Teach-to-Reason with Scoring: Self-Explainable Rationale-Driven Multi-Trait Essay Scoring

Anonymous ACL submission

Abstract

Multi-trait automated essay scoring (AES) systems provide a fine-grained evaluation of an essay's diverse aspects. While they excel in scoring, prior systems fail to explain why specific trait scores are assigned. This lack of transparency leaves instructors and learners unconvinced of the AES outputs, hindering their practical use. To address this, we propose a self-explainable Rationale-Driven Multi-trait automated Essay scoring (RaDME)[1] framework. RaDME leverages the reasoning capabilities of large language models (LLMs) by distilling them into a smaller yet effective scorer. This more manageable *student* model is optimized to sequentially generate a trait score followed by the corresponding rationale, thereby inherently learning to select a more justifiable score by considering the subsequent rationale during training. Our findings indicate that while LLMs underperform in direct AES tasks, they excel in rationale generation when provided with precise numerical scores. Thus, RaDME integrates the superior reasoning capacities of LLMs into the robust scoring accuracy of an optimized smaller model. Extensive experiments demonstrate that RaDME achieves both accurate and adequate reasoning while supporting high-quality multi-trait scoring, significantly enhancing the transparency of AES.

1 Introduction

Fine-grained feedback, grounded in an accurate assessment of writing quality, is crucial for enhancing learners' writing skills. While traditional holistic automated essay scoring (AES) models (Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017; Wang et al., 2022) provide only an overall score, recent research has shifted toward multi-trait scoring (Kumar et al., 2022; Do et al., 2024a,b) to enable a more granular evaluation of essays. With

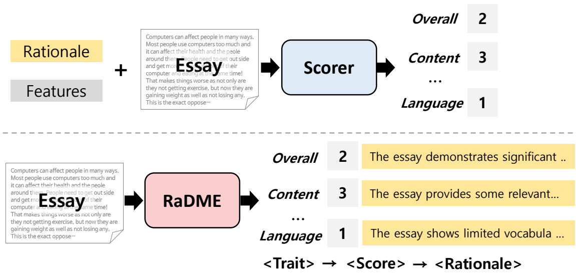

Figure 1: Comparison of existing multi-trait scoring methods (top) and RaDME (bottom). Existing methods take features or rationales as input, not allowing direct interpretation of the results; however, RaDME explicitly derives scores followed by its rationales, enhancing the reliability of the outcomes.

the introduction of the autoregressive score generation framework, ArTS (Do et al., 2024a), multi-trait scoring has made remarkable strides, achieving substantial agreement with human-expert ratings.

Despite advancements in AES, current systems remain opaque, as they fail to explain the rationale behind their scoring decisions. While these models deliver accurate score predictions, their lack of interpretability undermines the transparency and reliability of assessments (Kumar and Boulanger, 2020; Johnson and Zhang, 2024). Thus, educators and students, who require more than just numerical feedback, often find these outputs unconvincing, restricting the practical deployment of AES.

To interpret the model decisions, prior studies have attempted to derive scoring decisions by leveraging explicit grammatical or linguistic features (Wang and Hu, 2021; Sudoh et al., 2024). However, these approaches focus on model-driven explanations rather than providing human-centered justifications for assigned scores. More recently, Chu et al. (2024) utilized rubric guidelines to prompt large language models (LLMs) to generate evaluation rationales, which were then used as additional input for an ArTS-based (Do et al., 2024a) model.

[1]Codes and all generated results will be publicly available.

While they integrate rationales, their primary goal is to improve AES performance rather than enhance explainability, leaving the scoring model lacking an inherent mechanism to clarify the reasoning behind derived scores (Figure 1).

To address these, we propose a self-explainable, rationale-driven multi-trait essay scoring (RaDME) method, which *learns to reason with scoring*. Drawing inspiration from the human decision-making process, e.g., decide-with-reason, RaDME is designed to jointly generate a score and its corresponding rationale, ensuring each scoring decision is inherently grounded in clear, justifiable reasoning. To achieve this goal, we distill the reasoning capacity of LLMs into a smaller yet effective scoring model. Notably, while LLMs have struggled to achieve precise AES performance even with iterative or sophisticated prompting (Mizumoto and Eguchi, 2023; Mansour et al., 2024a; Lee et al., 2024), we find that they excel at reasoning, particularly when provided with explicit numeric scores; this also aligns with existing research (Huang and Chang, 2023; Ryu et al., 2024). In contrast, smaller domain-specific expert models excel in scoring but lack reasoning capabilities. RaDME bridges this gap by introducing rationale distillation that maximizes both advantages, allowing for effective rationale-driven assessment. Note that we construct a unified model capable of both reasoning and scoring across multiple traits and prompts by optimizing the model with trait-wise rationale-score pairs as a multi-task learning approach.

Extensive experimental results demonstrate that RaDME achieves outstanding scoring performance, even surpassing recent state-of-the-art methods while simultaneously generating high-quality rationales. This result is particularly noteworthy, as previous attempts to jointly perform feedback generation and scoring have largely failed to achieve reliable scoring (Stahl et al., 2024). Further discussions and analyses on both scoring and rationale generation results strongly support RaDME's ability to enhance both reasoning capabilities and scoring quality. Our findings include that RaDME with scoring-first and subsequent reasoning notably enhances both generations. Our work takes a crucial step toward enhancing the transparency of automated evaluation, laying the foundation for more interpretable and reliable AES. Our contributions can be summarized as follows:

- We propose RaDME, a self-explainable, rationale-driven multi-trait AES that explicitly outputs reasoning with scoring, ensuring both interpretability and scoring accuracy.
- By providing LLMs with explicit numeric trait scores, we extract clear, coherent, and well-structured rationales, effectively supporting the distilled student model in producing high-quality explanations.
- RaDME achieves efficient and scalable AES by distilling only the reasoning capabilities of scoring-inferior LLMs, enabling a lightweight model suitable for self-explaining and scoring in real-world deployment.
- Our findings highlight the efficacy of rationale-driven scoring, revealing that scoring-first modeling notably enhances both scoring consistency and explanation quality.

