CEAES: Bidirectional Reinforcement Learning Optimization for Consistent and Explainable Essay Assessment

Xia Li[1,2] **and Wenjing Pan**[1]
[1] School of Information Science and Technology
[2] Center for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies, Guangzhou, China
{xiali, wjpan}@gdufs.edu.cn

Abstract

Most current automated essay quality assessment systems treat score prediction and feedback generation as separate tasks, overlooking the fact that scores provide a quantitative evaluation of quality, while feedback offers a qualitative assessment. Both aspects reflect essay quality from different perspectives, and they are inherently consistent and can reinforce each other. In this paper, we propose a novel bidirectional reinforcement learning framework that effectively utilizes this consistency constraint to jointly optimize score prediction and feedback generation, ensuring mutual reinforcement and alignment between them. In this way, our model is hope to obtain a simultaneous accurate ratings and consistent text feedback. We conducted extensive experiments on publicly available datasets. The results demonstrate that our approach surpasses the current state-of-the-art models, enhancing both scoring accuracy and feedback quality.

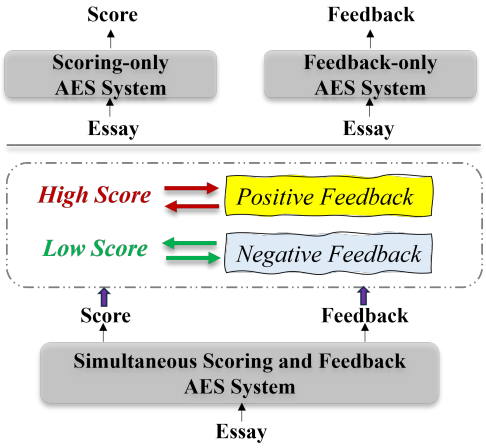


Figure 1: The diagram illustrates previous methods using separate models for score prediction and feedback generation. In contrast, our model jointly predicts scores and generates feedback, leveraging their consistency to enable mutual optimization.

1 Introduction

Automated Essay Quality Assessment aims to evaluate the quality of student essays. These systems enhance educators' efficiency and provide consistent scoring standards, reducing the subjectivity introduced by human bias. The assessment process encompasses both scoring (predicting scores) and evaluation (generating feedback[1]).

Most existing work focuses on the task of Automated Essay Scoring (AES), which includes prompt-specific AES and cross-prompt AES. This paper concentrates on prompt-specific AES. Current research in this area primarily aims to model textual content and discourse structures to optimize essay feature representation, thus improving scoring accuracy and reliability. Early studies (Larkey, 1998; Rudner and Liang, 2002; Yannakoudakis et al., 2011; Chen and He, 2013; Attali and Burstein, 2004; Phandi et al., 2015) focused on the extraction of handcrafted features to represent the surface semantics of essays. With the advent of deep learning, it has transitioned to neural network approaches that automatically extract features from essays. This shift has yielded promising results by effectively modeling essay structure (Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017), coherence (Tay et al., 2018), content (Wang et al., 2022; Shibata and Uto, 2022; Uto et al., 2023; Boquio and Naval, 2024), and rankings (Xie et al., 2022). However, these studies primarily concentrate on the scoring prediction task, without offering any feedback mechanisms.

Another line of research has thoroughly investigated the generation of essay feedback, utilizing comment corpora (Nagata et al., 2020) and predefined feedback labels (Cai et al., 2023) to construct classification models, as well as leveraging writing skill evaluations to develop end-to-end generation models (Liu et al., 2024). However, these

[1]In this paper, feedback refers to textual feedback.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26267–26279

approaches remain limited to the singular task of providing feedback and do not fully account for essay scoring.

Intuitively, the score provides a quantitative evaluation of essay quality, while feedback offers a qualitative assessment. Both aspects reflect essay quality from different perspectives, so the score and feedback of an essay should exhibit a constrained and consistent relationship, indicating that they are not independent of one another. As illustrated in Figure 1, essays that achieve high scores are more frequently associated with favorable assessment feedback, whereas essays that receive unfavorable evaluation feedback are more likely to obtain low scores. Therefore, the consistent correlation between rating scores and evaluation feedback should be thoroughly integrated into the AES model.

In the current literature, there are few studies that address the task of simultaneous rating and feedback. Among these, Gong et al. (2021) introduced the IFlyEA system, which incorporates independent scoring and review modules, while Li et al. (2023) developed a compact model for concurrent scoring and reasoning generation using ChatGPT. However, these approaches do not comprehensively consider the interrelationship among the essay, rating score, and evaluation feedback. In fact, both grading and feedback serve as quality evaluations of the same essay and should therefore exhibit consistency. So the question is: how can we ensure that the essay's representation fully accounts for the correlation between the essay and the score, the correlation between the essay and the feedback, and that the generated feedback accurately interprets the score while maintaining coherence between the score and the feedback?

To address this issue, we propose a novel Bidirectional Reinforcement Learning optimization strategy aimed at jointly optimizing score prediction and feedback generation by aligning their outputs and facilitating mutual reinforcement. As illustrated in Figure 1, we integrate scoring and feedback generation into a unified model, where feedback is refined by the predicted score and the score is adjusted based on the generated feedback. This approach ensures that the feedback accurately reflects the reasoning behind the assigned score, while the scoring predictions are enhanced through feedback signals. Consequently, this leads to superior consistency, interpretability, and improved performance for both tasks.

Based on the proposed strategy, we introduce a novel **C**onsistent and **E**xplainable **A**utomated **E**ssay **S**coring (**CEAES**) model. Our model employs a multi-task architecture that jointly predicts scores and generates feedback through shared representations and bidirectional optimization. Specifically, we first encode essays using a shared encoder to extract features pertinent to both tasks. Then, we utilize a feedback generator to produce detailed feedback and a scoring component to predict essay scores. The bidirectional reinforcement learning framework alternates between two optimization directions: generating feedback conditioned on the predicted score and refining scores based on the generated feedback. This alternating strategy ensures that each task reinforces the other by aligning feedback with scoring criteria and adjusting scores according to the quality of the feedback. After several optimization cycles, the model effectively generates actionable, detailed feedback while concurrently enhancing the accuracy and reliability of score predictions. The main contributions of this paper can be summarized as follows:

1) To the best of our knowledge, this is the first attempt to jointly optimize score prediction and feedback generation, enabling mutual reinforcement between the two tasks.

