Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070211
ISSN 2516-5003 Vol. 7 Num. 2

Prompt-Aware DeBERTa with Lightweight Discourse Modeling and Ordinal Contrastive Learning for Automated English Essay Scoring

Jiaying Gong

School of Foreign Languages, Nanjing Normal University Taizhou College, Taizhou, Jiangsu, China
**Corresponding author: 19852199022@163.com*

Keywords: Automated Essay Scoring; DeBERTa; Discourse Modeling; Ordinal Contrastive Learning

Abstract: To address challenges in Automated Essay Scoring (AES)—specifically regarding the alignment between writing prompts and essays, the representation of discourse structure, and the modeling of relationships between scoring levels—this paper proposes an AES method based on DeBERTa (Decoding-enhanced BERT with Disentangled Attention) that integrates prompt-awareness, lightweight discourse enhancement, and ordinal contrastive learning. First, a pre-trained language model is employed to jointly encode the writing prompt and the essay text, with prompt features pooled and projected to strengthen the thematic alignment. Second, paragraph-level attention and relative position modeling are introduced to extract features related to discourse organization, with multi-level information fused via residual connections. Finally, by combining ordinal threshold prediction with ordinal contrastive learning, the model establishes a mapping between the semantic representation of an essay and its assigned score level. Experiments were conducted on a dataset of 2,160 English essays, evaluated according to a structured scoring rubric, with various model configurations tested for performance comparison and ablation analysis. The results demonstrate that the full model achieves a Quadratic Weighted Kappa (QWK) of 0.965397 on a test set of 720 essays, with a Mean Absolute Error (MAE) of 0.173611 and a Root Mean Square Error (RMSE) of 0.416667. Compared to the baseline DeBERTa model, there is a slight improvement in scoring consistency—though the gains are limited—and the contributions of individual modules vary. These findings offer insights into feature fusion and the design of ordinal scoring mechanisms for automated essay assessment within controlled experimental settings.

1. Introduction

With the continuous advancement of Natural Language Processing (NLP) technology, Automated Essay Scoring (AES) has emerged as a significant research area within the field of intelligent education [1]. Traditional essay grading relies primarily on human evaluation; while this approach allows for a comprehensive assessment of linguistic expression and content organization, it faces

limitations regarding efficiency, cost, and inter-rater reliability [2]. In recent years, deep learning has introduced new paradigms for text comprehension and automated evaluation. Notably, methods based on pre-trained language models can extract richer semantic information from context, providing robust technical support for AES [3]. However, the quality of an essay is not determined solely by vocabulary and grammar; it is also deeply influenced by how well the content aligns with the assigned topic, the logical relationships between paragraphs, and the overall coherence of the writing [4]. Consequently, exploring how to effectively integrate semantic content, topical relevance, and discourse structure to achieve more accurate and reasonable automated scoring remains a valuable area for further research.

While existing deep learning-based AES methods can, to some extent, learn the mapping between text and score levels, several challenges persist. First, essay scoring must be contextualized within the specific writing prompt [5]. If a model relies exclusively on the essay text itself, it may struggle to distinguish between a well-written piece that deviates from the prompt and one that effectively addresses the assigned topic [6,7]. Second, essays possess a hierarchical discourse structure, where individual paragraphs serve distinct functions—such as introducing a topic, developing an argument, or summarizing points. Relying solely on a holistic semantic representation of the entire essay may fail to capture the organizational relationships between its constituent paragraphs. Finally, essay scores follow an ordinal scale, where the difference between adjacent score levels is typically smaller than that between more distant levels. Treating scores as independent, categorical labels ignores these inherent ordinal relationships, which can compromise the validity of the final scores [8]. Therefore, it is essential to further investigate AES methods by focusing on three key dimensions: topical alignment, discourse structure, and the ordinal nature of score levels.