2 Related Works

LLMs for AES. As LLMs continue to exhibit exceptional performance across diverse domains, their application to AES via zero-shot or few-shot prompting has garnered increasing attention (Mizumoto and Eguchi, 2023; Mansour et al., 2024b; Lee et al., 2024). However, these approaches often underperform compared to fine-tuned, domain-specific models. Lee et al. (2024) propose an iterative method in which an LLM first generates scoring criteria for multiple traits, then extracts textual evidence for each trait, and finally assigns trait scores that are aggregated into a holistic score. Despite its tailored design, the method underperforms compared to existing AES models (Xie et al., 2022) and incurs a high computational cost due to repeated prompting. Thus, instead of using LLMs directly for scoring, which proves both costly and less accurate, we propose leveraging their reasoning strengths and distilling these into a more efficient and effective scoring model.

Explainability in AES. To apply AI-based automated evaluation systems in real-world educational settings, transparency and explainability are crucial (Johnson and Zhang, 2024). While previous holistic AES studies have leveraged explicit grammatical or linguistic features to derive scoring outcomes (Wang and Hu, 2021; Sudoh et al., 2024), their focus on model-based explainability often does not adequately support human understanding of the model decisions. Stahl et al. (2024) explored the

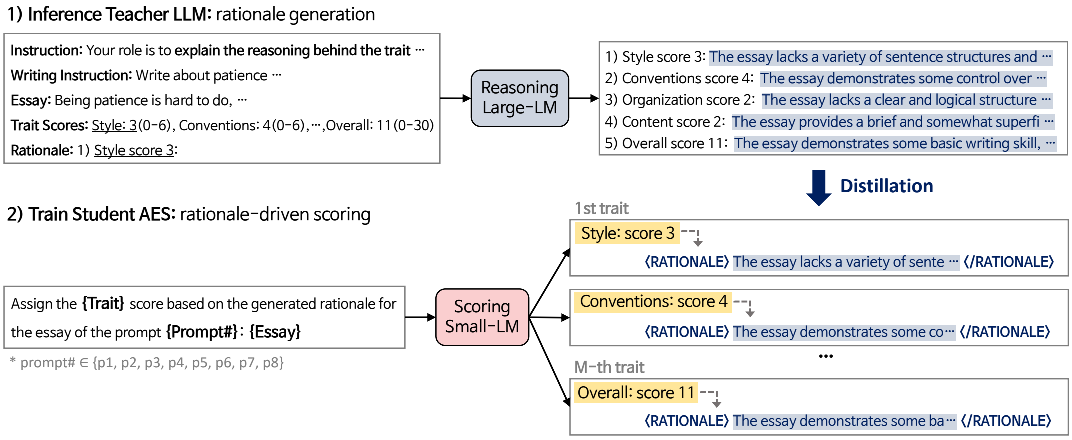

Figure 2: An overview of the RaDME framework.

use of LLMs for jointly providing feedback and AES, showing robust feedback quality but remarkably low scoring performance.

Recent research has shifted toward multi-trait scoring for fine-grained essay evaluation and intuitive feedback. ArTS (Do et al., 2024a) leverages trait dependencies within a text-to-text framework, achieving strong performance in multi-trait scoring. Chu et al. (2024) further improved scoring accuracy by incorporating LLM-generated rationales as additional input to ArTS-based models. However, these rationales were used solely as input features to improve scoring performance, not to explain the model's decisions, leaving it unable to articulate the reasoning behind its scores. Moreover, they prompted LLMs separately for each trait, requiring iterative prompting, and tasked the LLM with generating both rationales and the corresponding scores, an approach that is potentially unreliable (Mizumoto and Eguchi, 2023; Lee et al., 2024).

Independent of our work, Mohammadkhani (2024) presents a preliminary attempt to distill LLM reasoning into smaller models for interpretability, focusing on three traits and aiming to improve scoring performance. However, they do not consider the rationale–score relationship or assess rationale quality. In contrast, we thoroughly investigate this relationship across 11 traits and identify optimal learning orders, which enable smaller models to generate high-quality rationales comparable to those of the teacher LLM, thereby ensuring both transparency and reliability in scoring.

3 RaDME

In this work, we distinguish between using *explainability* to provide feedback for behavioral changes (i.e., how to revise writing) and providing justifications for scores given the essay (i.e., explaining why the essay received a certain score). Our work focuses on the latter, aiming to enhance the transparency and trustworthiness of scoring decisions.

In human decision-making, responses are often guided by implicit rationales derived from contextual understanding. Inspired by this, we propose a self-explainable Rationale-Driven Multi-trait automated Essay scoring (RaDME) to incorporate this reasoning process by training a student model that predicts both the ground-truth trait score and its corresponding rationale, which is distilled from teacher LLM (Figure 2). We hypothesize that this approach encourages the model to produce scores that are more likely to be accompanied by justifications, enhancing its ability to generate a score and its well-matched explanation. RaDME is implemented within a knowledge distillation framework, leveraging LLM as the teacher and a smaller language model as the student. The student RaDME is trained to first generate a trait score and then a corresponding rationale to justify that score. It internally guides the model to assign scores grounded in reasoning, rather than generating independent explanations, effectively leveraging LLM-derived rationales while maintaining resource efficiency.

3.1 Extracting rationale from teacher LLMs

To fully leverage the reasoning ability of teacher LLMs, we introduce a *score-guided prompting* strategy, which explicitly informs the model of the trait-specific scores assigned to each essay. As LLMs struggle with direct numerical scoring but excel at generating explanations (Ryu et al., 2024), we explicitly provide them with exact scores and

LLM Prompt	Generation Output
Your role is to explain the reasoning behind the trait scores assigned to the {narrative} type essays. Please review the essay and its trait scores, and provide a one-sentence rationale for each trait score. Words beginning with "@" have been intentionally replaced to anonymize personally identifying information in the essays. Writing Instructions: {Write about patience. Being patient means that you are understanding and tolerant. A patient person experience difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.} Essay: {Being patience is hard to do, @CAPS1 I do remember a time when I was patient. This was at @CITY1. Me and my group had to wait for everyone to get back to the bus. We were the first one's to the bus so it took a while before they came. @CAPS1 before they did I was as calm and patient as can be.} Trait Scores: {Style: 3 (0-6), Conventions: 4 (0-6), Organization: 2 (0-6), Content: 2 (0-6), Overall: 11 (0-30)} Rationale: {1) Style score 3: }	Rationale: 1) Style score 3: The essay lacks a variety of sentence structures and vocabulary, with the writer relying on simple sentences and basic words, which detracts from the overall style and tone of the essay. 2) Conventions score 4: The essay demonstrates some control over grammar, spelling, and punctuation, but there are noticeable errors, such as "Being patience" and "Me and my group," which affect the clarity and correctness of the writing. 3) Organization score 2: The essay lacks a clear and logical structure, with the writer jumping abruptly from stating the difficulty of being patient to recalling a specific experience, and failing to provide a clear conclusion or connection to the initial statement. 4) Content score 2: The essay provides a brief and somewhat superficial anecdote, but fails to fully develop the concept of patience, provide meaningful insights, or explore the significance of the experience, resulting in a lack of depth and substance. 5) Overall score 11: The essay demonstrates some basic writing skills, but is hindered by weaknesses in organization, content, and style, which limit its overall effectiveness in conveying the writer's message about patience.

