Rationale Behind Essay Scores: Enhancing S-LLM's Multi-Trait Essay Scoring with Rationale Generated by LLMs

Seong Yeub Chu[1*], **Jong Woo Kim**[2*], **Bryan Wong**[1], **Mun Yong Yi**[1†]
[1]Graduate School of Data Science, KAIST
[2]Department of Industrial & Systems Engineering, KAIST
{chseye7, gsds4885, bryan.wong, munyi}@kaist.ac.kr

Abstract

Existing automated essay scoring (AES) has solely relied on essay text without using explanatory rationales for the scores, thereby forgoing an opportunity to capture the specific aspects evaluated by rubric indicators in a fine-grained manner. This paper introduces Rationale-based Multiple Trait Scoring (RMTS), a novel approach for multi-trait essay scoring that integrates prompt-engineering-based large language models (LLMs) with a fine-tuning-based essay scoring model using a smaller large language model (S-LLM). RMTS uses an LLM-based trait-wise rationale generation system where a separate LLM agent generates trait-specific rationales based on rubric guidelines, which the scoring model uses to accurately predict multi-trait scores. Extensive experiments on benchmark datasets, including ASAP, ASAP++, and Feedback Prize, show that RMTS significantly outperforms state-of-the-art models and vanilla S-LLMs in trait-specific scoring. By assisting quantitative assessment with fine-grained qualitative rationales, RMTS enhances the trait-wise reliability, providing partial explanations about essays. The code is available at **https://github.com/BBeeChu/RMTS.git**.

1 Introduction

Multi-trait essay scoring, which evaluates essays on multiple dimensions such as *Content*, *Organization*, and *Style*, rather than on a single holistic score, has recently become a central issue in automated essay scoring (AES). Extensive research in this area has primarily utilized BERT and trait-wise layers to predict scores for individual traits (Mathias and Bhattacharyya, 2020; Ridley et al., 2021; Kumar et al., 2021; Do et al., 2023). Notably, Do et al. (2024) proposed using an autoregressive pre-trained language model, T5 (Raffel et al., 2020), for greater computational efficiency. Despite these efforts, most studies have used essay texts alone to predict labels as represented in Figure 1 (A), rather than extracting aspects evaluated by rubric indicators from the essays and using them.

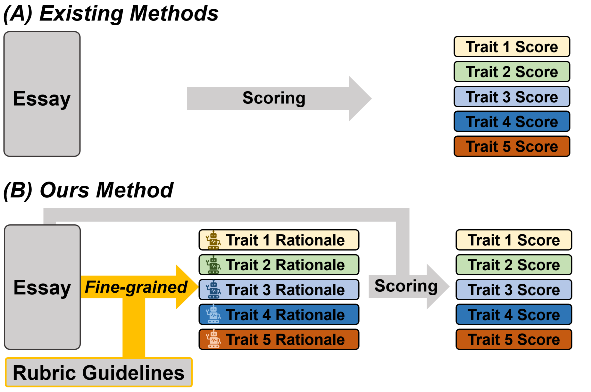


Figure 1: Unlike existing methods (A), we use multiple prompt-engineering LLMs to generate trait-specific rationales based on rubric guidelines as shown in (B), which are then combined with an S-LLM for comprehensive evaluation.

With the advent of LLMs, generating fine-grained rationales—explanations of how essays align with rubric criteria—has become feasible. As shown in Figure 1, incorporating rationales identifies relevant essay sections that demonstrate specific traits and links them directly to the rubric. This approach mirrors how human evaluators use rubrics to assess essays in real-world settings (Freeman and Miller, 2001). For instance, a rationale for *Organization* highlights transitions and structure, leading to more precise, rubric-aligned evaluations. Without rationales, the model may overlook key elements and score less accurately by focusing only on the semantic sequence.

To the best of our knowledge, few studies have attempted to use rationale-based evaluations derived from rubrics to assess essays (Lee et al., 2024; Li et al., 2023). Lee et al. (2024) used LLMs to pre-

[*]Both authors contributed equally to this research.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 5811–5829
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

dict holistic scores based on criteria synthesized from rubrics, but the models cannot be fine-tuned as this approach relies on prompt engineering. Li et al. (2023) combined LLM-generated rationales with human-assessed scores to create new labels for training sequence-to-sequence models. However, this method does not use the rationales as inputs to the encoders when predicting scores. Besides, it differs from our research as the work only focused on predicting overall score of short-answer responses rather than multiple trait scores of long-context essays.

In this paper, we propose a novel approach of effectively utilizing rationales in conjunction with rubics and essays to enable LLMs to better assess the various aspects of essays as outlined by the rubrics. These rationales, or qualitative assessments, are then used by an encoder-decoder-based smaller large language model, hereafter referred to as S-LLM, to predict scores more precisely. The S-LLM is fine-tuned using training data that includes human-rated scores. This approach aims to enhance the accuracy and reliability of automated multi-trait essay scoring.

The main contributions of this study are as follows:

- We introduce a novel approach to multi-trait scoring, **R**ationale-based **M**ulti-**T**rait essay **S**coring (**RMTS**), which combines an essay and a rationale together to predict multi-trait scores. This model utilizes the rationale to explicitly capture the elements assessed by the rubric from the essay.

- We compose trait-specific prompts using essays and rubrics to build an LLM-based trait-wise rationale generation system. This system generates rationales, which serve as the foundation for the multi-trait scores.

- We conducted a comprehensive analysis of the generated rationales and verified that they are sufficiently meaningful to be effectively utilized in essay scoring.

- Extensive experiments with five S-LLMs demonstrate that incorporating LLM-generated rationales significantly improves essay scoring, providing a model-agnostic incremental benefit to each S-LLM. Our approach advances the state-of-the-art baselines in essay scoring on the ASAP and ASAP++ benchmark datasets.

2 Related Work

2.1 Traditional and transformer-based automated essay scoring

Traditional automated essay scoring (AES) focused on holistic scoring, predicting an overall score using handcrafted features and linear regression models (Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017; Cozma et al., 2018). Particularly, transformer-based models like BERT (Devlin et al., 2018) significantly improved AES by capturing detailed language information (Yang et al., 2020a; Wang et al., 2022; Mayfield and Black, 2020). These models enhanced scoring accuracy but were primarily used for holistic scoring. Extending them to multi-trait scoring is inefficient due to the need for multiple models for different traits, increasing computational costs (Kumar et al., 2021; Do et al., 2024).

2.2 Multi-trait essay scoring approaches

Multi-trait AES evaluates essays across various dimensions with respect to different features existing in essays such as *Content*, *Organization*, and *Conventions*. Existing models used multiple linear layers or separate models for each trait, which require intensive resources (Mathias and Bhattacharyya, 2020; Ridley et al., 2021). Recent approaches introduced multi-task learning frameworks with shared models and trait-specific layers, improving efficiency (Kumar et al., 2021). However, handling trait dependencies and requiring specialized modules remains challenging. The autoregressive multi-trait scoring (ArTS) model addressed this by using a pre-trained T5 model to sequentially generate trait scores, leveraging inter-dependencies for better accuracy (Do et al., 2024). Yet, it still relied solely on essay texts alone for score prediction.

2.3 Rubric-based essay scoring using large language model

Recently, LLMs have been used to evaluate essays alongside assessment rubrics, showing competitive performance. For example, one study (Lee et al., 2024) divided criteria into multiple traits and generated sub-criteria for scoring, achieving moderate results. Another study (Li et al., 2023) used rubrics to score short answers and generated rationales, which were then used as labels to fine-tune the T5 model (Raffel et al., 2020) to produce both scores and rationales. However, this approach did not outperform fine-tuned models like BERT (Devlin et al., 2018) and Longformer (Beltagy et al., 2020a), which were trained using only scores as labels. Our approach differs by directly extracting rubric-based rationales from essays using LLMs and feeding them into a pre-trained S-LLM. This method explicitly considers detailed scoring criteria, improving alignment with human evaluators and enhancing both the reliability and transparency of automated essay scoring.

3 RMTS

RMTS is a framework that enhances the multi-trait essay scoring capabilities of an S-LLM, a pre-trained sequence-to-sequence model, by incorporating rationales. The framework consists of two parts: (1) generating trait-specific rationales using an LLM-based system with GPT-3.5 Turbo (OpenAI, 2022)[1] and Llama-3.1-8B-Instruct (Touvron et al., 2023)[2] (referred to as GPT and Llama respectively), and (2) extracting representations from both the essay and rationale using a shared encoder of the S-LLM. This dual-process approach improves the reliability of the scoring model. The detailed procedure is shown in Figures 2 and 3.

3.1 LLM-based trait-wise rationale generation system

As illustrated in Figure 2, individual trait-specific prompts are constructed using the essay and the rubric corresponding to each trait. Each trait-specific prompt is then provided to a separate LLM agent dedicated to that trait. This approach, referred to as the LLM-based trait-wise rationale generation system, relies on the LLM's demonstrated ability to effectively evaluate essays, as supported by prior research (Lee et al., 2024; Ho et al., 2022; Li et al., 2023). We have adopted the prompts used in Lee et al. (2024) as a basis for the task description and modified them to fit our context. We have also added trait-specific rubric to them. Our trait-wise LLM agents generate qualitative assessments based on the rubric, producing rationales in a text form. This method enables the generation of detailed, text-based rationales that are directly tied to the rubric, facilitating a subsequent S-LLM to decide the final numeric score in a more accurate manner.