2) We propose a novel bidirectional reinforcement learning strategy to dynamically align feedback generation with score predictions, ensuring that the generated feedback is consistent with scoring criteria.

3) We conduct extensive experiments on the ASAP++ dataset, and the results demonstrate that our approach outperforms state-of-the-art models.

2 Our Approach

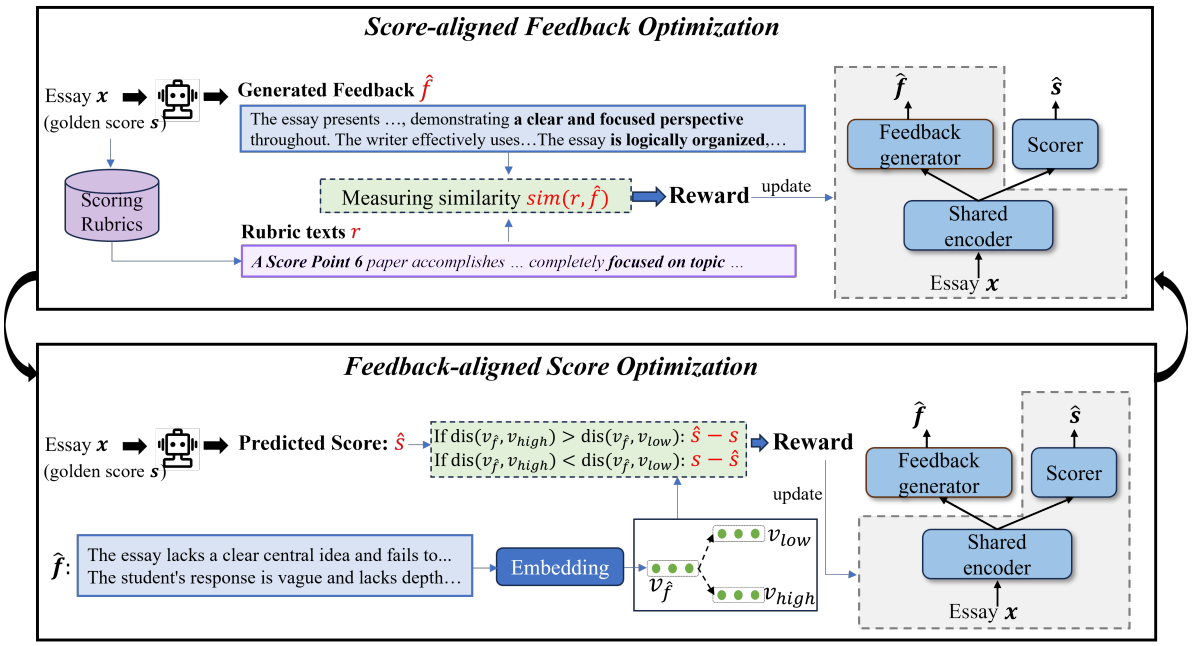
Our method aims to enable the model to predict essay scores effectively and provide consistent textual feedback simultaneously. The model architecture is shown in Figure 2.

2.1 Task Definition

Given a dataset of essays $\mathcal{D} = \{(e_i, s_i)\}_{i=1}^{N}$, where e_i represents the i-th essay and s_i is the corresponding score. Our objective is to train a model $\mathcal{M}$ that can predict both the score $\hat{s}$ and generates feedback $\hat{f}$ for an unseen essay e.

2.2 Scoring and Feedback Model

Our model comprises a shared encoder, a scoring component for the scoring task, and a feedback generator for the feedback generation task. This design



v_{low}, v_{high} is the centroids of low- and high-score essays' reference feedbacks in training data.

Figure 2: Overall architecture of CEAES. CEAES simultaneously predicts scores and generates feedback. Both tasks are trained jointly with alternating reinforcement learning optimization. When optimizing feedback, we measure the similarity between $\hat{f}$ and the rubric texts r corresponding to the golden score. We use the similarity as the reward to guide the model in generating score-aligned feedback. When optimizing scoring, we align $\hat{f}$ by comparing its distance to the centroids v_{high}, v_{low} of high- and low-score essays' feedback. We then use the difference of $\hat{s}$ and s as the reward to guide the scoring prediction.

allows the model to learn shared representations while addressing task-specific objectives.

Shared Encoder To ensure consistency and shared representations across both tasks, we employ a shared encoder architecture based on the pre-trained BART model. Specifically, the input essay e is tokenized into a sequence of tokens $T = \{t_1, t_2, \ldots, t_n\}$, where n is the length of the essay. The tokenized sequence is passed through the BART encoder to generate a sequence of contextualized embeddings:

$$H = \text{BART}_{\text{Encoder}}(T), \tag{1}$$

where $H = \{h_1, h_2, \ldots, h_n\}$ represents the contextual embeddings for each token.

Scoring Task The scoring task focuses on predicting scores for essays. Specifically, a multi-head self-attention mechanism is applied to the token embeddings H:

$$\text{head}_l = \text{softmax}\left(\frac{HW_{Q_l}(HW_{K_l})^\top}{\sqrt{d_k}}\right)HW_{V_l}, \tag{2}$$

$$A = \text{Concat}(\text{head}_1, \ldots, \text{head}_l)W_O, \tag{3}$$

where $W_{Q_l}, W_{K_l}, W_{V_l}$ are learnable weight matrices for queries, keys, and values in the l-th head, W_O is the output projection matrix, and d_k is the hidden size. The attention outputs are mean-pooled to obtain the essay representation h_s. The essay representation is then passed through a dense layer with sigmoid activation to predict the score:

$$\hat{s} = \sigma(W_r h_s + b_r), \tag{4}$$

where W_r and b_r are the dense layer's weights and bias, and σ is the sigmoid function. The optimization objective for this task is to minimize the mean squared error (MSE) loss:

$$\mathcal{L}_{\text{score}} = \frac{1}{N}\sum_{i=1}^{N}(s_i - \hat{s}_i)^2 \tag{5}$$

Feedback Generation Task The feedback generation task aims to generate targeted feedback based on the content of the essay.