Table 1: The used prompt for teacher LLM and its corresponding generation output. Notably, even without enforcing a structured format, simply indicating the first trait and its score ensured a consistent output format across all samples while effectively generating rationales for each trait score.

focus them solely on generating precise rationales. Our numeric guidance strengthens the coherence between ground-truth scores and obtained rationales, ensuring alignment with actual grading patterns. This approach can effectively support the student scorer in learning to generate more relevant and well-grounded explanations.

In particular, we prompt the LLM by including the following input components: *Instruction*, which defines general roles with broad conceptual guidelines; *Writing Instruction*, corresponding to the essay-writing prompt that specifies the topic and theme of the learner's essay; *Essay* (E), which is the learner's written submission; and *Trait Scores* ($S = \{s_t \mid t \in T\}$), comprising human-annotated trait-score (range) pairs. Given these elements, the model generates a set of rationales ($R = \{r_t \mid t \in T\}$) corresponding to the assigned trait scores. To ensure that the model generates responses in the fixed format *"N) {Trait} score {Score}: {Rationale},"* we inform a sample format for the first trait, such as *"1) Style score 3:"*. The detailed example of the used prompt and the corresponding output is described in Table 1.

3.2 Distillation for rationale-driven scoring

When making decisions, humans naturally rely on implicit reasoning shaped by their understanding of the surrounding context. Motivated by this, we train the student model to predict both a ground-truth trait t-th score (s_t) and its corresponding rationale r_t, which is distilled from the teacher LLM. Particularly, the teacher-generated multi-trait rationales R are separated by trait and employed to train a specialized scoring model optimized for multi-trait scoring with reasoning. Note that RaDME does not rely on LLMs at inference time, making it significantly more efficient and scalable for real-world deployment.

We design a unified model capable of predicting all trait scores across different prompts, leveraging an autoregressive score prediction method (Do et al., 2024a). However, distinct from their approach, which predicts all trait scores in a single sequence since it only generates trait scores, our model also produces long-form rationales. Therefore, we predict each trait independently in separate sequences to ensure stability.

When handling essays from multiple prompts within a single model, incorporating prompt guidance in the prefix has been shown to enhance scoring accuracy, as demonstrated in ArTS (Do et al., 2024a). Building on this, since our model predicts each trait in a separate sequence, we further incorporate trait name guidance alongside prompt guidance, ensuring that the model effectively differentiates between scoring criteria, leading to more consistent and reliable predictions.

Building on this, since our model predicts each trait independently in separate sequences, we further incorporate trait name guidance alongside prompt guidance. To predict the t-th trait score, the input comprises the essay E, trait name t, and

Pr	Traits	Es	Score Range (Overall / Trait)
1	Over, Cont, Org, WC, SF, Conv	1,783	2 - 12 / 1 - 6
2	Over, Cont, Org, WC, SF, Conv	1,800	1 - 6 / 1 - 6
3	Over, Cont, PA, Nar, Lang	1,726	0 - 3 / 0 - 3
4	Over, Cont, PA, Nar, Lang	1,772	0 - 3 / 0 - 3
5	Over, Cont, PA, Nar, Lang	1,805	0 - 4 / 0 - 4
6	Over, Cont, PA, Nar, Lang	1,800	0 - 4 / 0 - 4
7	Over, Cont, Org, Conv, Style	1,569	0 - 30 / 0 - 6
8	Over, Cont, Org, WC, SF, Conv, Voice	723	0 - 60 / 2 - 12

Table 2: Summarized statistics of the ASAP/ASAP++ dataset. Pr: prompt number, Es: the number of essays; Over: *Overall*, Cont: *Content*, Org: *Organization*, WC: *Word Choice*, SF: *Sentence Fluency*, Conv: *Conventions*, PA: *Prompt Adherence*, Nar: *Narrativity*, Lang: *Language*, Style: *Style*, Voice: *Voice*.

prompt number p, formatted as follows: *"Assign the {t} score based on the generated rationale for the essay of the prompt {p}: "*. The model then generates a trait name (t), a predicted score ($\hat{s}_t$), and a predicted rationale ($\hat{r}_t$), following the sequence:

$$P(t, \hat{s}_t, \hat{r}_t \mid E, t, p) = \prod_{i=1}^{N} P(y_i \mid y_{<i}, E, t, p) \tag{1}$$

where P indicates the probability distribution of the model's output in our autoregressive score-reasoning prediction, and N is the number of tokens in the output sequence. Specifically, the model is trained to generate the output in this structured format:

$$t \quad \hat{s}_t \quad \texttt{<RATIONALE>} \quad \hat{r}_t \quad \texttt{</RATIONALE>} \tag{2}$$

4 Experiments

Datasets. We use the most representative publicly available AES dataset, a combination of ASAP[2] and ASAP++[3] (Mathias and Bhattacharyya, 2018). All comparison multi-trait scoring models are also evaluated on this dataset. ASAP++ provides human-annotated multi-trait scores for essays written in English across eight distinct prompts, offering a more granular evaluation of writing quality. Notably, ASAP++ complements the original ASAP dataset by incorporating additional trait scores that were absent in the original one. As summarized in Table 2, each prompt is assessed using a distinct set of writing traits with varying score ranges. While most traits appear across multiple prompts, Style and Voice are exclusively evaluated in Prompts 7 and 8, respectively, resulting in a limited number of training samples for these traits.

