

Two Uses of Human-scored Anchors: Few-shot Direct Scoring and Anchor-based Comparative Grading.

Prashreet Poudel

Western Carolina University
Cullowhee, NC, USA
ppoudel1@catamount.wcu.edu

Huayue Gu

Kennesaw State University
Marietta, GA, USA
hgu2@kennesaw.edu

Collin Lynch

North Carolina State University
Raleigh, NC, USA
cflynch@ncsu.edu

Zhikai Gao

Western Carolina University
Cullowhee, NC, USA
zgao@wcu.edu

Abstract

This study evaluates direct few-shot scoring and anchor based comparative judgment for LLM essay grading. Using human scored anchor essays, we compare their accuracy and stability. Although both approaches produce competitive scores, comparative judgments demonstrate better stability and consistency, suggesting it as a better alternative for grading writing assessment.

1 Introduction

Automatic essay grading has long been studied as a way to support writing assessment, reduce grading workload, and provide students with timely feedback (Weigle, 2002). Traditional approaches often frame essay scoring as a score prediction task: given a student essay, the system predicts a numerical score that approximates human ratings (Weigle, 2002). With the rapid development of large language models, this task can now be approached through prompting rather than through training a task-specific scoring model (Brown et al., 2020). Because LLMs are capable of interpreting natural language, applying rubrics, and reasoning over examples, they offer a promising new direction for automatic essay grading (Mizumoto and Eguchi, 2023).

One common way to use LLMs for essay grading is few-shot direct scoring. In this approach, the model is provided with a rubric, several example essays with known human-assigned scores, and a target essay to be graded (Brown et al., 2020; Yancey et al., 2023; Mansour et al., 2024). The scored examples serve as anchors that demonstrate how the grading scale should be applied. The model is then asked to assign a numerical score to the target

essay. This approach is intuitive because it closely resembles the desired final output: a grade on the same scale used by human raters (Yancey et al., 2023; Mansour et al., 2024).

However, even when scored anchor essays are provided, direct LLM scoring may still face challenges. The model must interpret the rubric, compare the target essay against multiple examples, calibrate its judgment to the scoring scale, and produce a single numerical score within one inference. The final score may be affected by prompt wording, model randomness, the order or selection of examples, and the model's interpretation of score boundaries (Sclar et al., 2024). As a result, it is important to examine not only whether LLM-generated scores are close to human scores, but also whether they are stable across repeated grading conditions (Pack et al., 2024; Stahl et al., 2024).

This study first examines the performance and stability of few-shot direct LLM-based essay grading. In this direct scoring approach, the model is given a rubric, six human-scored anchor essays representing the score levels on the grading scale, and a target essay. The anchor essays and their scores are explicitly exposed to the model, and the model is asked to assign a holistic score to the target essay. This condition serves as the direct grading baseline and allows us to evaluate how accurately and consistently an LLM can produce numerical essay scores when provided with scoring references (Yancey et al., 2023; Mansour et al., 2024; Pack et al., 2024).

To explore an alternative use of the same anchor essays, this study introduces an anchor-based comparative grading approach. Rather than presenting all scored examples at once and asking the model

to generate a score directly, this approach compares the target essay with one human-scored anchor essay at a time. For each comparison, the model is asked to determine which essay is better overall according to the rubric. Each individual judgment is pairwise, but the goal is not to rank all essays against one another. Instead, the goal is to estimate the target essay’s numerical score by locating it relative to human-scored anchors (Liusie et al., 2024). Specifically, we try to answer the following research questions:

- RQ1: How accurate and stable is few-shot LLM essay scoring when using human-scored anchor examples?
- RQ2: Can anchor-based comparative judgments from LLM provide a reliable alternative for estimating essay scores?

By comparing these two anchor-based uses of LLMs, this study contributes to the growing discussion on how LLMs can be designed for educational assessment (Mizumoto and Eguchi, 2023; Pack et al., 2024). Rather than asking only whether LLMs can grade essays, this work examines how the structure of the grading task shapes model performance and stability. The findings may provide guidance for designing LLM-based essay assessment systems that are accurate, stable, interpretable, adaptable and practical for educational use.