Large language models (LLMs) are utilized to produce coherent and high-quality reference feedback for each essay. To ensure the generated feedback aligns with the scoring criteria (rubric guidelines), we design an instruction based on chain-of-thought reasoning and role-playing techniques.

The prompt directs the LLM to generate feedback grounded in the essay, its true score, and the rubric's evaluation criteria. Detailed descriptions of the prompt design are provided in the Appendix C.

The BART decoder acts as the feedback generator, taking the encoder output H and previously generated tokens $\hat{f}_{<t}$ to autoregressively predict the next token $\hat{f}_t$ at each step:

$$P(\hat{f}_t|\hat{f}_{<t}, H) = \text{softmax}(W_v \cdot D_t + b_v) \quad (6)$$

where D_t represents the decoder's intermediate output at time step t, W_v and b_v are learnable parameters of the decoder. The process iterates until an end-of-sequence token is generated or a maximum sequence length is reached. During training, cross-entropy loss is applied to optimize the token probability distribution generated by the feedback generator against the target distribution:

$$\mathcal{L}_{\text{feedback}} = -\sum_{t=1}^{m} f_t \log P(\hat{f}_t|\hat{f}_{<t}, H) \quad (7)$$

where f_t is the one-hot vector of the target token in reference feedback, $P(\hat{f}_t|\hat{f}_{<t}, H)$ is the predicted token probability, and m is the length of the feedback sequence.

The overall training loss is:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{score}} + \beta\mathcal{L}_{\text{feedback}}, \quad (8)$$

where $\mathcal{L}_{\text{score}}$ and $\mathcal{L}_{\text{feedback}}$ are defined in Equations 5 and 7, respectively. The weights α and β control the relative importance of the scoring and feedback tasks, allowing the model to balance learning effectively.

2.3 Bidirectional Reinforcement Learning Optimization

In practice, teacher feedback should explicitly justify the assigned score, and the score should be inferable from the provided feedback. To achieve this, we propose a Bidirectional Reinforcement Learning framework that optimizes feedback generation conditioned on the predicted score to align it with the scoring rubric while refining score predictions based on the generated feedback. This joint optimization of feedback generation and scoring enhances their consistency and reliability.

Score-aligned Feedback Optimization To ensure that the generated feedback and the corresponding essay score are consistent in evaluating essay quality, we propose a score-aligned reward function that acts as a weak supervision signal to update and optimize the parameters of the feedback generation model. Since textual feedback and scores cannot be directly compared, we utilize the text scoring criteria associated with the scores to assess their consistency with the generated feedback. This consistency is then employed as a reward to improve the feedback generation model through reinforcement learning. As illustrated in Figure 2, we measure semantic similarity between the two as an indicator of consistency and reward. The specific process is detailed as follows.

For a predicted score $\hat{s}$, its corresponding rubric text r is retrieved[2]. A pre-trained BERT model encodes the generated feedback $\hat{f}$ and rubric r into vector representations, $v_{\hat{f}}$ and v_r. The semantic consistency between the feedback and the rubric is measured using cosine similarity $\cos(v_{\hat{f}}, v_r)$. This enables the use of continuous signals grounded in human-defined scoring rubrics, thereby enhancing training stability. The reward function is defined as follows:

$$R_s = \begin{cases} \cos(v_{\hat{f}}, v_r), & \text{if } \cos(v_{\hat{f}}, v_r) \geq \psi, \\ -\lambda(1 - \cos(v_{\hat{f}}, v_r)), & \text{otherwise,} \end{cases} \quad (9)$$

where ψ is a similarity threshold and λ is a penalty coefficient. Our reinforcement learning approach employs the policy gradient algorithm. We treat feedback generation as the current policy $\pi_\theta(\hat{f}|e)$, which defines the probability of generating feedback $\hat{f}$ given the essay e. The policy is updated using the following gradient:

$$\nabla_\theta J(\theta) = \mathbf{E}_{\pi_\theta}[R_s \nabla_\theta \log \pi_\theta(\hat{f}|e)], \quad (10)$$

where R_s acts as the reward signal to guide the optimization. The parameters θ are updated as:

$$\theta \leftarrow \theta + \gamma \mathbf{E}_{\pi_\theta}[R_s \nabla_\theta \log \pi_\theta(\hat{f}|e)], \quad (11)$$

where γ denotes the learning rate and θ denotes the parameters in the shared encoder and feedback generator. Our method ensures that the generated feedback aligns with the scoring rubric by iteratively computing rewards and updating the policy.

[2]We preprocess the scoring rubrics from the ASAP dataset, extracting rubric descriptions corresponding to each score level and storing them in a structured dictionary. During retrieval, the predicted score is used as a key to obtain the corresponding human-annotated rubric text.

Feedback-aligned Score Optimization Preliminary experimental observations revealed that the model exhibits a bias towards predicting intermediate scores for both low- and high-score essays. This bias may be attributed to inherent class imbalance in training data distributions, characterized by underrepresentation of high/low-scoring essays (minority classes) and overrepresentation of medium-scoring samples (majority class). To mitigate this prediction bias that causes overestimation for low-quality essays and underestimation scores for high-quality ones, we propose a feedback-aligned reward function as a weak supervision signal to optimize the scoring model. As illustrated in Figure 2, we leverage the alignment between generated feedback and reference feedback for high- and low-score essays to design a reward function. This function encourages the scorer to predict scores that are more consistent with the evaluative implications of the feedback, thereby reducing bias in score prediction.

We first embed all reference feedback texts from the training set and partition them into two groups: feedback corresponding to high-scoring essays F_h and feedback corresponding to low-scoring essays F_l. For each group, we compute the euclidean centroids as follows:

$$v_{high} = \frac{1}{|F_h|} \sum_{f \in F_h} v_f, \quad v_{low} = \frac{1}{|F_l|} \sum_{f \in F_l} v_f \tag{12}$$

where v_{high}, v_{low} is the centroid of the reference feedback vectors for high- and low-quality essays; v_f denotes the embedding of reference feedback f, v_{high} and v_{low} provide a reference point for determining the alignment of generated feedback with positive or negative feedback.

