowledged. First, because the analysis focused on essays from a single prompt and grade level, the generalizability of the findings to other datasets, prompts, and grade levels may be limited. Second, the study used a specific prompting strategy for both GPT-5.1 and Claude Sonnet 4.6; therefore, the results may differ under alternative prompting strategies. Third, the study used only a temperature setting of 0.7 to introduce variation for the score-consistency method. Future studies could manipulate temperature values to ex-

amine whether score variability reflects meaningful disagreement rather than random noise or systematic bias. Last, this study focused on agreement between human and LLM-generated scores rather than the accuracy of LLM scoring. A high level of human–LLM agreement does not necessarily indicate high scoring accuracy. Future research could investigate approaches to evaluate the accuracy of LLM-generated scores instead of relying on human–LLM scoring agreement.

References

- Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. [Automatic text scoring using neural networks](#). *Preprint*, arXiv:1606.04289.
- Anthropic. 2026. [Claude Sonnet 4.6 \[large language model\]](#).
- Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. [A large-scale corpus for assessing source-based writing quality: ASAP 2.0. Assessing Writing](#), 65:Article 100954.
- Semire Dikli. 2006. [An overview of automated scoring of essays](#). *The Journal of Technology, Learning and Assessment*, 5(1).
- Robinson Ferrer, Damla Turgut, Zhongzhou Chen, and Shashank Sonkar. 2026. [When can we trust LLM graders? Calibrating confidence for automated assessment](#). *Preprint*, arXiv:2603.29559.
- Andrea Gaggioli, Giuseppe Casaburi, Leonardo Ercolani, Francesco Collova, Pietro Torre, and Fabrizio Davide. 2025. [Assessing the reliability and validity of large language models for automated assessment of student essays in higher education](#). *Preprint*, arXiv:2508.02442.
- Yue Huang and Joshua Wilson. 2025. [Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 71–83. National Council on Measurement in Education (NCME).
- Luke Korthals, Emma Akrong, Gali Geller, Hannes Rosenbusch, Raoul Grasman, and Ingmar Visser. 2026. [Towards reliable LLM grading through self-consistency and selective human review: Higher accuracy, less work](#). *Machine Learning and Knowledge Extraction*, 8(3):74.
- Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. [Unleashing large language models’ proficiency in zero-shot essay scoring](#). *Preprint*, arXiv:2404.04941.
- Hongli Li, Che Han Chen, Kevin Fan, Chiho Young-Johnson, Soyoung Lim, and Yali Feng. 2026. [Agreement between large language models and human raters in essay scoring: A research synthesis](#). *Preprint*, arXiv:2512.14561.
- Jerin George Mathew, Sumayya Taher, Anindita Kundu, and Denilson Barbosa. 2026. [LLMs do not grade essays like humans](#). *Preprint*, arXiv:2603.23714.
- OpenAI. 2025. [GPT-5.1 \[large language model\]](#).
- Ellis B. Page. 1966. [The imminence of grading essays by computer](#). *Phi Delta Kappan*, 47(5):238–243.
- Dadi Ramesh and Suresh Kumar Sanampudi. 2022. [An automated essay scoring systems: a systematic literature review](#). *The Artificial Intelligence Review*, 55(3):2495–2527.
- Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. [Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring](#). *Preprint*, arXiv:2008.01441.
- Erik Voss. 2025. [Comparison of traditional machine learning and neural network approaches for automated scoring of second language English essays](#). *Language Testing*, 42(4):369–396.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. [A framework for evaluation and use of automated scoring](#). *Educational Measurement: Issues and Practice*, 31(1):2–13.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs](#). *Preprint*, arXiv:2306.13063.
- Mingfeng Xue, Xingyao Xiao, Yunting Liu, and Mark Wilson. 2026. [On the consistency of automatic scoring with large language models](#). *Educational and Psychological Measurement*, 0(0).
- Kaixun Yang, Mladen Raković, Yuyang Li, Quanlong Guan, Dragan Gašević, and Guanliang Chen. 2024. [Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability](#). *Preprint*, arXiv:2401.05655.

A Prompt

<ROLE>

You are an experienced Grade 10 English teacher. You will evaluate a student argumentative essay about whether

classroom emotion-reading technology is valuable. You will be provided with the [ESSAY STEM], a scoring [RUBRIC], and [EXEMPLARS] with corresponding score levels. Your responsibility is to score the [TARGET ESSAY] according to the provided rubric. You should comprehensively consider every element in the rubric. You must score it on a 1–6 scale and report your confidence in that score on a 0–100 scale.

<ESSAY STEM>

<RUBRIC>

<EXEMPLARS>

<TARGET ESSAY>

You should comprehensively consider every element in the rubric.

Return your reply in this exact JSON format:

```
{
  "reasoning": "<your step-by-step analysis citing the rubric>",
  "score": <integer 1–6>,
  "confidence": <integer 0–100>
}
```