Given that the decoder of the S-LLM used in RMTS predicts subsequent tokens based on previous ones, the rationale was also constructed in a sequence that evaluates sub-dimensional, constituent traits of the essay first (e.g., *Content*, *Organization*, *Style*), followed by the overall trait. This approach capitalizes on the model's capability to boost predictive performance by replicating the sequential nature of human assessors when evaluating traits (Do et al., 2024).

3.2 Representation extraction and scoring

In the current study, we utilize various pre-trained encoder-decoder S-LLMs for scoring multi-traits of essays. We include five widely used models—T5, Flan-T5, BART, Pegasus, and LED (Longformer Encoder-Decoder model) (Raffel et al., 2020; Chung et al., 2024; Lewis et al., 2019; Zhang et al., 2019; Beltagy et al., 2020b)—as the S-LLMs for essay scoring. Figure 3 shows the RMTS architecture. Each component in RMTS framework's essay scoring model corresponds to the respective component of the individual S-LLM.

In RMTS, both the essay and the generated rationales are fed into a single encoder to extract their respective representations, which means that the two texts share a common encoder, allowing their representations to be projected into the same vector space. Inspired by (Do et al., 2024), we add the prompt *"Score the essay of the prompt N"* to the essay text to improve model inference. Special tokens, such as "<*Essay*>" and "<*Rationale*>," are inserted before the essay and rationale to help the tokenizer distinguish between the two. We also introduce tokens for multi-trait names (e.g., <*Content*>) to prevent them from being split into sub-words, preserving their meanings.

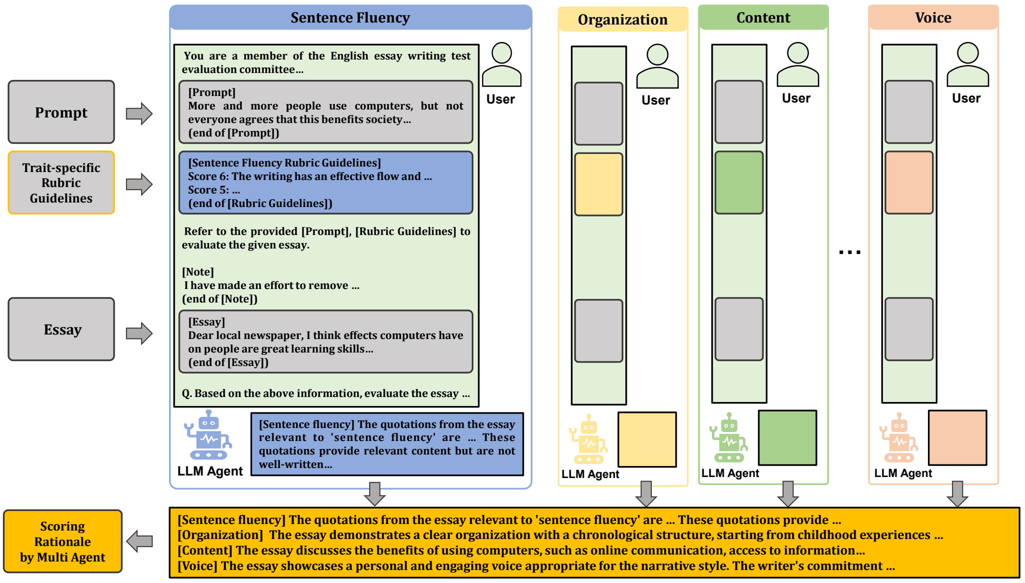
The encoder processes this combined input to generate dense representations, which are integrated into a unified feature vector by a linear layer for scoring. This vector is passed to a decoder, which predicts trait-specific scores. By leveraging both the essay and rationale, the model delivers detailed multi-trait scoring.

3.3 Score extraction

Since we use S-LLMs, which are sequence-to-sequence models, we predict and generate scores for multiple traits alongside their respective names from each essay one at a time, based on techniques from Do et al. (2024). The generated string of scores is transformed into a dictionary format, where trait names serve as keys to extract the scores. For accurate evaluation, we disregard predictions

[1] https://platform.openai.com/docs/models

[2] https://llama.meta.com/responsible-use-guide/



You are a member of the English essay writing test evaluation committee...

[Prompt]
More and more people use computers, but not everyone agrees that this benefits society...
(end of [Prompt])

[Sentence Fluency Rubric Guidelines]
Score 6: The writing has an effective flow and ...
Score 5: ...
(end of [Rubric Guidelines])

Refer to the provided [Prompt], [Rubric Guidelines] to evaluate the given essay.

[Note]
I have made an effort to remove ...
(end of [Note])

[Essay]
Dear local newspaper, I think effects computers have on people are great learning skills...
(end of [Essay])

Q. Based on the above information, evaluate the essay ...

[Sentence fluency] The quotations from the essay relevant to 'sentence fluency' are ... These quotations provide relevant content but are not well-written...

[Sentence fluency] The quotations from the essay relevant to 'sentence fluency' are ... These quotations provide ...
[Organization] The essay demonstrates a clear organization with a chronological structure, starting from childhood experiences ...
[Content] The essay discusses the benefits of using computers, such as online communication, access to information...
[Voice] The essay showcases a personal and engaging voice appropriate for the narrative style. The writer's commitment ...

Figure 2: Trait-specific rationales are constructed using the essay prompt, the essay, and the rubric guidelines corresponding to each trait. To generate the final rationale for each essay, we combine the trait-specific rationales in sequence.

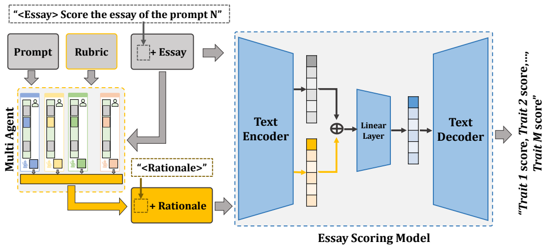


Figure 3: The final rationale generated by multiple LLM agents and the essay are fed into a shared encoder to extract their representations. These representations are then projected to a unified feature vector by a linear layer and passed through the decoder, which predicts trait-specific scores in sequence.

for traits whose ground truth values are *NaN*.

4 Experiment

In this study, we conducted extensive experiments to analyze the generated rationales and evaluate their effectiveness in scoring multiple essay traits, guided by the following research questions.

- **RQ1.** What are the key findings from the analysis of LLM-generated rationales for essay evaluation?

- **RQ2.** To what extent does incorporating rationales improve the reliability of multi-trait essay scoring using S-LLMs?

4.1 Datasets

In our main experiment, we utilized the ASAP[3] and ASAP++[4] (Mathias and Bhattacharyya, 2020) datasets, comprising English essays from American high school students (grades 7–10) across eight prompts. The ASAP dataset provides overall scores for all essays, but only prompts 7 and 8 have trait-specific scores. Thus, we included ASAP++ for rated trait scores on the remaining prompts, and this combined dataset will be referred to as "ASAP/ASAP++" throughout the paper. Additionally, the Feedback Prize dataset[5], which consists of argumentative essays written by American students (grades 6–12) and labeled with six traits, was used without distinguishing between prompts to examine the generalizability of the incremental effect of using essays and rationales together on vanilla S-LLMs. Due to space constraints, the dataset descriptions are provided in Table 1.

4.2 Rationale Analysis

To evaluate rationale quality, we performed various analyses. We evaluated the similarity of the generated rationales using ROUGE-L (Lin, 2004) on a sample of 100 essays to analyze the diversity in how LLMs generate them. Additionally, we measured the faithfulness of LLM-generated

[3] https://www.kaggle.com/competitions/asap-aes/data
[4] https://lwsam.github.io/ASAP++/lrec2018.html
[5] https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data

Dataset	Prompt	# Essays	Traits
ASAP/ ASAP++	1	1785	Over, Cont, WC, Org, SF, Conv
	2	1800	Over, Cont, WC, Org, SF, Conv
	3	1726	Over, Cont, PA, Nar, Lang
	4	1772	Over, Cont, PA, Nar, Lang
	5	1805	Over, Cont, PA, Nar, Lang
	6	1800	Over, Cont, PA, Nar, Lang
	7	1569	Over, Cont, Org, Conv, Style
	8	723	Over, Cont, WC, Org, SF, Conv, Voice
Feedback	-	3930	Coh, Syn, Voca, Phr, Gram, Conv

Table 1: Composition of the ASAP/ASAP++ combined dataset, listing writing traits per prompt. Traits include: Over: *Overall*, Cont: *Content*, WC: *Word Choice*, Org: *Organization*, SF: *Sentence Fluency*, Conv: *Conventions*, PA: *Prompt Adherence*, Nar: *Narrativity*, Lang: *Language*, with Feedback Prize traits being: Coh: *Cohesion*, Syn: *Syntax*, Voca: *Vocabulary*, Phr: *Phraseology*, Gram: *Grammar*, Conv: *Conventions*.

rationales to the predicted multi-trait scores, using a proxy method from prior studies (Wiegreffe et al., 2020; Jain et al., 2020; Li et al., 2023). Specifically, we fine-tuned S-LLMs to predict multi-trait scores using only the rationales as input to the models.