If the generated feedback $v_{\hat{f}}$ aligns with v_{high}, that is, $\text{dis}(v_{\hat{f}}, v_{high}) < \text{dis}(v_{\hat{f}}, v_{low}) + \delta$, and the corresponding essay is indeed of high quality, the scorer is encouraged to predict a higher score. Conversely, if the generated feedback $v_{\hat{f}}$ aligns with v_{low}, that is, $\text{dis}(v_{\hat{f}}, v_{high}) > \text{dis}(v_{\hat{f}}, v_{low}) + \delta$, and the essay quality is low, the scorer is encouraged to predict a lower score. The reward function is as follows:

$$R_f = \begin{cases} 1 - (s_{\text{high}} - \hat{s}), & \text{if } v_{\hat{f}} \text{ aligns with } v_{high}, \\ 1 - (\hat{s} - s_{\text{low}}), & \text{if } v_{\hat{f}} \text{ aligns with } v_{low}. \end{cases} \tag{13}$$

where s_{high} and s_{low} represent the true scores for high- and low-quality essays, respectively, $\hat{s}$ is the predicted score, and δ is a margin hyperparameter controlling the alignment threshold. We employ the policy gradient method to optimize the scorer. The optimization objective is:

$$\nabla_\theta J(\theta) = \mathbf{E}_{\pi_\theta}[R_f \nabla_\theta \log \pi_\theta(\hat{y}|\hat{f})], \tag{14}$$

where $\pi_\theta(\hat{y}|\hat{f})$ is the scoring policy parameterized by θ, $\nabla_\theta J(\theta)$ represents the gradient used to update the scorer's parameters.

Alternating Optimization Strategy As simultaneous optimization in both directions can result in conflicting gradients, hindering convergence, we adopt an alternating optimization strategy. Specifically, the scoring and feedback generation tasks are trained normally in each epoch, while reinforcement learning alternates between two optimization directions: in one epoch, feedback generation is optimized based on score-aligned rewards; in the next epoch, scoring is optimized based on feedback-aligned rewards. This approach facilitates effective collaboration between the two tasks while avoiding mutual interference, and significantly alleviates the inherent instability of reinforcement learning in natural language generation. By iteratively refining both tasks, the model learns to generate feedback that is coherent, detailed, and grounded in accurate scoring, while also producing score predictions that align with evaluative feedback content.

3 Experimental Setup

3.1 Dataset and Evaluation Metrics

We use the ASAP dataset (Mathias and Bhattacharyya, 2018), which contains 12,978 student essays across eight prompts. We conduct 5-fold cross-validation per prompt. The evaluation metric used is Quadratic Weighted Kappa (QWK). For feedback, we propose BERTScore (Hanna and Bojar, 2021) to measure semantic similarity with the scoring rubric. See Appendix A for more details.

3.2 Baseline Models

We compared our model with prompt-specific AES models. The baseline models include EASE[3], ALL-MTL-cTAP (Cummins et al., 2016), CNN+LSTM (Taghipour and Ng, 2016), LSTM-CNN-att (Dong et al., 2017), SKIPFLOW (Tay et al., 2018), HISK+BOSWE (Cozma et al., 2018), R^{2}BERT (Yang et al., 2020), NPCR(Xie et al., 2022) and BERT-ahs-wm-wcc (Boquio and Naval, 2024).

[3] https://github.com/openedx-unsupported/ease

Model	P1	P2	P3	P4	P5	P6	P7	P8	AVG
Score-only Models									
GPT-3.5-turbo* *(0-shot)*	0.264	0.492	0.351	0.437	0.516	0.489	0.153	0.307	0.376
EASE (SVR)	0.781	0.630	0.621	0.749	0.782	0.771	0.727	0.534	0.699
EASE (BLRR)	0.761	0.621	0.606	0.742	0.784	0.775	0.730	0.617	0.705
ALL-MTL-cTAP (2016)	0.816	0.667	0.654	0.783	0.801	0.778	0.787	0.692	0.747
CNN+LSTM (2016)	0.821	0.688	0.694	0.805	0.807	0.819	0.808	0.644	0.761
LSTM-CNN-att (2017)	0.822	0.682	0.672	0.814	0.803	0.811	0.801	0.705	0.764
SKIPFLOW (2018)	0.832	0.684	0.695	0.788	0.815	0.810	0.800	0.697	0.764
Self-att-LSTM (2018)	0.834	0.692	0.700	0.811	0.819	0.822	0.816	0.713	0.776
HISK+BOSWE (2018)	0.845	0.729	0.684	0.829	0.833	0.830	0.804	0.729	0.785
R^2BERT (2020)	0.817	0.719	0.698	0.845	0.841	0.847	0.839	0.744	0.794
NPCR (2022)	0.856	0.750	0.756	0.851	0.847	0.858	0.838	0.779	0.817
BERT-ahs-wm-wcc (2024)	0.834	0.688	0.708	0.821	0.832	0.828	0.817	0.771	0.788
Score&Feedback Models									
GPT3.5 *w/o rubric*	0.155	0.501	0.351	0.437	0.506	0.524	0.074	0.307	0.357
GPT3.5 *w/ rubric*	0.264	0.492	0.477	0.477	0.516	0.527	0.153	0.341	0.406
GPT4 *w/ rubric*	0.272	0.481	0.478	0.503	0.557	0.529	0.123	0.384	0.415
CEAES (ours)	**0.863**	**0.773**	**0.759**	**0.865**	**0.842**	**0.856**	**0.843**	**0.765**	**0.821**

Table 1: Scoring performance of models (measured by QWK). Underlining indicates top performance in score-only models; bolding denotes top performance in models with scoring and feedback.

Due to differing datasets and unavailable code from existing AES feedback generation methods (Liu et al., 2024; Gong et al., 2021), we cannot perform direct comparisons. Therefore, we develop three baseline models based on GPT-3.5 and GPT-4 for scoring and feedback generation: GPT3.5 w/o rubric, GPT3.5 w/ rubric, and GPT4 w/ rubric. We conduct comparisons in a zero-shot setting to facilitate comparisons.