To address these challenges, this paper proposes an automated essay scoring method that integrates topical prompt information, lightweight discourse modeling, and ordinal contrastive learning. For the foundational model, we employ DeBERTa (Decoding-enhanced BERT with disentangled attention) to extract contextual semantic features from essays. By explicitly distinguishing between the writing prompt and the essay body within the input representation, the model is enabled to leverage both the prompt requirements and the content of the essay. Building on this, we introduce a lightweight discourse modeling module to represent the content of individual paragraphs and their relative positions, integrating this information into the overall essay representation via residual fusion. Furthermore, by accounting for the ordinal relationships between scoring levels, we design a corresponding score prediction mechanism and a contrastive learning objective, ensuring that the model considers the distinctions between different score tiers during the feature-learning process. Through this design, we aim to explore the synergy between prompt information, discourse features, and the ordinal nature of scores, ultimately establishing a comprehensive framework for automated English essay scoring while maintaining controlled model complexity.

In this study, we conduct comparative experiments and component analyses to examine how different model architectures influence scoring consistency and prediction error. We also evaluate the specific contributions of individual components by analyzing the training process, score distributions, and prediction outcomes. The model's performance is assessed using the Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) as primary evaluation metrics, focusing on both scoring consistency and prediction deviation.

2. Proposed Method

2.1. Problem Formulation

Automated essay scoring aims to predict the quality level of a composition by considering both the writing prompt and the essay content [9-10]. Let q denote the writing prompt, x the corresponding

English essay, and y its reference score, where $y \in \{1,2,\ldots,6\}$. The scoring task is formulated as

$$\hat{y} = f_\theta(q, x) \tag{1}$$

where f_θ is the essay-scoring model parameterized by θ, and $\hat{y}$ is the predicted score. The model jointly captures prompt semantics, discourse structure, and ordinal score relationships to establish a mapping between essay content and score levels. Reference scores in this study are defined using predefined evaluation rules; the corresponding scoring settings are described in the experimental section [11].

2.2. Overall Architecture

To jointly exploit topic-related and discourse-level information, this study presents a DeBERTa-based essay-scoring framework that integrates prompt-aware representation learning, lightweight discourse modeling, and ordinal contrastive learning. As illustrated in Fig. 1, the writing prompt and essay are first encoded together, with dedicated markers identifying the prompt, essay body, and paragraph boundaries. A pretrained DeBERTa encoder then extracts contextual representations. Prompt features are incorporated into the global textual representation, while paragraph attention and positional relationships are used to capture discourse information. The resulting features are combined through residual fusion and passed to an ordered-threshold scoring head. During training, an additional ordinal contrastive objective uses score differences to constrain distances between essay representations, encouraging the model to capture relationships among score levels.