4.3 Baselines

To compare performance across the two datasets described earlier, we used five widely adopted vanilla S-LLMs, all encoder-decoder models (Raffel et al., 2020; Chung et al., 2024; Lewis et al., 2019; Zhang et al., 2019; Beltagy et al., 2020b) designed for text generation tasks. We also included baseline models with a string kernel based model and RNN-based architectures from the referenced papers: HISK (Cozma et al., 2018), STL-LSTM (Dong et al., 2017), MTL-BiLSTM (Kumar et al., 2021), PMAES (Chen et al., 2023), and PLAES (Chen and Li, 2024). These models align with our main task of multi-trait scoring (see Appendix A for details on each baseline). For a fair comparison with the traditional benchmark datasets (ASAP/ASAP++), we used the performance data of four baseline models—HISK (Cozma et al., 2018), STL-LSTM (Dong et al., 2017), MTL-BiLSTM (Kumar et al., 2021), and ArTS (Do et al., 2024)—as reported in (Do et al., 2024), along with two additional baseline models, PMAES (Chen et al., 2023) and PLAES (Chen and Li, 2024), as reported in their original papers. ArTS is a model that employs the vanilla T5-base model for scoring multi-trait essays.

4.4 Experimental Settings

In this study, we employed GPT-3.5-Turbo[6] and Llama-3.1-8B-Instruct[7] for rationale generation based on a prompt-engineering technique, and fine-tuned pre-trained S-LLMs from Huggingface[8]. Using the Seq2SeqTrainer from the same platform, models were trained over 15 epochs with a batch size of 4, and evaluations took place every 5000 steps, applying early stopping with a patience of 2. All experiments were conducted on a single NVIDIA A100 GPU using the PyTorch framework.

4.5 Evaluation

To ensure consistent evaluation, we utilized 5-fold cross-validation across all models, employing a 60/20/20 split for training, validation, and testing, following the methodology of Taghipour and Ng (2016) and Kumar et al. (2021) with the combined ASAP and ASAP++ dataset. For the Feedback Prize dataset, we applied the same 5-fold process but with stratified splitting based on label distribution. Assessment was conducted using quadratic weighted kappa (QWK) (Cohen, 1968), the dataset's designated metric, which effectively measures score disparities between human raters and model predictions. We chose the top two models from each fold and reported the highest QWK as the final result (Do et al., 2024).

5 Results

5.1 Rationale Analysis (RQ1)

We focus on analyzing the rationales from the ASAP/ASAP++ and Feedback Prize datasets in terms of similarity and faithfulness.

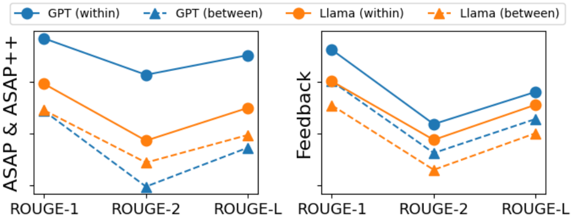


Figure 4: ROUGE scores of rationales *within* the same essay or *between* different essays across GPT and Llama.

[6] https://openai.com/index/openai-api/

[7] https://llama.meta.com/responsible-use-guide/

[8] https://huggingface.co/

	Trait (Prediction Order: ←)											
Model	**Overall**	**Cont**	**PA**	**Lang**	**Nar**	**Org**	**Conv**	**WC**	**SF**	**Style**	**Voice**	**AVG↑ (SD↓)**
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 (0.004)
STL-LSTM	0.750	0.707	0.731	0.640	0.699	0.649	0.505	0.621	0.612	0.609	0.544	0.642 (0.073)
MTL-BiLSTM	**0.764**	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.639 (0.057)
PMAES	0.671	0.567	0.584	0.545	0.614	0.481	0.421	0.584	0.582	-	-	0.614 (-)
PLAES	0.673	0.574	0.601	0.554	0.631	0.491	0.447	0.579	0.580	-	-	0.631 (-)
T5 (ArTS)	0.754	0.730	0.751	0.698	0.725	0.672	0.668	0.679	0.678	**0.721**	0.570	0.695 (0.018)
+ RMTS(G) (+%)	0.755 (+0.1)	**0.737** (+0.7)	**0.752** (+0.1)	**0.713** (+1.5)	**0.744** (+1.9)	0.682 (+1.0)	**0.690** (+2.2)	**0.705** (+2.6)	**0.694** (+1.6)	0.702 (-1.9)	0.612 (+4.2)	**0.708** (0.043)
+ RMTS(L) (+%)	0.754 (+0.0)	0.730 (+0.0)	0.749 (-0.2)	0.701 (+0.3)	0.737 (+1.2)	0.675 (+0.3)	0.684 (+1.6)	0.690 (+1.1)	0.684 (+0.6)	0.696 (-2.5)	0.640 (+7.0)	0.704 (0.042)
Flan-T5	0.662	0.645	0.615	0.539	0.577	0.646	0.636	0.694	0.667	0.578	0.624	0.626 (0.064)
+ RMTS(G) (+%)	0.732 (+7.0)	0.733 (+8.8)	0.750 (+13.5)	0.708 (+16.9)	0.737 (+16.0)	**0.684** (+3.8)	0.680 (+4.4)	0.691 (-0.3)	0.680 (+1.3)	0.688 (+11.0)	0.563 (-6.1)	0.695 (0.048)
+ RMTS(L) (+%)	0.723 (+6.1)	0.717 (+7.2)	0.736 (+12.1)	0.696 (+15.7)	0.722 (+14.5)	0.663 (+1.7)	0.662 (+2.6)	0.673 (-2.1)	0.663 (-0.4)	0.695 (+11.7)	0.620 (-0.4)	0.688 (0.054)
BART	0.701	0.672	0.711	0.664	0.705	0.600	0.588	0.624	0.601	0.646	0.547	0.642 (0.054)
+ RMTS(G) (+%)	0.720 (+1.9)	0.710 (+3.8)	0.731 (+2.0)	0.683 (+1.9)	0.720 (+1.5)	0.651 (+5.1)	0.637 (+4.9)	0.685 (+6.1)	0.655 (+5.4)	0.661 (+1.5)	**0.649** (+10.2)	0.674 (0.046)
+ RMTS(L) (+%)	0.724 (+2.3)	0.704 (+3.2)	0.732 (+2.1)	0.677 (+1.3)	0.714 (+0.9)	0.658 (+5.8)	0.647 (+5.9)	0.671 (+4.7)	0.662 (+6.1)	0.673 (+2.7)	0.596 (+4.9)	0.678 (0.037)
Pegasus	0.536	0.584	0.608	0.586	0.629	0.578	0.515	0.559	0.519	0.578	0.388	0.553 (0.065)
+ RMTS(G) (+%)	0.711 (+17.5)	0.651 (+6.7)	0.698 (+9.0)	0.674 (+8.8)	0.697 (+6.8)	0.615 (+3.7)	0.600 (+8.5)	0.618 (+5.9)	0.613 (+9.4)	0.619 (+4.1)	0.561 (+17.3)	0.641 (0.046)
+ RMTS(L) (+%)	0.713 (+17.7)	0.650 (+6.6)	0.698 (+9.0)	0.667 (+8.1)	0.699 (+7.0)	0.624 (+4.6)	0.605 (+9.0)	0.640 (+8.1)	0.626 (+10.7)	0.638 (+6.0)	0.570 (+18.2)	0.648 (0.041)
LED	0.709	0.677	0.706	0.666	0.707	0.627	0.633	0.643	0.640	0.655	0.522	0.653 (0.053)
+ RMTS(G) (+%)	0.736 (+2.7)	0.714 (+3.7)	0.733 (+2.7)	0.688 (+2.2)	0.719 (+1.2)	0.667 (+4.0)	0.663 (+3.0)	0.676 (+3.3)	0.674 (+3.4)	0.694 (+3.9)	0.597 (+7.5)	0.687 (0.038)
+ RMTS(L) (+%)	0.727 (+1.8)	0.711 (+3.4)	0.741 (+3.5)	0.674 (+0.8)	0.714 (+0.7)	0.656 (+2.9)	0.648 (+1.5)	0.658 (+1.5)	0.644 (+0.4)	0.684 (+2.9)	0.542 (+2.0)	0.673 (0.052)

Table 2: Average QWK scores across all prompts for each trait on the ASAP/ASAP++ datasets. The values in parentheses (%) represent the percentage of improvement in RMTS performance when incorporating rationales generated by GPT (G) or Llama (L) compared to the vanilla S-LLMs. Traits are predicted from right to left (←), and five-fold averaged standard deviation is reported (SD). The best results are highlighted in **bold**, and the second-best results are underlined. For PMAES and PLAES, style and voice traits, where contrastive learning is not feasible due to their presence in a single prompt, are marked as "-".