3.3 Implementation Details

CEAES employs `BART-base`[4] for shared representation learning and feedback generation. Task weights are set to balance losses, with reinforcement learning weights tuned for optimization. `BERT-base-uncased`[5] encodes feedback and rubric text for reward computation. All experiments were conducted on a single NVIDIA RTX8000 GPU. Training took approximately 2 hours for 60 epochs. The model has 141.8 million parameters. We select the best model based on the validation results and report the test results. More details are provided in Appendix B.

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

[5] https://huggingface.co/google-bert/bert-base-uncased

4 Results and Analysis

4.1 Results on Scoring

We first present the scoring performance of our model, comparing it with existing models that focus exclusively on essay scoring.As shown in Table 1, our model achieves an average QWK of 0.821, surpassing all baseline score-only models. Compared to the previous state-of-the-art model, NPCR, our approach shows a 2.3% improvement on Prompt 2. These results demonstrate the efficacy of our jointly optimization approach in automated essay scoring. Additionally, our method can also provide constructive feedback, making it superior to score-only models in both scoring performance and practical functionality.

4.2 Results on Feedback Generation

To assess whether our model provides meaningful feedback that correlates with essay quality, we then compare it with existing models capable of providing feedback. Due to the unavailability of codes and the different datasets used by previous feedback-only models, direct comparisons are infeasible. Therefore, we select GPT-3.5 and GPT-4 as baseline models for a simultaneous comparison in both scoring and feedback generation. The results are summarized as follows:

Model	Metric	P1	P2	P3	P4	P5	P6	P7	P8	AVG
	QWK	0.155	0.501	0.351	0.437	0.506	0.524	0.074	0.307	0.357
GPT3.5	BertScore(P)	0.462	0.597	0.507	0.511	0.531	0.529	0.423	0.584	0.518
w/o rubric	BertScore(R)	0.467	0.583	0.601	0.606	0.666	0.691	0.451	0.598	0.583
	BertScore(F1)	0.464	0.590	0.550	0.554	0.591	0.599	0.437	0.591	0.547
	QWK	0.264	0.492	0.477	0.477	0.516	0.527	0.153	0.341	0.406
GPT3.5	BertScore(P)	0.455	0.513	0.524	0.528	0.544	0.528	0.445	0.595	0.517
w/ rubric	BertScore(R)	0.524	0.599	0.612	0.631	0.686	0.716	0.486	0.632	0.611
	BertScore(F1)	0.487	0.553	0.565	0.575	0.607	0.608	0.465	0.613	0.559
	QWK	0.264	0.481	0.478	0.503	0.557	0.529	0.123	0.384	0.415
GPT4	BertScore(P)	0.451	0.592	0.589	0.500	0.517	0.528	0.600	0.579	0.545
w/ rubric	BertScore(R)	0.538	0.611	0.672	0.626	0.683	0.716	0.552	0.631	0.629
	BertScore(F1)	0.491	0.601	0.628	0.556	0.589	0.608	0.575	0.604	0.581
	QWK	0.863	0.773	0.759	0.865	0.842	0.856	0.843	0.765	0.821
CEAES	BertScore(P)	0.575	0.615	0.622	0.630	0.641	0.647	0.648	0.601	0.622
(ours)	BertScore(R)	0.561	0.611	0.621	0.632	0.697	0.753	0.618	0.638	0.641
	BertScore(F1)	0.568	0.613	0.621	0.631	0.668	0.696	0.633	0.619	0.631

Table 2: Feedback and scoring performance of models.

Quantitative analysis of feedback alignment with scoring rubrics. We use BERTScore to measure the similarity between the generated feedback and the rubric text corresponding to the score range. As demonstrated in Table 2, our CEAES model consistently achieves higher BERTScore values across various prompts, ranging from 0.622 to 0.641. This indicates that the generated feedback is more accurate and better aligned with the scoring rubrics. In contrast, the GPT-3.5 and GPT-4 models exhibit lower BERTScore (F1) values, ranging from 0.547 to 0.581, which reflects a decreased quality of feedback in relation to the scoring criteria. This is further evidenced by their relatively low QWK scores—for instance, GPT-4 achieves only a QWK of 0.415—suggesting that large language models (LLMs) may have a skewed understanding of essay quality. In comparison, our model achieves a much higher QWK of 0.821.

Qualitative comparison of feedback generated by our model and LLMs. In addition to the quantitative evaluation of feedback quality using BERTScore, we further conduct a qualitative analysis to better understand the advantages of CEAES over GPT-3.5. Table 3 presents a qualitative comparison of feedback generated by the two models for a high-quality essay (score: 12), alongside the corresponding human scoring rubric. While

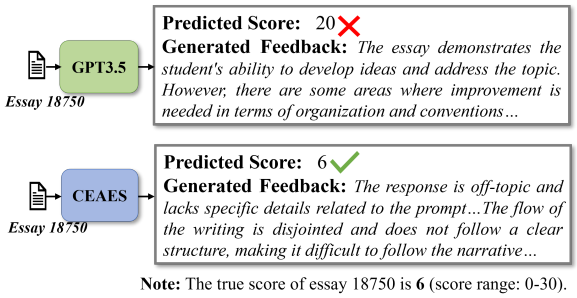


Figure 3: Case study on our model and GPT-3.5 in predicting scores and generating feedback.

GPT-3.5's feedback is generally reasonable, it overlooks several key evaluative criteria emphasized in the rubric. In contrast, CEAES-generated feedback explicitly integrates key rubric terms such as "*thoughtful position*" and "*fluent transitions*", offering a more rubric-aligned and interpretable explanation. This qualitative result underscores the superiority of our model in generating feedback that reflects the evaluation criteria more accurately and clearly.

Case study on our model and GPT-3.5 in predicting scores and generating feedback. We compare the performance of GPT-3.5 and our model CEAES on the same essay, as shown in Figure 3. GPT-3.5 significantly overestimates the score,

Feedback Source	Feedback Content
GPT	The essay presents a strong argument against the use of computers, highlighting the benefits of books over computers. The overall tone suggests a personal investment in the subject. However, there are areas for improvement in organization, word choice, sentence fluency, and conventions.
Our Model	The student has presented a well-developed response that takes a clear and **thoughtful position** on the effects of computers on society. The use of examples and reasoning throughout the essay enhances the overall effectiveness of the argument. **Transitions are fluent and sophisticated**, contributing to the overall flow of the essay.
Human Rubrics corresponding to 12 points	A well-developed response that takes a clear and **thoughtful position** and provides persuasive support. Typical elements: **a.** Has fully elaborated reasons with specific details. **b.** Exhibits strong organization. **c.** Is fluent and uses **sophisticated transitional** language. **d.** May show a heightened awareness of audience.