1.1 Prior Work

Automated Essay Scoring (AES) has traditionally been approached as a supervised learning problem that maps essays to predefined score categories. Research started with Page’s Project Essay Grade (PEG) in 1960’s. Page used multiple linear regression and simple measurable features like sentence length, character count, and word patterns without analyzing context of essay (Page, 1966). Even though these methods correlated well with human scoring, large-scale deployment was restricted by limited computing power.

Desktop computers became more accessible in the 1990s and early 2000s, and more sophisticated AES systems emerged. Vantage Learning’s IntelliMetric and ETS’s e-rater were the most noticeable ones (Attali and Burstein, 2004, 2006). These systems expanded beyond surface-level features to evaluate grammar, vocabulary, sentence structure, and writing style. AES research also incorporated

semantic analysis techniques such as Latent Semantic Analysis (LSA) to capture essay’s content and meaning (Landauer et al., 1998). Later work explored machine learning and deep learning approaches, including Bayesian models, RNNs, and LSTMs (Taghipour and Ng, 2016; Uto, 2021). Despite these advances, AES systems continued to face limitations, as many models were heavily relied on measurable writing features such as essay length, sentence complexity, and vocabulary usage. Because of that students could sometimes achieve high scores by writing essays that appeared sophisticated but contained weak reasoning or little meaningful content (Attali and Burstein, 2006). In addition, these systems required large collections of human-scored essays for each new prompt, making them costly to develop and difficult to adapt to new writing tasks (Ridley et al., 2020, 2021).

The development of LLMs has allowed AES to move beyond feature-engineered approaches that require extensive training data (Brown et al., 2020). By integrating detailed rubrics into prompts, LLMs can evaluate essay content directly, assign scores, explain their reasoning, and provide feedback, offering greater transparency than many traditional AES systems (Mizumoto and Eguchi, 2023; Stahl et al., 2024). However, LLM-based scoring also introduces new challenges, results vary depending on prompt wording, rubric presentation, and model settings, which raises concerns about scoring consistency and reliability (Sclar et al., 2024; Stahl et al., 2024). To overcome these issues, pairwise comparison of essays were also used for evaluating essay scores (Shibata and Miyamura, 2025). These pairwise judgments are assigned into rankings that can be converted into final grades. This reduces scoring instability while capitalizing the model’s ability to evaluate writing quality (Shibata and Miyamura, 2025). Comparative approaches may also reduce biases found in direct scoring, such as the tendency to favor longer essays simply because they contain more text (Zheng et al., 2023).

2 Dataset

In this study, we used a public dataset called PERSUADE 2.0 corpus (Crossley et al., 2024), which contains over 25,000 argumentative essays produced by 6th-12th grade students in the United States for 15 prompts on two writing tasks: independent and source-based writing. In this study, due to the limitation on our computing resources

and time, we concentrate only on three different prompts (Phones and Driving, Summer Projects, and Digital Learning) from the independent essays (n = 5075).

3 Methods

3.1 LLM Inference and Prompt Construction

Using Qwen 3.6 35B model (Qwen Team, 2026), We evaluated two LLM-based approaches for essay scoring: few-shot direct scoring and anchor-based comparative grading. Both approaches used the same grading rubric and the same set of human-scored anchor essays. The key difference was how the anchor essays were presented to the model and how the model was asked to use them. In the few-shot direct scoring condition, the model received all anchor essays and their human-assigned scores in a single prompt and was asked to assign a numerical score to the target essay. In the anchor-based comparative condition, the model received one target essay and one anchor essay at a time and was asked to determine which essay was better overall according to the rubric. The complete workflow is shown in Figure 1

For both approaches, the system prompt established the role of the model as a K-12 teacher grading student essays on the given writing task. The prompt included the assignment topic and the grading rubric, which defined the criteria used to evaluate essay quality. This ensured that both scoring approaches were grounded in the same assessment criteria.