5.1.1 Similarity of rationales

Figure 4 displays the similarity analysis results for rationales generated by two LLMs, presenting ROUGE scores for the ASAP/ASAP++ and Feedback Prize datasets. To evaluate consistency, we calculated ROUGE scores between rationales generated for the same essay across five iterations and averaged them, labeled as *"within"*. In RMTS, we used the first rationale generated from the five iterations. To gauge diversity, we computed ROUGE scores between the rationales used in RMTS for different essays, also averaged and labeled as *"between"*. As shown, *"within"* scores are higher than *"between"* scores, indicating each LLM captures and reflects the unique characteristics of each essay in its rationale while maintaining consistent features within the same essay.

5.1.2 Faithfulness of rationales

Figures 5 compares the reliability (measured by QWK against human-labeled scores) of each model in predicting essay scores for the ASAP/ASAP++ dataset, using either the essays or the LLM-generated rationales. Most models performed at over 80% of their essay-only performance in nearly entire traits when using the rationales, demonstrating that rationales make a meaningful contribution to S-LLMs' essay evaluations. Given that these rationales are qualitative free-text outputs, this also indicates that they provide partial explanations for the models' score predictions (Wiegreffe et al., 2020; Jain et al., 2020; Li et al., 2023). However, as shown in Figure 5, Pegasus achieved about 40% of its essay-only performance with GPT-generated rationales and 50% with Llama-generated rationales in average, suggesting that it relies more on the intrinsic features of essays than on qualitative evaluation data.

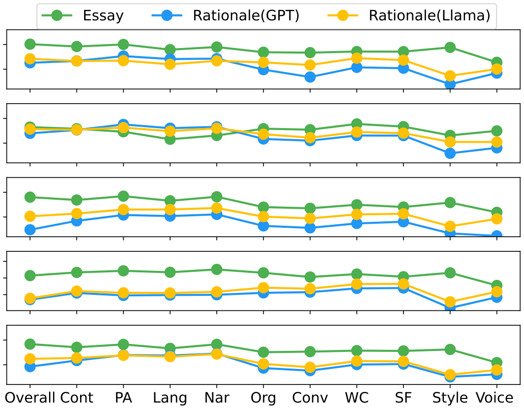


Figure 5: Performance comparison of S-LLMs based on QWK scores, averaged across all prompts for each trait with regard to the ASAP/ASAP++ dataset, using either the essays or the rationales generated by GPT or Llama.

	Prompt								
Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**AVG↑(SD↓)**
HISK	0.674	0.586	0.651	0.681	0.693	0.709	0.641	0.516	0.644 (0.004)
STL-LSTM	0.690	0.622	0.663	0.729	0.719	0.753	0.704	0.592	0.684 (0.055)
MTL-BiLSTM	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665 (0.048)
PMAES	0.656	0.553	0.598	0.606	0.626	0.572	0.386	0.530	0.566 (-)
PLAES	0.648	0.563	0.604	0.623	0.634	0.593	0.403	0.533	0.575 (-)
T5 (ArTS)	0.708	**0.706**	0.704	0.767	0.723	**0.776**	**0.749**	0.603	0.717 (0.025)
+ RMTS(G)(+%)	**0.716**(+0.8)	0.704(-0.2)	**0.723**(+1.9)	**0.772**(+0.5)	**0.737**(+1.4)	0.769(-0.7)	0.736(-1.3)	0.651(+4.8)	**0.726** (0.042)
+ RMTS(L)(+%)	0.705(-0.3)	0.692(-1.4)	0.714(+1.0)	0.766(-0.1)	0.726(+0.3)	0.773(-0.3)	0.726(-2.3)	**0.658**(+5.5)	0.720 (0.044)
Flan-T5	0.703	0.691	0.523	0.599	0.593	0.674	0.609	0.633	0.628 (0.056)
+ RMTS(G)(+%)	0.711(0.8)	0.666(-2.5)	**0.723**(+20.0)	0.771(+17.2)	0.736(+14.3)	0.762(+8.8)	0.723(+11.4)	0.642(0.9)	0.717 (0.055)
+ RMTS(L)(+%)	0.700(-0.3)	0.643(-4.8)	0.702(+17.9)	0.761(+16.2)	0.719(+12.6)	0.751(+7.7)	0.734(+12.5)	0.623(-0.1)	0.704 (0.055)
BART	0.647	0.602	0.658	0.727	0.713	0.713	0.624	0.534	0.652 (0.066)
+ RMTS(G)(+%)	0.707(+6.0)	0.667(+6.5)	0.702(+4.4)	0.751(+2.4)	0.718(+0.5)	0.737(+2.4)	0.684(+6.0)	0.595(+6.1)	0.695 (0.045)
+ RMTS(L)(+%)	0.698(+5.1)	0.658(+5.6)	0.691(+3.3)	0.744(+1.7)	0.720(+0.7)	0.748(+3.5)	0.690(+6.6)	0.614(+8.0)	0.695 (0.042)
Pegasus	0.639	0.520	0.518	0.562	0.636	0.597	0.539	0.478	0.561 (0.058)
+ RMTS(G)(+%)	0.672(+3.3)	0.631(+11.1)	0.683(+16.5)	0.725(+16.3)	0.718(+8.2)	0.695(+9.8)	0.593(+5.4)	0.573(+9.5)	0.661 (0.057)
+ RMTS(L)(+%)	0.670(+3.1)	0.637(+11.7)	0.679(+16.1)	0.714(+15.2)	0.708(+7.2)	0.714(+11.7)	0.611(+7.2)	0.587(+10.9)	0.665 (0.046)
LED	0.704	0.650	0.679	0.705	0.701	0.707	0.638	0.520	0.663 (0.064)
+ RMTS(G)(+%)	0.701(-0.3)	0.684(+3.4)	0.693(+1.4)	0.762(+5.7)	0.721(+2.0)	0.742(+3.5)	0.715(+7.7)	0.620(+10.0)	0.705 (0.004)
+ RMTS(L)(+%)	0.694(-1.0)	0.654(+0.4)	0.688(+0.9)	0.754(+4.9)	0.724(+2.3)	0.745(+3.8)	0.714(+7.6)	0.592(+7.2)	0.696 (0.049)

Table 3: Average QWK scores across all traits for each prompt on the ASAP/ASAP++ datasets. The values in parentheses (%) represent the percentage of improvement in RMTS performance when incorporating rationales generated by GPT (G) or Llama (L) compared to the vanilla S-LLMs. Five-fold averaged standard deviation is reported (SD). The best results are highlighted in **bold**, and the second-best results are underlined.

5.2 Performance Comparison (RQ2)

To address the second research question, we first evaluate RMTS against the baseline models and vanilla S-LLMs on the ASAP/ASAP++ dataset, the standard for essay scoring. We then assess RMTS against vanilla S-LLMs on the Feedback Prize dataset to demonstrate its broader applicability. Finally, to gain deeper insights into the role of rationales, we fine-tuned three S-LLMs with rationales and essays, removing one trait at a time during the process to observe performance variations.

5.2.1 Performance with ASAP/ASAP++

Since our target is to predict individual trait scores, we will focus on trait scoring rather than *Overall* scores. Owing to space constraints, the results are presented in Table 2 and 3. Table 2 shows model performance on the ASAP/ASAP++ dataset. Using GPT-generated rationales, RMTS applied to each of the five S-LLMs outperforms their respective vanilla versions across nearly all traits, except for *Style* in T5 and *Word Choice* and *Voice* in Flan-T5. T5 model shows incremental improvements with rationales, ranking first or second in every trait, including *Overall*. Additionally, RMTS with T5, BART, and LED outperforms the best traditional models—HISK, STL-LSTM, and MTL-BiLSTM—in every trait using GPT rationales. Although MTL-BiLSTM has a higher *Overall* score, the gap with RMTS-T5 is small, and RMTS focuses more on individual trait scoring (Additional performance comparisons with baselines specifically designed for overall score assessment are provided in Appendix D.2).

On top of that, Table 3 shows prompt-wise performance. RMTS improves the vanilla S-LLMs for the majority of prompts. With GPT-generated rationales, RMTS using T5, Flan-T5, and LED generally outperforms their vanilla counterparts across most prompts, and other S-LLMs show improvements with the same rationales in every prompt. Additionally, RMTS with T5 and Flan-T5 using GPT rationales outperforms the best traditional baseline models.

5.2.2 Performance with Feedback Prize Dataset

To assess the broader applicability of using rationales to enhance S-LLMs in essay scoring, we conducted additional experiments with the Feedback Prize dataset, as shown in Table 4. Despite the small dataset size of 2.3K samples—about one-third of the ASAP/ASAP++ dataset—rationales improve the performance of the four S-LLMs (T5, BART, Pegasus, and LED) across most traits. However, integrating rationales does not improve the vanilla Flan-T5 model. We attribute this to the model's inherent characteristics from instructionfine-tuning (Chung et al., 2024), which may prevent it from effectively incorporating rationales with such a small dataset. Nevertheless, these results indicate that rationales generally enhance model performance, even in data-scarce environments.