Models and settings. For the teacher LLM, we select Llama3.1-70B (Dubey et al., 2024), an open-source model, demonstrating competitive performance to GPT-4o (Hurst et al., 2024) and Claude 3.5 Sonnet (Anthropic, 2024) on the massive multi-task language understanding benchmark, to avoid the reliance on costly proprietary LLMs. As the student scoring-expert model, we employ T5-large (770M) (Raffel et al., 2020), a Transformer-based model. The generation process follows the hyper-parameter settings as ArTS, using Seq2SeqTrainer with 5,000 evaluation steps, a batch size of 4, and 15 epochs. Experiments are performed on A100-SMX4-8 GPUs.

Evaluations and baseline models. In line with previous studies (Taghipour and Ng, 2016; Do et al., 2024a; Chu et al., 2024), we perform five-fold cross-validation using the same dataset splits as their work. For evaluation, we adopt QWK, the official metric of the ASAP dataset, and report both the five-fold average scores and their standard deviations. To ensure a fair comparison, we also compute QWK scores separately for each prompt, following previous systems (Taghipour and Ng, 2016; Do et al., 2024a; Chu et al., 2024). As baseline models, we primarily compare our approach with the robust ArTS model (Do et al., 2024a) and its stronger extensions models, SaMRL (Do et al., 2024b) and RMTS (Chu et al., 2024). Details on baseline models are provided in Appendix A.

5 Results

5.1 Quality of multi-trait scoring

Our experimental results demonstrate that RaDME achieves robust scoring performance across multiple traits and prompts while offering explainability. As shown in Table 3 and Table 4, RaDME outperforms other strong and state-of-the-art models, achieving the highest average QWK score in both trait-wise and prompt-wise evaluations. It is noteworthy that our method for enhancing the interpretability of scoring could also jointly improve the assessment quality. Remarkably, under the same training conditions, the sequential generation of the score and rationale (RaDME) consistently outperforms its counterpart without rationale distillation (RaDME-w/o R), achieving significant performance improvements across all traits and prompts.

[2]https://www.kaggle.com/c/asap-aes
[3]https://lwsam.github.io/ASAP++/lrec2018.html

Model	Explainability	Traits											
		Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑
HISK	✗	0.718	0.679	0.697	0.605	0.659	0.610	0.527	0.579	0.553	0.609	0.489	0.611
STL-LSTM	✗	0.750	0.707	0.731	0.640	0.699	0.649	0.605	0.621	0.612	0.659	0.544	0.656
MTL-BiLSTM	✗	**0.764**	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.638
ArTS-large (←)	✗	0.751	0.730	0.750	0.701	0.728	0.675	0.682	0.680	0.680	0.715	0.603	0.700
ArTS-ind	✗	0.723	0.717	0.752	0.695	0.713	0.649	0.659	0.662	0.675	**0.722**	0.548	0.683
RMTS-GPT	✗	0.755	0.737	0.752	**0.713**	**0.744**	0.682	0.690	**0.705**	**0.694**	0.702	0.612	0.708
RMTS-Llama	✗	0.754	0.730	0.749	0.701	0.737	0.675	0.684	0.690	0.684	0.696	**0.640**	0.704
SaMRL-large	✗	0.754	0.735	0.751	0.703	0.728	0.682	0.685	0.688	0.691	0.710	0.627	0.705
RaDME-w/o R	✗	0.713	0.700	0.728	0.655	0.683	0.636	0.654	0.647	0.652	0.684	0.548	0.664
RaDME	✓	0.754	**0.744**	**0.759**	0.706	0.736	**0.701**	**0.692**	0.693	0.692	0.719	0.623	**0.711**

Table 3: Trait-wise effects of RaDME on ASAP/ASAP++ averaged over prompts. The numerical values denote QWK scores. Bolded and underlined scores highlight the highest and the second-highest performance, respectively.

Model	Explainability	Prompts								
		1	2	3	4	5	6	7	8	AVG↑
HISK	✗	0.674	0.586	0.651	0.681	0.693	0.709	0.641	0.516	0.644
STL-LSTM	✗	0.690	0.622	0.663	0.729	0.719	0.753	0.704	0.592	0.684
MTL-BiLSTM	✗	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665
ArTS-large (←)	✗	0.701	0.698	0.705	0.766	0.725	0.773	0.743	0.635	0.718
ArTS-ind	✗	0.695	0.679	0.705	0.762	0.721	0.756	0.734	0.578	0.704
RMTS-GPT	✗	**0.716**	0.704	**0.723**	**0.772**	**0.737**	0.769	0.736	0.651	0.726
RMTS-Llama	✗	0.705	0.692	0.714	0.766	0.726	0.773	0.726	**0.658**	0.720
SaMRL-large	✗	0.702	0.711	0.708	0.766	0.722	0.773	0.743	0.649	0.722
RaDME-w/o R	✗	0.665	0.669	0.664	0.731	0.690	0.735	0.704	0.605	0.683
RaDME	✓	0.705	**0.716**	0.715	**0.772**	0.731	**0.774**	**0.762**	0.654	**0.729**

Table 4: Prompt-wise effects of RaDME on ASAP/ASAP++ averaged over traits.

Note that RaDME-w/o R is designed to isolate the effect of rationale generation, as it predicts each trait in a separate sequence as RaDME, but without generating rationales. This outcome suggests that training models to consider succeeding rationales, rather than solely relying on encoder outputs, can enhance the accuracy of score predictions, underscoring the efficacy of integrating the reasoning process into the scoring decision. Since our study focuses on prompt-specific multi-trait scoring, it is not directly comparable to cross-prompt models (Do et al., 2023; Chen and Li, 2023, 2024) and is therefore excluded from comparisons.

Single vs. sequential trait prediction. As observed in the comparison between ArTS (i.e., predicting all traits in a sequence) and ArTS-ind (i.e., using individual models for separated trait sequences), incorporating trait-wise dependencies within an autoregressive decoding strategy has been revealed to improve performance (Do et al., 2024a). However, in our experiments on a single fold, predicting the score-rationale sequence for all traits within a single forward pass resulted in significantly lower performance, with a trait-wise average of 0.454 and a prompt-wise average of 0.504. Since RaDME generates a rationale for each trait score, predicting all traits at once can cause instability in subsequent trait predictions. Additionally, as each essay can be evaluated on up to seven traits, later predictions may suffer from information loss regarding the essay content. This comparison emphasizes that in scenarios requiring explanations, our approach, i.e., predicting one trait at a time, can better assist the model in understanding the contexts to make precise decisions. It is noteworthy that despite ArTS, SaMRL (Do et al., 2024b), and RMTS (Chu et al., 2024) being explicitly designed to leverage trait dependencies, giving them an inherent advantage in multi-trait AES tasks, RaDME still outperforms them. This result provides strong evidence that rationale distillation itself can enhance the model's decision-making capabilities, further validating our approach.