In the few-shot direct scoring condition, the prompt included six human-scored anchor essays, one from each score level on the holistic scoring scale. Each anchor essay was presented together with its corresponding human-assigned score. After reviewing these examples, the model was given a target essay and asked to assign a holistic score on the same scale. This prompt design was intended to provide the model with explicit examples of how the scoring scale should be applied. The model’s output was parsed as the predicted essay score for the target essay.

In the anchor-based comparative grading condition, the model was not asked to directly assign a score. Instead, each target essay was compared with six anchor essays, one from each score level. For each comparison, the target essay was presented as Essay A and the anchor essay was presented as Essay B. The model was asked to

compare the two essays using the rubric and select which essay was better overall. The prompt instructed the model to return only a structured JSON object indicating whether Essay A or Essay B was better. This constrained output format was used to reduce parsing ambiguity and ensure that each inference produced a binary comparative judgment.

To support a fair comparison between the two approaches, the direct scoring and comparative grading conditions used the same pool of human-scored anchor essays. Thus, the experimental contrast was not whether the model had access to scoring examples, but whether the model used those examples through direct score prediction or through a sequence of comparative judgments. This design allowed us to examine whether changing the form of LLM inference, while holding the rubric and anchor information constant, affected the accuracy and stability of essay score estimation.

3.2 Bayesian Aggregation Algorithm for Comparative Grading

To convert the LLM-generated comparisons results into the actual essay scores, we developed a Bayesian aggregation procedure. The goal of this procedure is to infer the most likely holistic score of a target essay based on its comparison outcomes against human-score anchor essays. For each target essay, the LLM compared the essay with six human-scored anchor essay, one from each score level. Let $k \in 1, 2, 3, 4, 5, 6$ denote the score level of the anchor essay. For each comparison, the output from LLM is a threshold-style observation. If the LLM selected the target essay as better than the scored-k anchor, the observation is interpreted as evidence that the target essay’s score was at least k. If the LLM selected the anchor essay as better, the observation was interpreted as evidence that the target essay’s score was at most (k). Formally, each comparison produced an observation (O_k), where:

$$O_k = \begin{cases} \text{score} \geq k, & \text{if target essay is preferred} \\ \text{score} \leq k, & \text{if anchor is preferred} \end{cases}$$

Given the six observations ($O_1, \dots, O_6$), we estimated the posterior probability of each possible score ($i \in 1, 2, \dots, 6$). Using a naïve Bayes assumption, the comparison observations were treated as conditionally independent given the true score. The posterior probability was computed as:

$$P(\text{score} = i \mid O_1, \dots, O_6) \propto P(\text{score} = i) \prod_{k=1}^6 P(O_k \mid \text{score} = i)$$

Where $(P(\text{score} = i))$ represents the prior probability of score (i), and $(P(O_k | \text{score} = i))$ represents the likelihood of observing the comparison outcome against the score-(k) anchor if the true score of the target essay were (i).

The likelihood term $(P(O_k | \text{score} = i))$ was determined by whether the observation was consistent with the hypothesized score (i). If the observation was $(\text{score} \geq k)$, then it was considered consistent when $(i \geq k)$. If the observation was $(\text{score} \leq k)$, then it was considered consistent when $(i \leq k)$. Importantly, when $(i=k)$, both directions were treated as consistent. This reflects the fact that two essays with the same holistic score may differ in local qualities, and a forced-choice comparison may reasonably prefer either essay while still being compatible with the same score level.