	Trait (Prediction Order: ←)						
Model	**Conv**	**Gram**	**Phr**	**Voc**	**Syn**	**Coh**	**AVG↓ (SD↑)**
HISK	0.279	0.279	0.270	0.261	0.252	0.241	0.264(0.012)
STL	0.544	0.440	0.534	0.515	0.518	0.459	0.502(0.024)
MTL-BiLSTM	0.527	0.484	0.505	0.519	0.507	0.462	0.501(0.027)
T5	0.521	0.479	0.512	0.454	0.497	0.467	0.488(0.027)
+ RMTS(G) (+%)	0.568 (+4.7)	0.550 (+7.1)	0.543 (+3.1)	0.430 (-2.4)	0.543 (+4.6)	0.498 (+3.1)	0.522 (0.024)
+ RMTS(L) (+%)	0.570 (+4.9)	**0.557** (+7.8)	0.535 (+2.3)	0.443 (-1.1)	0.534 (+3.7)	0.490 (+2.3)	0.522 (0.024)
Flan-T5	0.539	0.512	0.527	**0.466**	0.531	0.491	0.511 (0.025)
+ RMTS(G) (+%)	0.520(-1.9)	0.507(-0.5)	0.497(-3.0)	0.440(-2.6)	0.513(-1.8)	0.472(-1.9)	0.492 (0.034)
+ RMTS(L) (+%)	0.479(-6.0)	0.476(-3.6)	0.513(-1.4)	0.407(-5.9)	0.496(-3.5)	0.477(-1.4)	0.475 (0.123)
BART	0.396	0.357	0.440	0.363	0.288	0.314	0.360 (0.050)
+ RMTS(G) (+%)	0.565 (+16.9)	0.477 (+12.0)	**0.596** (+15.6)	0.461 (+9.8)	0.507 (+21.9)	**0.573** (+25.9)	0.530 (0.051)
+ RMTS(L) (+%)	0.439 (+4.3)	0.410 (+5.3)	0.366 (-7.4)	0.329 (-3.4)	0.341 (+5.3)	0.172 (-14.2)	0.343 (0.085)
Pegasus	0.273	0.233	0.265	0.304	0.270	0.264	0.268 (0.021)
+ RMTS(G) (+%)	0.290 (+1.7)	0.237 (+0.4)	0.309 (+4.4)	0.315 (+1.1)	0.313 (+4.3)	0.334 (+7.0)	0.299 (0.031)
+ RMTS(L) (+%)	0.327 (+5.4)	0.273 (+4.0)	0.359 (+9.4)	0.363 (+5.9)	0.350 (+8.0)	0.369 (+10.5)	0.340 (0.033)
LED	0.520	0.479	0.486	0.428	0.505	0.476	0.482 (0.029)
+ RMTS(G) (+%)	**0.586** (+6.6)	0.552 (+7.3)	0.540 (+5.4)	0.462 (+3.4)	**0.550** (+4.5)	0.505 (+2.9)	**0.533** (0.039)
+ RMTS(L) (+%)	0.565 (+4.5)	0.531 (+5.2)	0.507 (+2.1)	0.428 (+0.0)	0.520 (+1.5)	0.461 (-1.5)	0.502 (0.045)

Table 4: Average QWK scores across all prompts for each trait on the Feedback Prize dataset. The values in parentheses (%) represent the percentage of improvement in RMTS performance when incorporating rationales generated by GPT (G) or Llama (L) compared to the vanilla S-LLMs. Traits are predicted from right to left (←), and five-fold averaged standard deviation is reported (SD). The best results are highlighted in **bold**, and the second-best results are underlined. PMAES and PLAES, which utilize prompt-wise contrastive learning, were excluded as they cannot be trained on datasets where all essays share a single prompt.

5.3 Trait Rationale Ablation Study (RQ2)

To assess the effectiveness of trait-specific rationales, we conducted an ablation study by removing a trait commonly shared across prompts, as illustrated in Figure 6. We focused on traits present in at least four prompts. Specifically, *Content* is evaluated in all prompts, while *Organization* and *Conventions* are scored in Prompts 1, 2, 7, and 8. *Prompt Adherence*, *Language*, and *Narrativity* are assessed in Prompts 3 to 6. For comparison, we divided the prompts into two groups and excluded *Word Choice*, *Sentence Fluency*, *Style*, and *Voice* since they appear in fewer prompts. We used T5, Flan-T5, and BART in the experiments, fine-tuning each model by removing one trait at a time.

A consistent decline in performance is observed across all traits when the corresponding rationale is removed, confirming that trait rationales significantly influence their respective assessments. For instance, removing the rationale for *Conventions* results in a performance drop for that trait, particularly when compared to RMTS, which utilizes rationales for all traits. Although the performance of models without a trait rationale lag behind RMTS, it still outperforms vanilla models without any rationale input. This suggests that trait rationales not only influence their own assessments but also interact with and affect the evaluation of other traits (Canale and Swain, 1980). For example, when the rationale for *Conventions* is removed, performance still surpass that of vanilla models.

Interestingly, performance does not always drop significantly when certain trait rationales are removed, particularly in the case of T5 when *Organization* and *Prompt Adherence* rationales are excluded. This implies that the effectiveness of rationales can vary by trait and model.

Overall, the results show that rationales generally have an incremental effect on performance. Essay scoring performs worse when trait rationales are removed, highlighting the crucial role these rationales play in predicting trait scores.

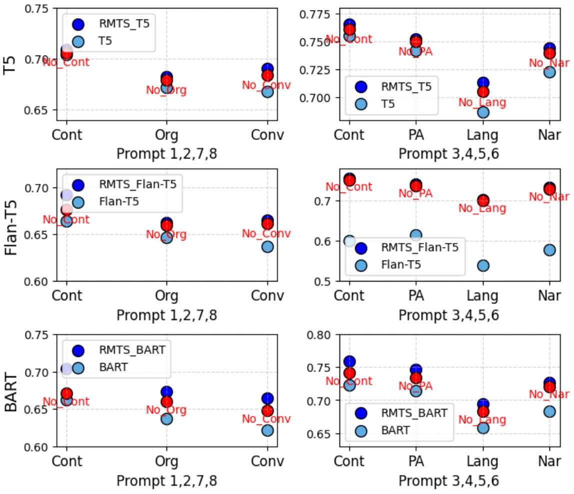


Figure 6: Ablation study of rationale when removing each of the trait in **(A) Prompt 1,2,7,8** and **(B) Prompt 3,4,5,6**. The performances generally drop when any one trait is omitted, underscoring the importance of incorporating all traits in rationale generation.

6 Conclusion

This paper introduces RMTS, a framework that uses prompt-engineering-based LLMs to improve multi-trait essay scoring in S-LLMs by generating trait-specific rationales aligned with rubric guidelines and incorporating them into the scoring process. Our results show that RMTS with S-LLMs significantly improves the performance of each vanilla model, with RMTS using T5 even outperforming state-of-the-art baselines. Additionally, removing rationales negatively impacts performance. These study findings highlight the substantial benefits of utilizing trait-specific rationales generated by LLMs, which has been untapped by prior research. From this view point, RMTS can be seen as opening up new horizons for automated essay scoring with S-LLMs.

Limitations

In this study, we have identified two primary limitations. First, our model's performance could be affected by the sequence order of traits due to the use of autoregressive models like T5 and BART. Future research should explore models like XLNet (Yang et al., 2019), which are better suited for handling sequence orders. Secondly, we focused exclusively on multi-trait scoring of English writing. To evaluate the scalability of our model for general language education, further studies are needed on languages other than English.

Ethical Statement

This study utilizes only publicly available benchmark datasets, including ASAP, ASAP++, and Feedback Prize.

Acknowledgements

This research was supported by Chunjae Education Inc. (Project Number G01250064).

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020a. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020b. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Michael Canale and Merrill Swain. 1980. Theoretical bases of com-municative approaches to second language teaching and testing. *Applied linguistics*, 1(1):1–47.

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.

Yufei Chen, Yuhan Zhang, Yiming Liu, Yang Liu, and Bin Wang. 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*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Mădălina Cozma, Andrei M Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. *arXiv preprint arXiv:1804.07954*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt-and trait relation-aware cross-prompt essay trait scoring. *arXiv preprint arXiv:2305.16826*.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2024. Autoregressive score generation for multi-trait essay scoring. *Preprint*, arXiv:2403.08332.

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.

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.

Liz Freeman and Andy Miller. 2001. Norm-referenced, criterion-referenced, and dynamic assessment: what exactly is the point? *Educational Psychology in Practice*, 17(1):3–16.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. *arXiv preprint arXiv:2005.00115*.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Many hands make light work: using essay traits to automatically score essays. *arXiv preprint arXiv:2102.00781*.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Prompting large language models for zero-shot essay scoring via multi-trait specialization. *arXiv preprint arXiv:2404.04941*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling chatgpt for explainable automated student answer assessment. *arXiv preprint arXiv:2305.12962*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? *In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.

OpenAI. 2022. Chatgpt.

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.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762*.

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.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020a. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020b. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Preprint*, arXiv:1912.08777.

Appendix

A Details of the baselines

- HISK (Cozma et al., 2018): is a string kernel based on histogram intersection, used in combination with a support vector regressor.

- STL-LSTM (Dong et al., 2017)[9]: uses a combination of LSTM and CNN to infer essay scores of every trait individually.