Self-explaining vs. injecting rationales. We investigate the impact of rationales as input (i.e., RMTS (Chu et al., 2024)), versus generating rationales as an output (RaDME). While RMTS incorporates rationales as additional context for AES, our RaDME generates them internally, allowing the model to self-explain its scoring decisions. Despite RMTS being based on trait dependencies and rationale injection, which highly benefits scoring, RaDME's self-explanatory mechanism surpasses it in overall scoring performance. The results suggest that training the model to learn reasoning alongside scoring not only enhances model interpretability

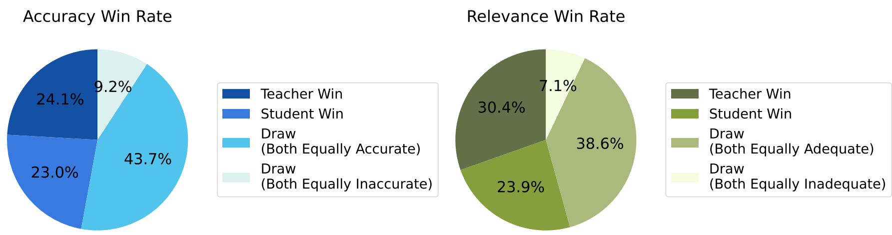

Figure 3: Evaluation of win rates for accuracy and relevance between rationales generated by the student model and those generated by the LLM on the test set.

but also improves scoring efficacy. Notably, our system achieves higher performance in *Content*, *Prompt Adherence*, and *Organization*, which require a comprehensive understanding of the essay's contextual coherence and logic rather than just identifying isolated elements. Beyond its robustness in scoring, a key advantage of RaDME is that it does not require LLMs at inference time, making it significantly more efficient and scalable for practices.

5.2 Quality of the generated rationales

To validate the quality of the rationale generated by the RaDME, we measured the winning rate using GPT-4o, randomly selecting 1,000 samples. The evaluation involved comparing two rationales: the teacher LLM's rationale and the RaDME-generated (i.e., student model's) rationale. Evaluators selected one of four possible outcomes: Teacher (Rationale 1) Win, Student (Rationale 2) Win, Draw (Both Good), or Draw (Both Poor). A detailed example of the prompt is provided in Appendix C (Table 7). We evaluated two dimensions: *accuracy*, which measures whether the rationale contains only correct information, and *relevance*, which assesses whether the rationale adequately includes the necessary information.

Figure 3 results showed that only 9.2% of samples were rated as poor for both rationales. Surprisingly, 66.7% of the student rationales have results more accurate or equally accurate compared to the teacher-generated ones, suggesting that the scoring procedure itself may influence the quality of the rationale. For relevance, only 8.8% of the samples were rated as poor for both models, and the majority (62.5%) were judged to be better or as good as the teacher model. These findings demonstrate that RaDME, even with an efficient student model, can generate effective rationales that accurately convey the reasoning behind essay scores, making it suitable for practical settings. Detailed qualitative analyses for rationales are provided in Appendix B.

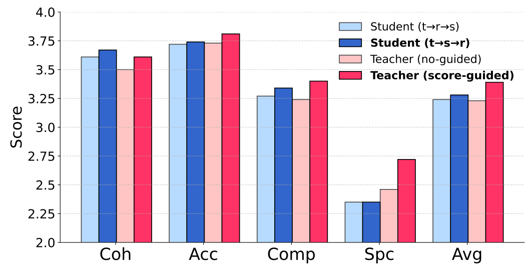

Figure 4: Evaluation results with G-Eval.

6 Discussions

Effect of score-rationale generation order. In our framework, RaDME generates a trait score first, followed by its rationale. To examine whether this prediction order is optimal, we conduct additional experiments where the model first generates the rationale and then the score.

Our results in Table 5 and Table 6 demonstrate that the score-first approach consistently outperforms the rationale-first approach across both trait-wise and prompt-wise evaluations. Determining the score before generating the rationale anchors the explained output to a concrete decision, ensuring alignment between the two and leading to stable predictions. Contrarily, without a predefined score to guide the rationale, the model may struggle to produce explanations that align with appropriate numerical assessment, resulting in greater variance in score predictions. These results suggest that in AES tasks where both accuracy and explainability are vital, deciding the score first is more effective.

We further investigated whether rationale quality itself benefits from being predicted first (Figure 4; blue). For evaluation, we utilized G-Eval (Liu et al., 2023), an evaluation framework proven robustness in natural language generation using GPT-4. We compared RaDME ($t \rightarrow s \rightarrow r$) with its reverse variant ($t \rightarrow r \rightarrow s$) to assess the impact of generation order on rationale quality. G-Eval automatically generates a Chain-of-Thought (CoT)

Model	Traits											
	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑
Teacher (No-guided)	0.405	0.406	0.358	0.381	0.353	0.471	0.395	0.459	0.456	0.235	0.491	0.401
RaDME ($t \rightarrow r \rightarrow s$)	0.728	0.727	0.750	0.671	0.720	0.678	0.686	0.642	0.673	0.702	0.524	0.682
RaDME ($t \rightarrow s \rightarrow r$)	**0.754**	**0.744**	**0.759**	**0.706**	**0.736**	**0.701**	**0.692**	**0.693**	**0.692**	**0.719**	**0.623**	**0.711**

Table 5: Trait-wise QWK results of Teacher w/o score-guidance, RaDME ($t \rightarrow r \rightarrow s$), and the original RaDME ($t \rightarrow s \rightarrow r$). Bolded scores highlight the highest performance.