For each anchor score level (k), we introduced a reliability parameter (p_k), representing the probability that the LLM’s comparison against a score-(k) anchor produces a threshold observation consistent with the target essay’s human score. The likelihood was therefore defined as:

$$P(O_k | \text{score} = i) = \begin{cases} p_k, & \text{if } O_k \text{ is consistent with } \text{score} = i \\ 1 - p_k, & \text{if } O_k \text{ is inconsistent with } \text{score} = i \end{cases}$$

The reliability parameters (p_k) were estimated using a validation set of 200 essays with known human scores. Each validation essay was compared with one anchor essay from each of the six score levels, producing 1200 validation comparisons in total. For each anchor level (k), we counted the number of comparisons (C_k) whose threshold interpretation was consistent with the validation essay’s human score. Since each validation essay was compared once with a score-(k) anchor, ($N_k = 100$) comparisons were used to estimate each (p_k). To avoid reliability estimates of exactly 0 or 1, we applied Laplace smoothing (Kikuchi et al., 2015):

$$p_k = \frac{C_k + 1}{N_k + 2}$$

This smoothing procedure adds one pseudo-correct and one pseudo-incorrect observation to each anchor level, producing a more conservative estimate of the LLM’s reliability for that comparison level.

After estimating the reliability parameters from the validation set, we applied the Bayesian aggregation procedure to the test essays. For each test essay, the six LLM comparison outcomes were converted

into threshold observations and used to compute the posterior probability for each possible score. After computing the posterior probability for each possible score, we estimated the final essay score using the posterior expected value rather than selecting the single most probable score. Specifically, the predicted score was computed as:

$$\hat{s} = \sum_{i=1}^6 iP(\text{score} = i | O_1, O_2, \dots, O_6)$$

This formulation uses the full posterior distribution over score levels and preserves uncertainty across adjacent scores. Because essay scoring is ordinal and neighboring score levels often reflect subtle differences in writing quality, the posterior probability may be distributed across multiple plausible scores. The expected-score formulation captures this uncertainty by producing a continuous score estimate. To obtain a final discrete score on the original 1–6 scale, we rounded ($\hat{s}$) to the nearest valid score level.

3.3 Evaluation Methodology

To evaluate the performance of automated essay scoring system we used two different evaluation frameworks: accuracy based evaluation and stability based evaluation. As the frameworks are complimentary, these provide comprehensive picture of both how well the model predicts the score and how consistent it is across the repeated runs.

Accuracy was accessed using six metrics. Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual score. R^2 captures the proportion of variance in actual score explained by the model’s prediction. Quadratic Weighted Kappa (QWK) and Cohen’s Kappa both are statistical metrics used to measure the agreement between two raters or classification models, designed to account for agreement that happens purely by chance. However, QWK it penalizes the larger mistakes heavily due to its quadratic weighing (Cohen, 1968). Exact Match Accuracy reports the percentage of prediction where the actual score is the same as the predicted score, while Adjacent Match Accuracy is more forgiving on the criterion to include the predictions within one point of true score.

Stability across multiple runs was also evaluated using six measures. Perfect Consistency Rate (PCR) is the proportion of essays receiving identical score across all runs, and Acceptable Consistency Rate (ACR) is more flexible to prediction

differing by no more than one points. Average Standard Deviation (Avg SD) measures the mean spread of predicted score around each essay standard deviation. Worst Case Difference catches the largest observed score deviation for any single essay across runs. Mean Absolute Deviation (MAD) is the average absolute deviation of individual predictions from each essay’s median predicted score. Finally, Pairwise Pooled QWK looks at every pair of scores given to the same essay, treats them like they came from two different raters, and then calculates how consistent those ratings are. This is similar to measuring consistency between two human raters(Cohen, 1960).

4 Results

We evaluated the two LLM prompting architectures for AES and the results are stated in Table 2 and Table 1.

4.1 Accuracy and Stability RQ1 & RQ2

The anchor-based comparative framework was consistently better than direct prompting in ordinal agreement and variance explanation. On the Summer Project prompt, direct grading showed an R^2 of 0.1020 and a QWK of 0.5880, while the comparative method had an R^2 of 0.2701 and a QWK of 0.7157. Similarly, on Distance Learning, direct scoring achieved an R^2 of 0.3069 and a QWK of 0.6640, compared to the comparative model’s 0.4731 and 0.7704. These gains show that pairwise comparisons yield superior rubric understanding.