- MTL-BiLSTM (Kumar et al., 2021)[10]: employs trait-specific BiLSTM layers to score multi-trait, ultimately predicting the overall score.

- R2BERT (Yang et al., 2020b): enhances AES by fine-tuning pre-trained language models, combining regression and ranking objectives to improve scoring accuracy.

- NPCR (Xie et al., 2022)[11]: employs pairwise contrastive regression to learn relative score differences between essays, thereby refining the scoring process.

- PMAES (Chen et al., 2023)[12]: utilizes prompt-mapping contrastive learning to generalize scoring across various prompts, enhancing the model's adaptability.

- PLAES (Chen and Li, 2024): introduces a prompt-generalized and level-aware learning framework, improving cross-prompt AES performance by considering prompt variations and essay complexity levels.

- T5[13] (Raffel et al., 2020): is a transformer-based model that frames NLP tasks as a text-to-text problem. In this study, we used the "google-t5/t5-base" model.

- Flan-T5[14] (Chung et al., 2024): builds upon T5 (Raffel et al., 2020) by introducing fine-tuning on instruction-based datasets. In this study, we used the "google/flan-t5-base" model.

- BART[15] (Lewis et al., 2019): is a sequence-to-sequence models trained by corrupting text and learning to reconstruct the original text. In this study, we used the "facebook/bart-base" model.

- Pegasus[16] (Zhang et al., 2019): is designed specifically for abstractive summarization tasks, focusing on predicting whole sentences that have been masked. In this study, we used "google/pegasus-x-base" model.

- LED[17] (Beltagy et al., 2020b): extends the transformer architecture to handle longer documents efficiently by using sparse attention mechanisms. In this study, we used "allenai/led-large-16384" model.

B Length Statistics of Rationales

The generated rationales were tokenized using each model's corresponding tokenizer. As shown in Figure 7, aside from the rationales from the Feedback dataset generated by Llama (which reached a maximum of 586 tokens when tokenized by the T5 tokenizer), the maximum number of tokens in the rationales produced by either GPT or Llama did not exceed 512. This is the typical limit that transformer-based language models can process. This suggests that the rationale lengths are manageable and should not impede the models' ability to capture contextual information. For Llama-generated rationales in the Feedback Prize dataset, any rationales exceeding 512 tokens were truncated to comply with the limit. Interestingly, Llama tended to generate longer rationales than GPT in the Feedback dataset.

C Faithfulness of Rationales from ASAP/ASAP++ and Feedback Prize Datasets

This section presents the faithfulness analysis of rationales generated by LLMs on the ASAP/ASAP++ and Feedback Prize datasets. We compare model performance using Quadratic Weighted Kappa (QWK) scores when scoring essays versus LLM-generated rationales.

[9] https://github.com/feidong1991/aes.git

[10] https://github.com/ASAP-AEG/MTL-Essay-Traits-Scoring.git

[11] https://github.com/CarryCKW/AES-NPCR.git

[12] https://github.com/gdufsnlp/PMAES.git

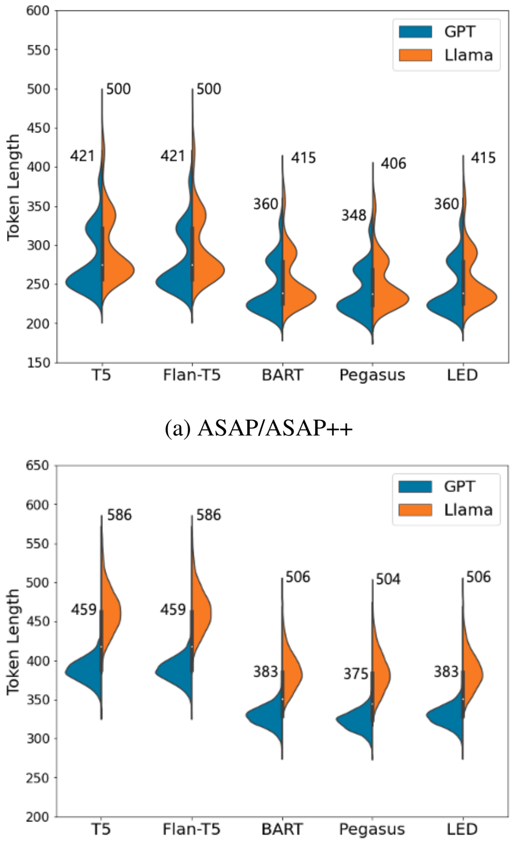
[13] https://huggingface.co/google-t5/t5-base

[14] https://huggingface.co/google/flan-t5-base

[15] https://huggingface.co/facebook/bart-base

[16] https://huggingface.co/google/pegasus-x-base

[17] https://huggingface.co/allenai/led-large-16384



(a) ASAP/ASAP++

(b) Feedback

Figure 7: Violin plots of the number of tokens per rationale depending on each individual model's tokenizer. (A) refers to the rationale from the ASAP/ASAP++ dataset and (B) refers to the rationale from the Feedback Prize dataset.

C.1 ASAP/ASAP++ Dataset

Figure 8 shows the QWK performance of five S-LLMs on the ASAP/ASAP++ dataset using only LLM-generated rationales.

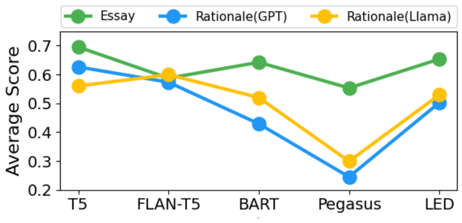


Figure 8: Performance comparison based on QWK scores, averaged across all traits, for the ASAP/ASAP++ dataset, using either the essays or the rationales generated by GPT or Llama.

C.2 Feedback Prize Dataset

Figures 10 and 9 illustrate the model performance in QWK on the Feedback Prize dataset using only LLM-generated rationales.

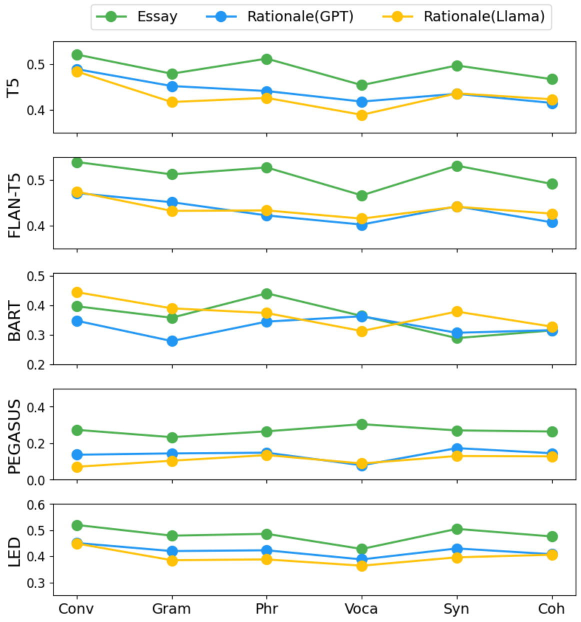


Figure 9: Performance comparison of each model based on QWK scores, averaged across all prompts for each trait with regard to the Feedback Prize dataset, using either the essays or the rationales generated by GPT or Llama.

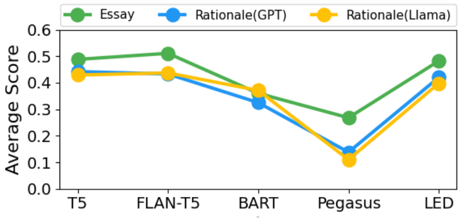


Figure 10: Performance comparison based on QWK scores, averaged across all traits, with regard to the Feedback Prize dataset using either the essays or the rationales generated by GPT or Llama.

D Additional Experiments

D.1 Result of Compressed Rationales

As shown in Appendix B, some rationales were relatively long and could have hindered S-LLMs' scoring performance. To address this, we designed a prompt to guide the LLM in removing redundant information, making the rationales more concise while retaining key details.

An analysis of the ASAP dataset showed that the average rationale length was significantly higher for the original rationales than for the shortened versions across all tested models. For instance, in

Table 5: Performance of Compressed ASAP Rationale. The models are abbreviated as follows: G = GPT-based original rationale, G+C = GPT-based rationale with compression, L = Llama-based original rationale, L+C = Llama-based rationale with compression.

Model	Overall	Con	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG
T5 RMTS(G)	0.755	0.737	0.752	0.713	0.744	0.682	0.690	0.705	0.694	0.702	0.612	0.708
T5 RMTS(G+C)	0.754	0.733	0.757	0.703	0.740	0.680	0.696	0.691	0.688	0.697	0.604	0.704
T5 RMTS(L)	0.754	0.730	0.749	0.701	0.737	0.675	0.684	0.690	0.684	0.696	0.640	0.704
T5 RMTS(L+C)	0.752	0.722	0.746	0.694	0.734	0.672	0.679	0.682	0.682	0.692	0.621	0.698

Table 6: Performance of Compressed Feedback Rationale. The models are abbreviated as follows: G = GPT-based, G+C = GPT-based with Compression, L = Llama-based, L+C = Llama-based with Compression.