Table 3: Qualitative comparison of feedback generated by GPT and our model. ***Note:*** Essay ID: 346. Score: 12. Score Range: 2–12.

assigning a 20 compared to the actual score of 6, highlighting its accuracy issues. In contrast, our model predicts the score accurately, demonstrating greater reliability. Additionally, CEAES provides precise feedback addressing specific issues like prompt deviations and structural weaknesses, while GPT-3.5, despite offering coherent feedback, lacks specificity and accuracy, resulting in misaligned feedback with its scores. These findings confirm that our model effectively delivers meaningful feedback that correlates well with essay quality.

4.3 Discussion

Ablation Studies. To validate the proposed bidirectional reinforcement learning framework, we conduct an ablation study by removing the reinforcement learning components for the scoring task (RL(S)) and feedback generation task (RL(F)). As presented in Table 4, the following findings can be derived: 1) Removing RL(S) results in a decrease in QWK from 0.821 to 0.805, indicating a significant dependency on RL(S). 2) Eliminating RL(F) leads to a decline in feedback generation performance, with Rouge-1 dropping from 0.654 to 0.646, underscoring reliance on RL(F). 3) When both RL(F) and RL(S) are removed, performance further deteriorates, with QWK decreasing by 1.6 and Rouge-1 by 1.0, suggesting a cooperative effect between the two optimizations.

Case study analysis of the impact of feedback on scoring performance. To investigate how integrating feedback generation into the automated scoring model improves prediction accuracy, we compare a score-only model using only the CEAES system with our full CEAES model, which incorporates both scoring and feedback generation. As detailed in Figure 4, we analyze specific cases to assess feedback's impact on scoring accuracy. Our findings show that the score-only model inaccurately assigned scores, overestimating in Case #1 and underestimating in Case #2, while the full CEAES model accurately evaluated scores and provided constructive feedback. This integration enhances content analysis and aligns better with rubric criteria, facilitating score refinement. These results confirm that feedback generation significantly improves scoring prediction performance.

5 Related Work

5.1 Automated Essay Scoring

Early studies (Larkey, 1998; Rudner and Liang, 2002; Yannakoudakis et al., 2011; Chen and He, 2013; Attali and Burstein, 2004; Phandi et al., 2015) use machine learning approaches, such as classification, ranking, and regression, with handcrafted features. With neural networks, more advanced architectures emerge (Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017; Tay et al., 2018; Wang et al., 2022; Shibata and Uto, 2022; Uto et al., 2023; Li and Pan, 2025). For example, Dong and Zhang (2016) propose the hierarchical models to capture word and essay-level textual structure; Tay et al. (2018) propose to improve the contextual understanding through sentence depen-

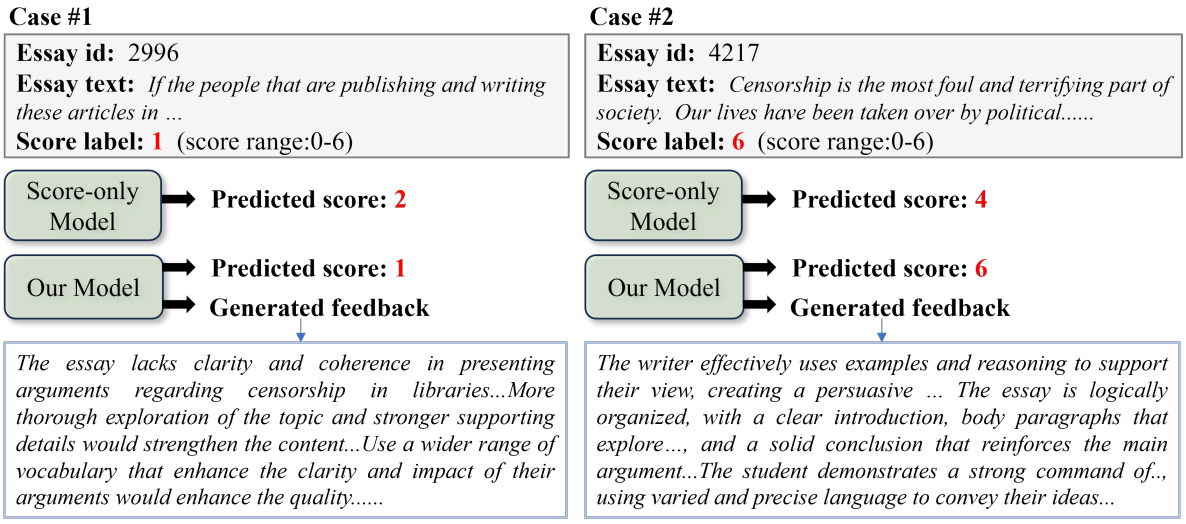


Figure 4: Case study analysis of the impact of feedback on scoring performance.

dency modeling. Recent transformer-based models (Wang et al., 2022; Uto et al., 2023) further enhance the scoring performance by leveraging deep contextual representations.These developments illustrate a transition from feature-based models to neural approaches, with most studies concentrating on scoring prediction without providing other textual feedback.

5.2 Automated Feedback Generation

Recent advancements explore integrating feedback generation to enhance assessment. Nagata et al. (2020) highlight the importance of generating effective feedback, including grammatical corrections, for language learning. Gong et al. (2021) introduce IFlyEA, an AES system with multilayered analysis (grammar, rhetoric, discourse) and independent scoring and feedback modules. Cai et al. (2023) propose a soft label-driven approach using 30 predefined feedback labels. Liu et al. (2024) develope GEEF, an encoder-decoder model integrating writing skill assessment for targeted feedback. Li et al. (2023) fine-tune a smaller model using ChatGPT-generated scores and rationales for simultaneous grading and rationale generation.