Model	Prompts								
	1	2	3	4	5	6	7	8	AVG↑
Teacher (No-guided)	0.407	0.526	0.353	0.412	0.395	0.353	0.331	0.425	0.400
RaDME ($t \rightarrow r \rightarrow s$)	0.699	0.688	0.696	0.763	0.708	0.747	0.740	0.605	0.706
RaDME ($t \rightarrow s \rightarrow r$)	**0.705**	**0.716**	**0.715**	**0.772**	**0.731**	**0.774**	**0.762**	**0.654**	**0.729**

Table 6: Prompt-wise QWK results of Teacher w/o score-guidance, RaDME ($t \rightarrow r \rightarrow s$), and the original RaDME.

evaluation based on the target dimension's criteria. We define four key dimensions for assessing rationale quality: *Coherence*, *Accuracy*, *Completeness*, and *Specificity*. *Coherence* evaluates whether the rationale presents a clear, structured, and logically connected explanation, with higher scores indicating smoother and well-organized reasoning. *Accuracy* measures how correctly the rationale justifies the assigned score. *Specificity* assesses the level of detail, where higher scores reflect more precise explanations. *Completeness* examines whether the rationale fully addresses all relevant aspects of the essay trait being scored. Following these criteria, we generate a CoT-based assessment and assign scores on a 1–5 scale for randomly selected 100 samples. Specific evaluation methods and the prompts used are detailed in Appendix C.3. Interestingly, comparison results (Figure 4) suggest predicting the score first also leads to superior rationale quality across all dimensions. The results indicate that the preceding score decision can induce more structured and coherent reasoning, while the opposite increases variance in explanation quality.

Impact of score-guided prompting. To fully leverage the capabilities of the teacher LLM, we proposed a *score-guided prompting* strategy. To verify whether our guidance effectively led to high-quality outputs, we conducted a comparative analysis with no guidance prompting. As shown in Table 5 and Table 6, the teacher model without score guidance (Teacher No-Guided), where the LLM is directly tasked with generating both scores and rationales, exhibits poor scoring performance. This result aligns with previous research findings (Lee et al., 2024), which highlight the inherent limitations of LLMs in accurate score prediction, supporting the necessity of our score-guidance method.

We also conducted further evaluations of the generated rationales using G-Eval (Liu et al., 2023), following the same four evaluation criteria defined earlier. Our results (Figure 4; red) reveal that providing exact trait scores significantly enhances rationale quality across all evaluation dimensions. Specifically, our score-guided generation yields rationales that are more coherent, accurate, and complete compared to those generated without explicit score guidance. Notably, we observe a substantial improvement in *Specificity*, suggesting that grounding rationale generation in predefined correct answers enhances the model's ability to produce more precise and well-structured explanations. These findings highlight the critical role of explicit score guidance in improving rationale generation, which then subsequently affects the distillation efficacy for optimizing the student model.

7 Conclusion

We propose RaDME, a self-explainable, rationale-driven multi-trait AES method that enhances both transparency and scoring accuracy. Unlike existing AES systems that lack an explanation for assigned scores, RaDME explicitly generates rationales alongside trait scores, making its decisions interpretable. By distilling LLMs' reasoning capabilities into a scoring-efficient student model, RaDME achieves both high-quality scoring and clear, coherent explanations. Our extensive experiments reveal that while LLMs struggle with direct scoring, they excel in rationale generation when provided with precise numerical scores. RaDME successfully integrates this strength, producing accurate, well-structured, and detailed rationales while maintaining outstanding scoring performance.

Limitations

While RaDME demonstrates strong technical and empirical performance, its practical impact in real-world educational environments, particularly in human-centered interactions, remains an area for future exploration. Assessing its effect on students in interactive settings would further enhance the significance of this research. Nevertheless, we believe this limitation is not unique to RaDME but applies to most recent AES systems, which often lack explicit evaluation of human-centered effects. Therefore, future research could extensively explore how educators and students perceive, interpret, and utilize rationale-based feedback provided by RaDME, presenting an important direction for further investigation.

Ethical Statement

We used publicly available datasets of automated essay scoring in this study, ensuring compliance with ethical guidelines and data usage policies. The dataset does not contain personally identifiable or sensitive information.

Acknowledgments

References

AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3.

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

Yuan Chen and Xia Li. 2024. PLAES: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12775–12786, Torino, Italia. ELRA and ICCL.

SeongYeub Chu, JongWoo Kim, Bryan Wong, and MunYong Yi. 2024. Rationale behind essay scores: Enhancing s-llm's multi-trait essay scoring with rationale generated by llms. *arXiv preprint arXiv:2410.14202*.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Lee. 2024a. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian's, Malta. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Lee. 2024b. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA. Association for Computational Linguistics.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Matthew Johnson and Mo Zhang. 2024. Examining the responsible use of zero-shot ai approaches to scoring essays. *Scientific Reports*, 14(1):1–10.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. In *Frontiers in education*, volume 5, page 572367. Frontiers Media SA.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024a. Can large language models automatically score proficiency of written essays? *arXiv preprint arXiv:2403.06149*.

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024b. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia. ELRA and ICCL.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

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.

Ali Ghiasvand Mohammadkhani. 2024. Rdbe: Reasoning distillation-based evaluation enhances automatic essay scoring. *Preprint*, arXiv:2407.13781.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sangwon Ryu, Heejin Do, Daehee Kim, Yunsu Kim, Gary Geunbae Lee, and Jungseul Ok. 2024. Guide-to-explain for controllable summarization. *arXiv preprint arXiv:2411.12460*.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

Katsuhito Sudoh, Satoshi Nakamura, and 1 others. 2024. Automated essay scoring using grammatical variety and errors with multi-task learning and item response theory. *arXiv preprint arXiv:2406.08817*.

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.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*.

Yupei Wang and Renfen Hu. 2021. A prompt-independent and interpretable automated essay scoring method for chinese second language writing. In *China National Conference on Chinese Computational Linguistics*, pages 450–470. Springer.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Baseline Models

We primarily compare our method with ArTS (Do et al., 2024a) and its enhanced multi-trait scoring extensions, SaMRL (Do et al., 2024b), and RMTS (Chu et al., 2024). Among them, SaMRL incorporates scoring-aware reinforcement learning, while RMTS leverages rationales extracted from LLMs as additional input.

Following prior multi-trait scoring studies, we also report results for other multi-trait scoring models (Kumar et al., 2022) and holistic scoring models (Cozma et al., 2018; Dong et al., 2017), where each holistic model is applied independently for trait-specific predictions. Specifically, the multi-trait scoring MTL model (Kumar et al., 2022) employs trait-specific layers while using all other trait layers as auxiliary components for target trait training and prediction. For holistic scoring, HISK utilizes a support vector regressor with a histogram intersection string kernel, while STL-LSTM applies an LSTM-CNN-based structure, with each model iteratively deployed for independent trait scoring.