However, Exact Match Accuracy (EMA) shows a contradiction. Direct scoring achieved higher exact matches on all three prompts tested: Phones and Driving (47.22% vs. 39.70%), Summer Project (43.20% vs. 32.42%), and Distance Learning (43.71% vs. 40.22%). However, when direct scoring missed, it tended to miss by a larger margin. The comparative framework achieved adjacent precision (within ± 1 point) of 97.42%, 98.44%, and 98.78% across the three prompts, while direct scoring ranged from 89.99% to 93.04%. Even the lowest adjacent-match accuracy of the comparative framework (97.42%) exceeded the highest achieved by direct scoring (93.04%).

It is important to note that the LLM demonstrated very high accuracy in pairwise comparisons, achieving an accuracy of above 90% for all score levels. Hence, the comparative method still requires a better model to convert these highly accu-

rate binary comparisons into final predicted scores. The current Bayesian aggregation framework does not fully preserve the LLM’s native discriminative strength.

4.2 Stability RQ2

Comparative grading was also more operationally stable. On the Summer Project prompt, the PCR of direct scoring was 28.46% with a MAD of 0.2283 and an Avg SD of 0.3740. With comparative grading, PCR rose to 67.03%, MAD decreased to 0.0837, and Avg SD decreased to 0.1648. Pairwise Pooled QWK increased from 0.7990 to 0.9129, reflecting substantially stronger agreement across repeated grading runs. On Distance Learning, comparative grading achieved a PCR of 65.32%, a MAD of 0.0838, an Avg SD of 0.1670, and a Pairwise Pooled QWK of 0.9324, compared to 35.65%, 0.1922, 0.3255, and 0.8462 for direct scoring.

The downside is that direct scoring limited the maximum observed error to 2 points across all prompts. In contrast, comparative grading produced occasional outlier scores, with worst-case differences of 4 points on Phones and Driving and 3 points on both Summer Project and Distance Learning. These outliers appear to result from weaknesses within the Bayesian aggregation algorithm rather than the pairwise comparisons themselves.

4.3 Discussion

Few-Shot Direct Grading is a conservative estimator that improves exact-match accuracy and limits maximum error to 2 points, but it does not approximate the distribution of true scores particularly well (low R^2 values) and displays higher volatility across repeated evaluations. Anchor-Based Comparative Grading takes advantage of the LLM’s robust discriminative ability. Still, despite the LLM’s exceptional pairwise accuracy, the Bayesian aggregation framework remains a limitation in this setting. A superior mathematical model is needed to transform the LLM’s accurate comparison judgments into scores without introducing error propagation.

Nevertheless, comparative grading yields strong overall performance, with an R^2 as high as 0.4731, near-perfect adjacent-match accuracy (up to 98.78%), and high consistency (MAD < 0.09). Unless limiting worst-case error to 2 points is the primary objective, comparative grading demonstrates better overall stability and score calibration, which may be more valuable for large-scale K–12

essay assessment.

5 Limitations and Future Work

This study was not perfect and certainly has some limitations. Starting out, we used one fixed LLM model to compare and grade essays. Other stable models, like Llama or deepseek, might perform differently and adaptive service models like GPT-4 or Claude, may vary day by day. This in turn may limit the generality of these results.

Following that, the Bayesian aggregation model is not the best choice to find the final grade for comparative essays. The LLM model was accurate on identifying the better essay when compared throughout the score range. However, the Bayesian method did not preserve this accuracy when converting comparisons into final scores. It introduced errors and occasional outliers (up to 4 points off) and Exact-Match accuracy of just 37.44%. The Bayesian approach is not the best way to aggregate comparative judgments.

To improve limitations, our future goal is to make results generalized for different LLMs as well as incorporating better aggregation models. We want to capitalize on the high accuracy demonstrated by the LLM on identifying which essay is better and correctly predict the actual score of an essay.