Model	Conv	Gram	Phr	Voc	Syn	Coh	Avg
T5 RMTS (G)	0.568	0.550	0.543	0.430	0.543	0.498	0.522
T5 RMTS (G+C)	0.521	0.479	0.512	0.414	0.497	0.467	0.482
T5 RMTS (L)	570	0.557	0.535	0.443	0.534	0.490	0.522
T5 RMTS (L+C)	0.553	0.521	0.527	0.412	0.515	0.479	0.501

the GPT-T5 setting, the original rationales averaged 277.55 words, with a maximum of 431 words, whereas the compressed versions averaged 133.04 words, with a maximum of 228 words. Similarly, in the Llama-T5 setting, original rationales averaged 298.62 words (max 500), compared to 143.35 words (max 238) for the shortened versions.

To assess the impact of rationale length, we compared T5's performance using original versus shortened rationales. As shown in Table 5 and Table 6, models using original rationales generally outperformed those using compressed versions on average. This suggests that the original rationales provided richer information that improved the model's ability to score essays reliably.

D.2 Comparison of *Overall* Score Prediction

While RMTS is inherently designed for multi-trait scoring, we adapted it to generate a single *overall* score for direct comparison with models that do not support multi-trait scoring.

Table 7 shows the QWK performance of our model compared to two baselines designed for *overall* score prediction: R2Bert (Yang et al., 2020b) and NPCR (Xie et al., 2022). T5 RMTS (GPT-3.5) achieves the highest average QWK score of 0.758, outperforming NPCR, the best-performing baseline, by a significant margin (7.7%).

Additionally, RMTS maintains strong performance across different prompts, demonstrating its generalization ability. These results confirm that RMTS excels not only in multi-trait scoring but also in *overall* score prediction.

Table 7: Overall QWK Score Comparison Across Models

Model	1	2	3	4	5	6	7	8	AVG
R2Bert	0.627	0.533	0.517	0.757	0.610	0.685	0.683	0.428	0.605
NPCR	0.719	0.586	0.650	0.751	0.766	0.753	0.683	0.543	0.681
T5 RMTS(G)	**0.768**	**0.655**	**0.712**	**0.810**	**0.810**	**0.808**	**0.792**	**0.710**	**0.758**

D.3 Ablation study of Feedback Prize Dataset

To assess the impact of trait-specific rationales, we conducted an additional experiment on the Feedback dataset (Figure 11). Using T5, Flan-T5, and BART, we fine-tuned each model following the same procedure described in Section 5.3. The results show a consistent performance drop for a given trait when its corresponding rationale is removed, confirming its importance in trait-specific assessment. Notably, models without a specific trait rationale still outperform baseline models without rationales, suggesting that rationales contribute beyond their assigned trait evaluation. However, in some cases, removing a trait rationale improved scoring performance on Flan-T5, indicating that rationale integration does not enhance its performance on the Feedback dataset, likely due to its small size, as discussed in Section 5.2.2. Although the extent of performance degradation varies across traits and models, the overall trend confirms that trait-specific rationales improve scoring accuracy, as their removal generally weakens performance.

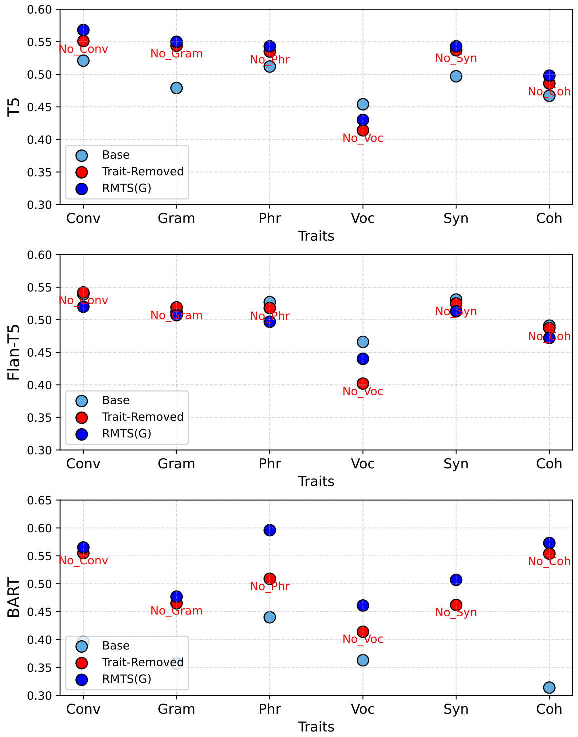


Figure 11: Ablation study on the Feedback dataset when removing each trait's rationale.

E LLM Settings

For RMTS, we used GPT-3.5-Turbo and LLama3.1-8B-Instruct provided by OpenAI and Meta. GPT-3.5-Turbo was used in the form of API[18], and LLama 3.1-8B-Instruct was employed by utilizing the official code shared by Meta[19]. Regarding GPT, we performed the experiments with *gpt-3.5-turbo-0125*. When this study was conducted, the cost for processing input tokens with the model was $0.5 per 1M tokens, while generating output tokens was priced at $1.5 per 1M tokens. We consistently used identical hyperparameters: a temperature of 0, frequency and presence penalties both set to 0, and a Top-p value of 1 for the cumulative probability cutoff in nucleus sampling. Given that the temperature hyperparameter is set to 0, we conducted the experiment a single time. For prompts 3 to 6 of ASAP++, excerpts were excluded in both LLMs.

F Prompts and System Message

Examples of prompts and system messages used by the LLMs to generate rationales can be found in Appendix G. We revised and supplemented (Lee et al., 2024) by adding trait-specific rubric and additional prompts to generate rationales, designing a comprehensive prompt template.

[18] https://openai.com/index/openai-api/

[19] https://llama.meta.com/responsible-use-guide/

G Examples of System Messages and Predefined Template

G.1 System Message

The system message corresponding to each agent used in our experiment are as follows.
System message: You are a member of the English essay writing test evaluation committee. Please, evaluate given essay using following information.

G.2 Predefined Template (*Prompt 1*, *Content*)

[Prompt]
More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.
(end of [Prompt])
[Trait-Specific Rubric Guidelines]
This property checks for the amount of content and ideas present in the essay.
Score 6: The writing is exceptionally clear, focused, and interesting. It holds the reader's attention throughout. Main ideas stand out and are developed by strong support and rich details suitable to audience and purpose. The writing is characterized by
• clarity, focus, and control.
• main idea(s) that stand out.
• supporting, relevant, carefully selected details; when appropriate, use of resources provides strong, accurate, credible support.
• a thorough, balanced, in-depth explanation / exploration of the topic; the writing makes connections and shares insights.
• content and selected details that are well-suited to audience and purpose.

Score 5: The writing is clear, focused and interesting. It holds the reader's attention. Main ideas stand out and are developed by supporting details suitable to audience and purpose. The writing is characterized by
• clarity, focus, and control.
• main idea(s) that stand out.
• supporting, relevant, carefully selected details; when appropriate, use of resources provides strong, accurate, credible support.
• a thorough, balanced explanation / exploration of the topic; the writing makes connections and shares insights.
• content and selected details that are well-suited to audience and purpose.

Score 4: The writing is clear and focused. The reader can easily understand the main ideas. Support is present, although it may be limited or rather general. The writing is characterized by
• an easily identifiable purpose.
• clear main idea(s).
• supporting details that are relevant, but may be overly general or limited in places; when appropriate, resources are used to provide accurate support.
• a topic that is explored / explained, although developmental details may occasionally be out of balance with the main idea(s); some connections and insights may be present.
• content and selected details that are relevant, but perhaps not consistently well-chosen for audience and purpose.

Score 3: The reader can understand the main ideas, although they may be overly broad or simplistic, and the results may not be effective. Supporting detail is often limited, insubstantial, overly general, or occasionally slightly off-topic. The writing is characterized by
• an easily identifiable purpose and main idea(s).
• predictable or overly-obvious main ideas; or points that echo observations heard elsewhere; or a close retelling of another work.
• support that is attempted, but developmental details are often limited, uneven, somewhat off-topic, predictable, or too general (e.g., a list of underdeveloped points).
• details that may not be well-grounded in credible resources; they may be based on clichés, stereotypes or questionable sources of information.
• difficulties when moving from general observations to specifics.

Score 2: Main ideas and purpose are somewhat unclear or development is attempted but minimal. The writing is characterized by
• a purpose and main idea(s) that may require extensive inferences by the reader.
• minimal development; insufficient details.
• irrelevant details that clutter the text.
• extensive repetition of detail.

Score 1: The writing lacks a central idea or purpose. The writing is characterized by
• ideas that are extremely limited or simply unclear.
• attempts at development that are minimal or nonexistent; the paper is too short to demonstrate the development of an idea.
(end of [Trait-Specific Rubric Guidelines])

Refer to the provided [Prompt], and [Trait-Specific Rubric Guidelines] to evaluate the given essay.