Although previous methods achieve promising results, they have not considered the correlation between scoring and feedback. In contrast, our model leverages bidirectional reinforcement learning to jointly optimize both tasks, enabling a more synergistic learning process. By simultaneously refining scoring accuracy and enhancing feedback relevance, our approach ensures a more consistent and informative assessment framework.

Model	AVG QWK	AVG Rouge-1
CEAES	0.821	0.654
CEAES w/o RL(S)	0.805	0.652
CEAES w/o RL(F)	0.816	0.646
CEAES w/o RL(F+S)	0.806	0.644

Table 4: Comparison of models with and without reinforcement learning components

6 Conclusion

In this paper, we propose CEAES, a novel model for joint score prediction and feedback generation using bidirectional reinforcement learning. Specifically, we design a unified framework that aligns feedback with scoring criteria and refines scores based on feedback, ensuring consistency and mutual reinforcement. Experimental results on the public dataset demonstrate that CEAES outperforms state-of-the-art models in scoring accuracy. We further validate its ability to generate feedback consistent with scores while maintaining reliable scoring, proving its practicality for real-world educational applications.

7 Limitations

Our approach achieves significant improvements in simultaneous scoring and feedback generation but still has some limitations: 1) Our method relies heavily on the stability of reinforcement learning. We use RL to optimize scoring-feedback consistency, but RL training is typically unstable, highly sensitive to hyperparameters, and prone to issues

such as mode collapse and reward bias, particularly in text generation tasks. 2) Since RL relies on reward signals, and the scoring-feedback consistency reward is inherently based on similarity metrics, the model may learn to generate overly templated feedback. Future work could explore more stable RL training strategies, such as reward shaping or adversarial training, to improve robustness. Additionally, incorporating diverse feedback objectives beyond similarity-based rewards could help mitigate the risk of template-driven feedback generation.

Acknowledgements

This work is supported by the National Natural Science Foundation of China [grant number: 61976062] and Social Sciences Foundation of the Ministry of Education of China [grant nunber: 24YJA740014].

References

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v.2.0. *ETS Research Report Series*, 2004(2):i–21.

Eujene Nikka V. Boquio and Prospero C. Naval, Jr. 2024. Beyond canonical fine-tuning: Leveraging hybrid multi-layer pooled representations of BERT for automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2285–2295, Torino, Italia. ELRA and ICCL.

Yuzhe Cai, Shaoguang Mao, Chenshuo Wang, Tao Ge, Wenshan Wu, Yan Xia, Chanjin Zheng, and Qiang Guan. 2023. Enhancing detailed feedback to chinese writing learners using a soft-label driven approach and tag-aware ranking model. In *Natural Language Processing and Chinese Computing*, pages 576–587, Cham. Springer Nature Switzerland.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA. Association for Computational Linguistics.

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.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–799, Berlin, Germany. 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.

Jiefu Gong, Xiao Hu, Wei Song, Ruiji Fu, Zhichao Sheng, Bo Zhu, Shijin Wang, and Ting Liu. 2021. IFlyEA: A Chinese essay assessment system with automated rating, review generation, and recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 240–248, Online. Association for Computational Linguistics.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Leah S. Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 90–95, New York, NY, USA. Association for Computing Machinery.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6007–6026, Singapore. Association for Computational Linguistics.

Xia Li, Minping Chen, Jianyun Nie, Zhenxing Liu, Ziheng Feng, and Yingdan Cai. 2018. Coherence-based automated essay scoring using self-attention. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 386–397, Cham. Springer International Publishing.

Xia Li and Wenjing Pan. 2025. Kaes: Multi-aspect shared knowledge finding and aligning for cross-prompt automated scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24476–24484.

Yuanchao Liu, Jiawei Han, Alexander Sboev, and Ilya Makarov. 2024. Geef: A neural network model for automatic essay feedback generation by integrating

writing skills assessment. *Expert Systems with Applications*, 245:123043.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating corpora for research in feedback comment generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345, Marseille, France. European Language Resources Association.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

Lawrence M. Rudner and Tahung Liang. 2002. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

Takumi Shibata and Masaki Uto. 2022. Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2917–2926, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Masaki Uto, Itsuki Aomi, Emiko Tsutsumi, and Maomi Ueno. 2023. Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on Learning Technologies*, pages 1–18.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

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. 2020. 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, Online. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

A Dataset and Evaluation Metrics

We use the ASAP dataset (Mathias and Bhattacharyya, 2018) in our experiment, a widely recognized open-source dataset for the AES task. It includes 12,978 English essays written by students in response to eight prompts. Detailed statistics of the dataset are presented in Table 5. Following previous studies, experiments are conducted separately for each prompt using 5-fold cross-validation, allocating 60% of the data for training, 20% for validation, and 20% for testing. The evaluation metric for scoring is Quadratic Weighted Kappa (QWK), which assesses the agreement between the predicted scores and human-labeled scores. For the feedback, given the absence of public metrics, we propose utilizing BERTScore (Hanna and Bojar, 2021) to assess the semantic similarity between the generated feedback and the corresponding segment of the scoring rubric. Higher BERTScore values indicate better alignment with the rubrics, suggesting more accurate and relevant feedback.

B Implementation Details

CEAES uses the `BART-base` model[6] as the encoder for shared representation learning and the decoder for feedback generation. The scoring task weight is set to $\beta = 1.0$, and the feedback generation task weight is set to $\alpha = 0.2$, ensuring that the respective losses are balanced to the same magnitude. For reinforcement learning, the weight for optimizing feedback generation is set to $\gamma = 0.01$, and

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

Prompt	Genre	#Essay	Avg.len	Grade	Range
P1	Arg	1,783	350	8	2 - 12
P2	Arg	1,800	350	10	0 - 6
P3	SD	1,726	150	10	0 - 3
P4	SD	1,772	150	10	0 - 3
P5	SD	1,805	150	8	0 - 4
P6	SD	1,800	150	10	0 - 4
P7	Nar	1,569	300	7	0 - 30
P8	Nar	723	650	10	0 - 60

Table 5: Statistics of ASAP dataset. The Range column presents the score range. Arg denotes Argumentative, SD denotes Source-Dependent, and Nar denotes Narrative.

the weight for optimizing score prediction is set to $\eta = 0.04$. In reinforcement learning optimization, the similarity threshold ψ is set to 0.4, the penalty factor λ is set to 0.1, and the margin parameter δ is set to 0.1. The `BERT-base-uncased`[7] model is employed to encode feedback and rubric text for reward computation. Our experiments are conducted on an NVIDIA RTX8000 GPU. We train our model for 60 epochs with a batch size of 8 and a learning rate of 5×10^{-5}. The best model is selected based on the highest average QWK on the validation set. We run our model three times with different seeds and report the average results on the test set.