B Qualitative Analysis of Rationales

Figure 5, 6, and 7 compare the actual generated rationales from four different models, discussed in Section 6. As illustrated in these figures, our proposed methods yield more detailed and clearly articulated rationales.

Trait: Language (Score: 2.0)

Student (t → r → s): The essay demonstrates limited control over language, with frequent errors in grammar, spelling, and punctuation, such as "modify this top as a mooring mast" and "costed money.

Student (t → s → r): The essay contains numerous grammatical errors, such as "do modify", "accomadate", "costed", and "hit something or caught fire", which detract from the overall clarity and effectiveness of the writing.

Teacher (no-guided): The essay contains several errors in spelling, grammar, and punctuation, which detract from the overall clarity of the writing.

Teacher (score-guided): The essay contains several grammatical errors ("accomadate", "costed", "if the shell broke, the dirigible could hurt many people") and simplistic vocabulary, which detract from the overall clarity and effectiveness of the writing.

Figure 5: Comparison of rationales generated by different models for the *Language* trait, in the case of a score of 2. Bolded models represent our proposed methods, while green highlights indicate well-specified phrases within the rationales.

Trait: Word Choice (Score: 3.0)

Student (t → r → s): The essay's word choice is limited, with frequent use of basic vocabulary and a lack of precise language to convey the writer's ideas effectively."

Student (t → s → r): The essay's vocabulary is limited, with frequent use of basic words and phrases, such as "a lot of affects" and "you don't really have a social life," which fail to add depth or nuance to the writing.

Teacher (no-guided) :The essay uses basic vocabulary, but the language is often informal and lacks precision, with some words being repeated or used inappropriately.

Teacher (score-guided): The writer uses a limited range of vocabulary, often relying on basic, informal words and phrases that lack precision and nuance, with some instances of awkward phrasing ("Computers have a lot of affects on people").

Figure 6: Comparison of rationales generated by different models for the *Word Choice* trait, in the case of a score of 3.

Trait: Style (Score: 5.0)

Student (t → r → s): The essay demonstrates a clear and engaging narrative voice, with vivid descriptions of the scene and characters, but the writing could benefit from more varied and sophisticated sentence structures to elevate the overall style.

Student (t → s → r): The writer uses vivid and engaging descriptions, such as "dirty blond hair whips her face," "hunches her back against the wind," and "I was cold, bord, and hungry," to bring the story to life, demonstrating a strong sense of style.

Teacher (no-guided): The essay has a conversational tone, and the writer uses vivid descriptions, such as "dirty blond hair whips her face," to create a sense of atmosphere.

Teacher (score-guided): The writer uses some vivid descriptions of the scene, such as "Frezing cold," "dirty blond hair whips her face," and "yellow lab," which create a sense of atmosphere, but the writing lacks variety in sentence structure and vocabulary.

Figure 7: Comparison of rationales generated by different models for the *Style* trait, in the case of a score of 5.

C Detailed Prompts

C.1 Prompt for the winning rate evaluation

The prompt used to evaluate the winning rate between the teacher's and the student's rationales is described in Table 7. The results are shown in Figure 3 and discussed in Section 5.2.

C.2 Prompt for the teacher LLM without score-guided prompting

The prompt used to generate the rationale without our score-guided prompting strategy, as a comparison results, are shown in Table 8. The results are shown in Table 5 and Table 6 and discussed in Section 6.

C.3 Prompts for G-Eval

Based on the criteria of each of the four defined dimensions, we generate a CoT for rationale evaluation. We conduct n evaluations using its own steps and then compute a weighted summation to determine the final score: $\sum_{i=1}^{n} p(s_i) \times s_i$, where p denotes the probability of each score and s denotes the score set. The used prompts are illustrated in Figure 8, 9, 11, 10.

Prompt for Accuracy	**Prompt for Relevance**
[system prompt] Your task is to evaluate which rationale most **accurately** explains the assigned scores for the essay. [input prompt] Please review the essay, trait score, and each rationale carefully, and then choose one of the following options: ### Writing Instruction: {instruction} ### Essay: {essay} ### {trait} Trait Score: {score} ### Rationale 1: {rationale1} ### Rationale 2: {rationale2} 1. Rationale 1 most **accurately** explains the trait quality of the essay. 2. Rationale 2 most **accurately** explains the trait quality of the essay. 3. Draw (both rationales are equally **accurate**): Both rationales provide an equally **accurate** explanation of the assigned scores. 4. Draw (both rationales are equally **inaccurate**): Both rationales fail to provide an **accurate** explanation of the assigned scores. Provide only the corresponding option number:	[system prompt] Your task is to evaluate which rationale most **adequately** explains the assigned scores for the essay. [input prompt] Please review the essay, trait score, and each rationale carefully, and then choose one of the following options: ### Writing Instruction: {instruction} ### Essay: {essay} ### {trait} Trait Score: {score} ### Rationale 1: {rationale1} ### Rationale 2: {rationale2} 1. Rationale 1 most **adequately** explains the trait quality of the essay. 2. Rationale 2 most **adequately** explains the trait quality of the essay. 3. Draw (both rationales are equally **adequate**): Both rationales provide an equally **adequate** explanation of the assigned scores. 4. Draw (both rationales are equally **inadequate**): Both rationales fail to provide an **adequate** explanation of the assigned scores. Provide only the corresponding option number:

Table 7: Evaluation prompts for comparing the winning rates of rationales generated by teacher and student models. An example of the *accuracy* aspect.

You will be given a rationale generated based on the score assigned to a specific essay.
Your task is to rate the rationale on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence (1-5) – Logical Structure and Flow. Evaluates whether the rationale follows a clear, structured, and logically connected explanation. Higher scores indicate smooth progression with well-organized reasoning, while lower scores reflect disjointed, confusing, or contradictory justifications.

Evaluation Steps:

1. Read the rationale entirely to understand its argument's full context.
2. Evaluate if the argument flows logically, if the points made are well-structured and connected.
3. Consider whether the reviewer explains each point convincingly. There should not be any contradictions in what they say.
4. Assess whether the rationale provides a clear and orderly explanation or if it is disorderly, confusing, or contradicts itself.
5. Score the rationale on the scale of 1-5. A high coherence score (close to 5) indicates a smooth and well-structured explanation, while a low coherence score (close to 1) indicates a disjointed or confusing rationale. For instance, missing steps in logic, jumping from one point to another, statements that contradict each other, etc., will lead to a lower score.