[Note]
I have made an effort to remove personally identifying information from the essays using the Named Entity Recognizer (NER). The relevant entities are identified in the text and then replaced with a string such as "@PERSON", "@ORGANIZATION", "@LOCATION", "@DATE", "@TIME", "@MONEY", "@PERCENT", "@CAPS" (any capitalized word) and "@NUM" (any digits). Please do not penalize the essay because of the anonymizations.
(end of [Note])

[Essay]
{essay}
(end of [Essay])

Q. List the quotations from the [Essay] that are relevant to "content" and evaluate whether each quotation is well-written or not. Provide a literal evaluation of the [Essay], returning text-based feedback rather than scores. Refrain from offering additional commentary. Please keep your response within 50 words.

H Examples of Rationales

H.1 A Rationale of an Essay for Prompt 1 of ASAP++

- **Essay:** Dear @CAPS1 @CAPS2, I believe that using computers will benefit us in many ways like talking and becoming friends will others through websites like facebook and mysace. Using computers can help us find coordibates, locations, and able ourselfs to millions of information. Also computers will benefit us by helping with jobs as in planning a house plan and typing a @NUM1 page report for one of our jobs in less than writing it. Now lets go into the wonder world of technology. Using a computer will help us in life by talking or making friends on line. Many people have myspace, facebooks, aim, these all benefit us by having conversations with one another. Many people believe computers are bad but how can you make friends if you can never talk to them? I am very fortunate for having a computer that can help with not only school work but my social life and how I make friends. Computers help us with finding our locations, coordibates and millions of information online. If we didn't go on the internet a lot we wouldn't know how to go onto websites that @MONTH1 help us with locations and coordinates like @LOCATION1. Would you rather use a computer or be in @LOCATION3. When your supposed to be vacationing in @LOCATION2. Million of information is found on the internet. You can as almost every question and a computer will have it. Would you rather easily draw up a house plan on the computers or take @NUM1 hours doing one by hand with ugly erazer marks all over it, you are garrenteed that to find a job with a drawing like that. Also when appling for a job many workers must write very long papers like a @NUM3 word essay on why this job fits you the most, and many people I know don't like writing @NUM3 words non-stopp for hours when it could take them I hav an a computer. That is why computers we needed a lot now adays. I hope this essay has impacted your descion on computers because they are great machines to work with. The other day I showed my mom how to use a computer and she said it was the greatest invention sense sliced bread! Now go out and buy a computer to help you chat online with friends, find locations and millions of information on one click of the button and help your self with getting a job with neat, prepared, printed work that your boss will love.

- **Ground truth trait scores:** *Content: 4/6, Organization: 4/6, Word Choice: 4/6, Sentence Fluency: 3/6, Conventions: 4/6*)

- **Rationale (GPT-3.5-Turbo): [sentence fluency]:** The essay lacks sentence fluency due to inconsistent structure, awkward phrasing, and errors in grammar and punctuation. The ideas are presented in a disjointed manner, making it challenging for the reader to follow smoothly. The essay would benefit from improved sentence variety and coherence.**[word choice]:** The essay lacks precision and variety in word choice. There are numerous spelling and grammar errors, misuse of words, and awkward phrasing. The vocabulary is limited, and the language is not engaging or effective in persuading the audience. The essay would benefit from significant revisions for clarity and impact.**[conventions]:** The essay demonstrates significant issues with standard writing conventions. There are frequent errors in punctuation, spelling, grammar, and capitalization, which impede readability and clarity. The lack of control over basic conventions hinders effective communication and requires substantial editing for improvement. **[organization]:** The essay lacks clear organization. Ideas are scattered and not well-developed, making it difficult to follow a logical flow. The structure of a letter to the editor is attempted but not effectively maintained. Improvement is needed in organizing paragraphs and connecting ideas cohesively. **[content]:** The essay discusses the benefits of using computers, such as online communication, access to information, and assistance with tasks like job applications. However, the ideas lack depth and coherence, with limited development and organization, impacting the overall clarity and effectiveness of the argument.

- **Rationale (Llama-3.1-8B-Instruct): [sentence fluency]**: The essay has a choppy flow with abrupt sentence transitions and inconsistent sentence structure. Many sentences are short and lack variety, resulting in a monotonous rhythm. The use of fragments and run-on sentences further disrupts the flow, making it difficult to follow at times. **[word choice]**: The essay demonstrates a limited vocabulary, relying heavily on simple and generic words. The writer uses colloquial expressions and slang ("garrenteed," "I hav an") and misuses words ("coordibates," "able ourselfs"). The text also features overused phrases and clichés ("the greatest invention since sliced bread"). **[conventions]**: The essay demonstrates limited control of standard writing conventions, with numerous errors in punctuation, spelling, capitalization, and grammar. Internal punctuation is frequently incorrect, and spelling errors, such as "coordibates" and "mysace," significantly impede readability. **[organization]**: The essay lacks a clear and logical structure. Ideas are presented in a disjointed manner, jumping between topics such as the benefits of computers, social media, and job applications. Transitions between ideas are absent, making the essay difficult to follow. **[content]**: The essay lacks depth and balance, relying heavily on general statements and clichés. It fails to provide nuanced arguments, instead using simplistic and repetitive points to support its claim that computers benefit society. Relevant details are often overly general or limited.

H.2 A Rationale of an Essay for Prompt 3 of Feedback Prize

- **Essay:** I think that students would benefit from learning at home,because they wont have to change and get up early in the morning to shower and do there hair. taking only classes helps them because at there house they'll be pay more attention. they will be comfortable at home. The hardest part of school is getting ready. you wake up go brush your teeth and go to your closet and look at your cloths. after you think you picked a outfit u go look in the mirror and youll either not like it or you look and see a stain. Then you'll have to change. with the online classes you can wear anything and stay home and you wont need to stress about what to wear.most students usually take showers before school. they either take it before they sleep or when they wake up. some students do both to smell good. that causes them do miss the bus and effects on there lesson time cause they come late to school. when u have online classes u wont need to miss lessons cause you can get everything set up and go take a shower and when u get out your ready to go.when your home your comfortable and you pay attention. it gives then an advantage to be smarter and even pass there classmates on class work. public schools are difficult even if you try. some teacher dont know how to teach it in then way that students understand it. that causes students to fail and they may repeat the class.

- **Ground Truth Trait Scores:** *Cohesion: 3.5/5.0, Syntax: 3.5/5.0, Vocabulary: 3.0/5.0, Phraseology: 3.0/5.0, Grammar: 4.0/5.0, Conventions: 3.0/5.0* cohesion syntax vocabulary phraseology grammar conventions

- **Rationale (GPT-3.5-Turbo): [conventions]:** "they wont have to change and get up early in the morning to shower and do there hair." - Spelling and grammar errors. "most students usually take showers before school." - Spelling errors and lack of punctuation. **[grammar]:** "they wont have to change and get up early in the morning to shower and do there hair." (Grammar errors: "wont" should be "wont́," "there" should be "their"). **[phraseology]:** "get up early in the morning to shower and do there hair." - Incorrect grammar and spelling. "you wake up go brush your teeth and go to your closet and look at your cloths." - Incorrect grammar and spelling. "most students usually take showers before school." - Relevant to the topic but lacks variety in phrases.**[vocabulary]:** "they wont have to change and get up early in the morning to shower and do there hair." - Incorrect word usage and spelling errors. "most students usually take showers before school." - Simple vocabulary, lacks complexity. **[syntax]:** "The hardest part of school is getting ready." - Well-written, simple sentence. "most students usually take showers before school." - Minor errors in sentence structure. "when your home your comfortable and you pay attention." - Basic errors in sentence structure. **[cohesion]:** "taking only classes helps them because at there house theyĺl be pay more attention." (Poorly written - lacks proper grammar and cohesion) "most students usually take showers before school." (Fairly written - lacks cohesion but clear in meaning) "when your home your comfortable and you pay attention." (Poorly written - lacks proper grammar and cohesion)

- **Rationale (Llama-3.1-8B-Instruct): [cohesion]:** The essay demonstrates a limited range of cohesive devices, primarily relying on repetition of ideas and phrases to connect sentences and paragraphs. There is a lack of transitional words and phrases, resulting in abrupt shifts between ideas, and some sentences seem disconnected from the rest of the text. **[syntax]:** The essay demonstrates a limited range of syntactic structures, primarily using simple sentences with some compound structures. There are occasional errors in sentence formation, such as incorrect subject-verb agreement ("they wont have to change") and missing articles ("at there house"). **[vocabulary]:** The essay demonstrates a narrow range of vocabulary, relying on basic words and phrases. Topic-related terms are occasionally used, but often inappropriately. There are errors in word formation and choice, such as "do there hair" and "pay more attention," which distort meanings. **[phraseology]:** The essay demonstrates a narrow range of phraseological devices, primarily relying on simple collocations and lexical bundles, such as "take classes," "get up early," and "pay attention." Repetitions and misuses of phrases are evident, with an over-reliance on basic sentence structures. **[grammar]:** The essay demonstrates inconsistent grammar usage, with several errors in verb tense, subject-verb agreement, and punctuation. Incorrect word order, missing articles, and misuse of pronouns are also prevalent. However, the writer attempts to convey their ideas and uses basic sentence structures effectively. **[conventions]:** The essay demonstrates a developing use of conventions to convey meaning. While there are some errors in spelling (e.g., "there" instead of "their"), capitalization (e.g., missing capitalization in sentences), and punctuation (e.g., missing periods), the overall structure and syntax are mostly coherent.