C Instruction Template for LLM Feedback Generation

We utilize LLMs to produce coherent and high-quality reference feedback for each essay. To ensure the generated feedback aligns with the scoring criteria (rubric guidelines), we design an instruction based on chain-of-thought reasoning and role-playing techniques. The instrution directs the LLM to generate feedback grounded in the essay, its true score, and the rubric's evaluation criteria. The instruction template is provided as follows:

You are an English teacher. Your task is to evaluate the student's English essay and generate feedback.

Please follow these steps carefully to evaluate the essay:

Step 1: Understand the Evaluation Criteria (Rubric)

Begin by thoroughly reading and understanding the rubric provided. Ensure that you are familiar with the specific attributes to evaluate. These attributes will be key in your assessment, and each one needs to be addressed in your feedback.

The attributes you need to evaluate include:{Insert the attributes corresponding to this prompt here}

The evaluation criteria are as follows:

—Start of Evaluation Criteria—

{Insert the overall rubric guideline and the rubric guidelines for multiple attributes here}

—End of Evaluation Criteria—

Step 2: Review the Essay Sentence by Sentence

Review the student's essay carefully, analyzing each sentence based on the rubric. Consider how well the essay responds to the prompt, develops ideas, is organized logically, and uses language effectively. As you review, pay attention to areas where the student performs well and areas that may need improvement, keeping in mind each of the rubric categories.

Note: In the essay, the named entities (people, places, dates, times, organizations, etc.) are replaced by placeholders (e.g., @NAME1, @LOCATION1, etc.), and capitalized phrases such as @CAP1, @CAP2, etc., are anonymized.

The essay prompt is as follows:

—Start of Essay Prompt—

{Insert essay prompt text here}

—End of Essay Prompt—

The student's essay that you need to evaluate is as follows:

—Start of Essay—

{Insert student's essay text here}

—End of Essay—

Step 3: Refer to the Provided Scores After reviewing the essay, refer to the overall score and the individual scores given for each attribute. These scores will help you contextualize your evaluation and feedback.

The scores for this essay are as follows:

{'score': '6', 'content': '3', 'organization': '3', 'word_choice': '3', 'sentence_fluency': '3', 'conventions': '3'}

Step 4: Provide Feedback Based on your review and the provided scores, offer feedback for the essay overall and for each attribute. Make sure your feedback is aligned with the rubric and reflects the strengths and weaknesses of the essay in each specific area.

[7] https://huggingface.co/google-bert/bert-base-uncased

Your final evaluation should be in the following JSON format:

{"overall_feedback": "A paragraph giving general feedback on the essay.", "content_feedback": "Feedback on how well the student addressed the topic and developed ideas.", "organization_feedback": "Feedback on the logical structure and flow of the essay.", "word_choice_feedback": "Feedback on the variety and appropriateness of vocabulary used.", "sentence_fluency_feedback": "Feedback on the readability and flow of sentences.", "conventions_feedback": "Feedback on grammar, punctuation, and spelling." }

D Feedback Generation Performance

Table 6 presents the performance of our model in generating feedback across different prompts, measured at the point where the feedback loss on the validation set is minimized. The table reports the BLEU scores and ROUGE values between the model-generated feedback and the reference feedback on the test set. It can be observed our model achieves the best performance on prompt 6, with a BLEU score of 26.66, ROUGE-1 of 0.680, and ROUGE-L of 0.459, all of which are the highest among the eight prompts. The essays in prompt 6 are short, source-dependent essays, making it easier for the model to capture their structure and core content. In contrast, prompt 8 exhibits the lowest performance in terms of BLEU (17.87) and ROUGE-L (0.389). Essays in prompt 8 have the longest average length and are narrative in nature, involving extensive details and complex structures. The diversity of content and longer text length likely contribute to the increased difficulty for feedback generation.

Prompt	Bleu (%)	Rouge-1	Rouge-L
P1	18.05	0.658	0.393
P2	18.04	0.652	0.385
P3	21.70	0.668	0.426
P4	20.02	0.644	0.402
P5	22.73	0.672	0.427
P6	26.66	0.680	0.459
P7	21.73	0.625	0.413
P8	17.87	0.632	0.389

Table 6: Model performance of feedback generation across prompts.

Essay id	Generated Feedback	Predicted Score	True Score
4454 *(P2)*	*The essay lacks a clear central idea. The ideas pre-sented are simplistic and lack depth... There are instances of unclear or inappropriate word choices……*	w/o RL: 2 w/ RL: 1	1 *(Range: 0-6)*
19278 *(P7)*	*The essay demonstrates a strong understanding of the prompt and effectively conveys a story with clear anddetailed descriptions. The use of dialogue and languagehelps to……*	w/o RL: 22 w/ RL: 24	24 *(Range: 0-30)*

Figure 5: Comparison of predicted scores with and without reinforcement learning (RL)

E Scoring Optimization Based on Feedback

Figure 5 presents a case study analyzing the model's predicted scores before and after applying reinforcement learning (RL) optimization under both low- and high-score scenarios. For the low score scenario, in Essay 4454, the RL optimization effectively guides the model to adjust the score based on the negative feedback, imposing stricter penalties for severe issues and preventing score overestimation. For the high score scenario, in Essay 19278, the RL optimization enables the model to better leverage the positive feedback, accurately evaluating the strengths of the essay and avoiding score underestimation. These results demonstrate that the RL mechanism performs effectively under both positive and negative feedback conditions, enhancing the dynamic interaction between feedback and scoring through reward and penalty mechanisms. Consequently, the RL-based approach improves the scoring reliability of our model.