6 Conclusion

This study looked at two ways to use automated systems to score essays. Firstly, directly score them, and secondly to compare them to other essays. The study used three prompts from the PERSUADE 2.0 corpus to see which way worked better. The comparative method worked better overall: it better matched human graders, more accurately ranked essays by quality, and remained more stable when the same essay was scored multiple times. These results challenge the traditional approach of scoring essays individually. Instead, comparing each essay against others proves to be a more effective strategy that leverages the natural strengths of automated grading systems.

However, we could see a contradiction on Exact-Match Accuracy. Direct scoring achieved higher exact match accuracy with human raters. Conversely, the comparative method excelled at avoiding large errors and exhibited substantially greater operational stability across multiple runs. It produced more consistent scores, lower variance, and

fewer wild fluctuations. While direct scoring may occasionally land on the correct score, the comparative approach is more reliable overall and serves as a safer, more dependable strategy for automated essay grading. Despite the automated system's strong performance in pairwise comparisons, a limitation emerged. The Bayesian aggregation model used to convert these comparisons into final scores proved inadequate. It failed to fully leverage the system's natural discriminative strength, effectively becoming a bottleneck rather than an asset which explains the lower accuracy on Exact Match Accuracy. Improving the mathematical calculation model is one of our future goals.

For educational institutions seeking to deploy automated essay scoring, the comparative approach remains the recommended strategy. It is more stable, more consistent across repeated trials, and better at preventing large grading errors. Unless a school requires that predicted scores never deviate from human ratings by more than two points, the comparative method is the superior choice. Future work should focus on developing better mathematical models to replace the Bayesian aggregator. Improving this post-processing step will allow comparison based automated systems to more fully realize their potential for accurate and reliable essay grading.

References

- Yigal Attali and Jill Burstein. 2004. [Automated essay scoring with e-rater® v.2.0](#). *ETS Research Report Series*, 2004(2):i–21.
- Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment*, 4(3):1–30.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Jacob Cohen. 1968. [Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit](#). *Psychological Bulletin*, 70(4):213–220.

- Scott A Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Meg Benner, and Ulrich Boser. 2024. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61:100865.
- Masato Kikuchi, Mitsuo Yoshida, Masayuki Okabe, and Kyoji Umemura. 2015. Confidence interval of probability estimator of laplace smoothing. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE.
- Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. [An introduction to latent semantic analysis](#). *Discourse Processes*, 25(2–3):259–284.
- Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–151. Association for Computational Linguistics.
- Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786. ELRA and ICCL.
- Atsushi Mizumoto and Masaki Eguchi. 2023. [Exploring the potential of using an ai language model for automated essay scoring](#). *Research Methods in Applied Linguistics*, 2(2):100050.
- Austin Pack, Alex Barrett, and Juan Escalante. 2024. [Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability](#). *Computers and Education: Artificial Intelligence*, 6:100234.
- Ellis B. Page. 1966. The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.
- Qwen Team. 2026. Qwen3.6: Large language model series. <https://github.com/QwenLM/Qwen3.6>.
- 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. *arXiv preprint arXiv:2008.01441*.
- Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. [Automated cross-prompt scoring of essay traits](#). In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, volume 35, pages 13745–13753.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.
- Takumi Shibata and Yuichi Miyamura. 2025. LCES: Zero-shot automated essay scoring via pairwise comparisons using large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29976–29989, Suzhou, China. Association for Computational Linguistics.
- Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298. Association for Computational Linguistics.
- Kaveh Taghipour and Hwee Tou Ng. 2016. [A neural approach to automated essay scoring](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.
- Masaki Uto. 2021. [A review of deep-neural automated essay scoring models](#). *Behaviormetrika*, 48(2):459–484.
- Sara Cushing Weigle. 2002. *Assessing Writing*. Cambridge University Press, Cambridge, UK.
- Kevin P. Yancey, Geoffrey Laffair, Anthony Verardi, and Jill Burstein. 2023. Rating short 12 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 576–584. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Figures and Tables