Writing Instruction:
{{Instruction}}

Essay:
{{Essay}}

Rationale (Trait: {{Trait}}, Score: {{Score}}):
{{Rationale}}

Evaluation Form (scores ONLY):

- Coherence:

Figure 8: Evaluation for *Coherence*.

You will be given a rationale generated based on the score assigned to a specific essay.
Your task is to rate the rationale on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Accuracy (1-5) – The correctness of the rationale in reflecting the essay's quality for the given trait. A rationale should provide an objective and faithful assessment, aligning with the essay's actual strengths and weaknesses.

Evaluation Steps:

1. Read the rationale carefully and understand the evaluator's reasoning behind the given score.
2. Refer to the original essay if necessary, to determine if the evaluator's justification accurately represents the essay's performance.
3. Assess whether the justification for the score is factual and aligned with the scoring criteria. Does it accurately reflect the strengths and weaknesses of the essay?
4. If the rationale corresponds with the essay's strengths and weaknesses, and is aligned with the scoring criteria, it gets a high accuracy score - close to 5.
5. If the rationale doesn't align with the essay's actual performance, the accuracy score should be low - close to 1.
6. Accuracy rating is not about whether you agree with the evaluator's opinion, but whether their reasoning accurately mirrors the essay's performance based on scoring criteria.
7. Finally, assign a score for the accuracy of the rationale on a five-point scale (1-5). A score of 1 represents a mainly inaccurate justification and a score of 5 represents a highly accurate justification.

Writing Instruction:
{{Instruction}}

Essay:
{{Essay}}

Rationale (Trait: {{Trait}}, Score: {{Score}}):
{{Rationale}}

Evaluation Form (scores ONLY):

- Accuracy:

Figure 9: Evaluation for *Accuracy*.

You will be given a rationale generated based on the score assigned to a specific essay.
Your task is to rate the rationale on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Completeness (1-5) – Coverage of Necessary Aspects. Checks whether the rationale fully addresses all relevant aspects of the essay trait being scored. A complete rationale covers all key elements, while an incomplete one omits critical details, leading to a less informative explanation.

Evaluation Steps

1. Start by understanding the trait of the essay being scored.
2. Read the rationale carefully.
3. While reading, identify if the rationale is explaining all the relevant aspects about the essay trait being scored.
4. Note down if you find any important element pertaining to the scored trait that has been omitted from the explanation.
5. Check whether the rationale is informative and provides a clear explanation about the score assigned to the essay.
6. Based on your assessment, assign a score between 1 and 5. Assign 1 if the rationale is not at all complete and omits many critical details related to the scoring trait. Assign 5 if you believe the rationale is highly complete and covers all relevant aspects of the essay trait being scored, providing a clear and informative explanation accordingly.
7. Make a final decision about the score and submit your rating. If you're unsure, reevaluate the rationale against the evaluation steps mentioned above before making your final choice.

Writing Instruction:
{{Instruction}}

Essay:
{{Essay}}

Rationale (Trait: {{Trait}}, Score: {{Score}}):
{{Rationale}}

Evaluation Form (scores ONLY):

- Completeness:

Figure 10: Evaluation for *Completeness*.

You will be given a rationale generated based on the score assigned to a specific essay.
Your task is to rate the rationale on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Specificity (1-5) – The inclusion of essay-specific reasoning. A rationale should provide concrete details directly related to the essay rather than generic or vague explanations. Annotators should penalize rationales that rely on broad, non-specific justifications.

Evaluation Steps:

1. Carefully read the provided rationale.
2. Check for the level of detail provided in the explanation. Look for specifics such as quotes or references from the essay, precise aspects related to the essay's content, style, structure, grammar, etc., or detailed explanations indicating the scorer's thought process.
3. Also verify whether this detail ties directly into the trait being discussed. Pure detail is not enough if it does not provide reasoning linked to the trait score.
4. Refer to only the level of detail in the explanation when scoring. Do not consider other factors, such as whether you agree with the assessment, etc.
5. Assign a score from 1 to 5. A score of 1 indicates a very generic rationale with no supporting details, and a score of 5 indicates a very detailed, specific rationale with clear, concrete references or explanations.

Writing Instruction:
{{Instruction}}

Essay:
{{Essay}}

Rationale (Trait: {{Trait}}, Score: {{Score}}):
{{Rationale}}

Evaluation Form (scores ONLY):

- Specificity:

Figure 11: Evaluation for *Specificity*.

LLM Prompt	**Generation Output**
Your role is to assign trait scores to the narrative type essay and explain the reasoning behind the assigned scores. Please read the essay and assign its trait scores, each followed by a one-sentence rationale. Words beginning with '@' have been intentionally replaced to anonymize personally identifying information in the essays. Writing Instructions: {Write about patience. Being patient means that you are understanding and tolerant. A patient person experience difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.} Essay: {Being patience is hard to do, @CAPS1 I do remember a time when I was patient. This was at @CITY1. Me and my group had to wait for everyone to get back to the bus. We were the first one's to the bus so it took a while before they came. @CAPS1 before they did I was as calm and patient as can be.} Traits (Score ranges): {Style (0-6), Conventions (0-6), Organization (0-6), Content (0-6), Overall (0-30)} Please score each trait only once with the following format: 'trait score [score]: [rationale].' Assigned Scores and Rationales: {1) Style score }	Assigned Scores and Rationales: 1) Style score 2: The writing lacks a clear and engaging tone, and the language is simple and repetitive. 2) Conventions score 3: The essay contains several grammatical errors, including incorrect verb tense and subject-verb agreement, but the writer demonstrates some control over basic sentence structure. 3) Organization score 3: The narrative is loosely structured and lacks transitions between sentences, but the writer attempts to tell a story with a clear beginning and end. 4) Content score 3: The essay provides a brief anecdote about being patient, but the story is thin and lacks descriptive details that would make it more engaging and memorable. 5) Overall score 11: The essay demonstrates some effort to tell a story about patience, but it is marred by errors in grammar, syntax, and organization, and lacks a clear and engaging tone.

Table 8: The prompt used for the teacher LLM without our score-guidance and its corresponding generation output. Although the ground-truth scores are *Style: 3, Conventions: 4, Organization: 2, Content: 2, Overall: 11*, the scores generated without score guidance deviate significantly from the actual values. This result contrasts with the findings presented in Table 1.