	Accuracy Measurement					
	MAE	R^2	QWK	Cohen's Kappa	Exact-Match Acc	Adjacent-Match Acc
Phones & Driving	0.5987	0.0327	0.5280	0.2309	47.22%	93.04%
Summer Project	0.6736	0.1020	0.5880	0.2382	43.20%	89.99%
Distance Learning	0.6496	0.3069	0.6640	0.2579	43.71%	91.78%

	Stability Measurement					
	PCR	ACR	Avg SD	Worst-case Diff	MAD	Pairwise Pooled QWK
Phones & Driving	27.23%	93.41%	0.3731	2	0.2247	0.7238
Summer Project	28.46%	94.23%	0.3740	2	0.2283	0.7990
Distance Learning	35.65%	97.03%	0.3255	2	0.1922	0.8462

Table 1: Evaluation results for few-shot direct grading. PCR: Perfect Consistency Rate (same exact score every run); ACR: Acceptable Consistency Rate (score variance within ± 1 point).

	Accuracy Measurement					
	MAE	R^2	QWK	Cohen's Kappa	Exact-Match Acc	Adjacent-Match Acc
Phones & Driving	0.6300	0.1017	0.5825	0.1365	39.70%	97.42%
Summer Project	0.6917	0.2701	0.7157	0.1521	32.42%	98.44%
Distance Learning	0.6100	0.4731	0.7704	0.2415	40.22%	98.78%

	Stability Measurement					
	PCR	ACR	Avg SD	Worst-case Diff	MAD	Pairwise Pooled QWK
Phones & Driving	33.39%	80.82%	0.3722	4	0.1934	0.5931
Summer Project	67.03%	95.43%	0.1648	3	0.0837	0.9129
Distance Learning	65.32%	97.82%	0.1670	3	0.0838	0.9324

Table 2: Evaluation results for anchor-based comparative grading. PCR: Perfect Consistency Rate (same exact score every run); ACR: Acceptable Consistency Rate (score variance within ± 1 point).

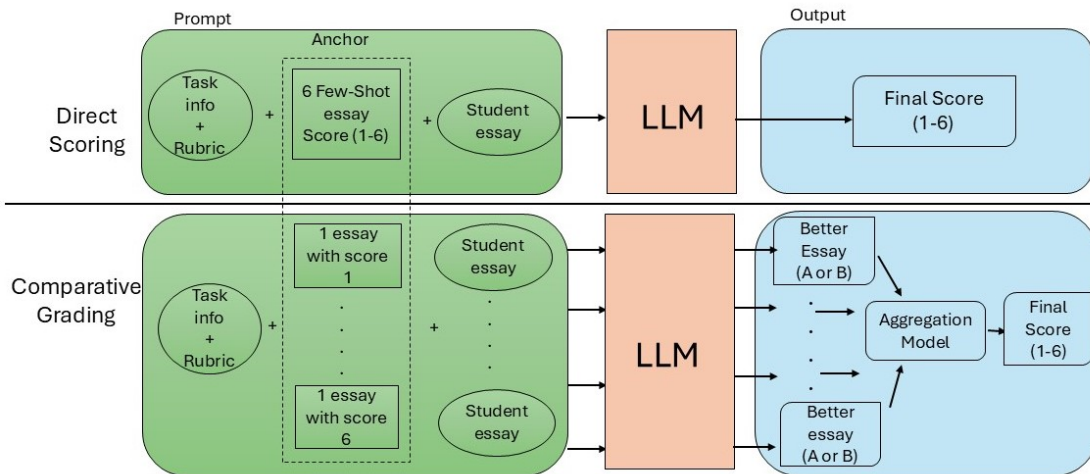


Figure 1: Two anchor-based LLM grading methods: direct scoring predicts in one step using all anchors, while the comparative method uses pairwise comparisons and Bayesian aggregation