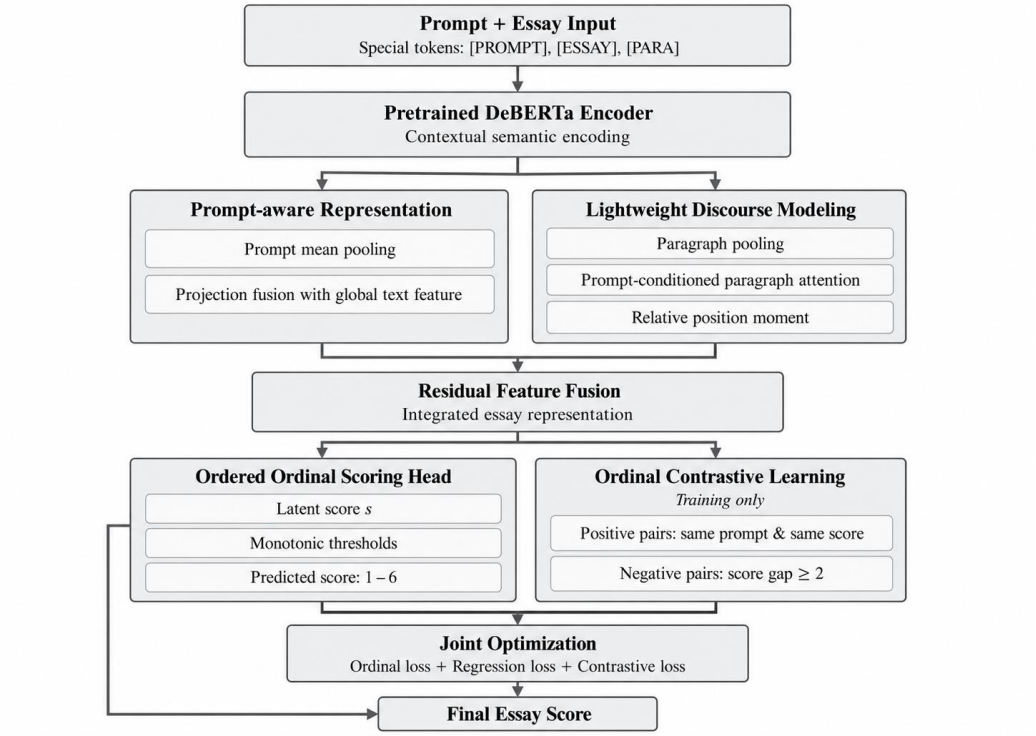


Figure 1: Overall Framework of the Entire Text

2.3. Prompt-Aware Representation Learning

Because essay quality must be assessed against the assigned writing task, a prompt-aware mechanism is employed to represent the relationship between the prompt and its response. The input sequence uses [PROMPT], [ESSAY], and [PARA] markers to distinguish the prompt, essay body,

and paragraph boundaries, respectively. After DeBERTa encoding, the contextual feature matrix is $H \in \mathbb{R}^{L \times d}$, where L denotes sequence length and d the feature dimension. Mean pooling over valid prompt tokens produces a prompt representation, which is linearly projected and fused with the global representation:

$$\begin{aligned} H &= DeBERTa(q, x) \\ \mathrm{q} &= \frac{1}{|M_q|} \sum_{i \in M_q} H_i \\ \mathrm{b} &= H_{CLS} + W_q \mathrm{q} + \mathrm{b}_q \end{aligned} \tag{2}$$

Here, M_q is the set of valid prompt-token positions; H_{CLS} is the representation at the first sequence position; and W_q and $\mathbf{b}_q$ are learnable projection weights and bias. With this design, prompt information participates not only in the encoder's contextual interactions but also directly in the final scoring representation, providing a foundation for the subsequent discourse fusion.

2.4. Lightweight Discourse Modeling

After obtaining the prompt-aware global representation, the model further considers the organization of paragraphs within the essay. A lightweight discourse module aggregates paragraph features and models their relative positions. The contextual token features of each valid paragraph are first mean-pooled to obtain $\mathbf{p}_j$ [12-13]. Prompt-conditioned paragraph attention is then used to calculate the relative contribution of each paragraph:

$$\begin{aligned} e_j &= \mathrm{w}_a^T \tanh(W_k \mathrm{p}_j + W_a \mathrm{q} + \mathrm{c}_a) \\ a_j &= \frac{\exp(e_j)}{\sum_{r=1}^{n} \exp(e_r)} \\ \mathrm{u} &= \sum_{j=1}^{n} a_j \mathrm{p}_j \end{aligned} \tag{3}$$

Here, n denotes the number of valid paragraphs, e_j the attention logit for paragraph j, a_j its normalized attention weight, and $\mathbf{u}$ the aggregated discourse representation. To preserve paragraph-order information, a normalized relative-position coefficient and a paragraph positional moment are further defined as

$$\begin{aligned} \rho_j &= \frac{2(j-1)}{\max(n-1,1)} - 1 \\ \mathrm{m} &= \frac{\sum_{j=1}^{n} \rho_j \mathrm{p}_j}{\max(1, \sum_{j=1}^{n} |\rho_j|)} \end{aligned} \tag{4}$$

where ρ_j indicates the relative paragraph position and $\mathbf{m}$ describes how paragraph features vary across the essay. The attention-pooled representation and positional moment are concatenated and transformed into discourse features [14-15]. A learnable residual coefficient then controls their contribution to the prompt-aware representation:

$$\begin{aligned} \mathrm{r} &= LN(W_2 GELU(W_1[\mathrm{u}; \mathrm{m}] + \mathrm{b}_1) + \mathrm{b}_2) \\ \alpha &= 0.2 \tanh(a) \\ \mathrm{f} &= LN(Dropout(\mathrm{b} + \alpha \mathrm{r})) \end{aligned} \tag{5}$$

In these expressions, LN denotes layer normalization, GELU is the nonlinear activation function, α controls the strength of discourse fusion, and $\mathbf{f}$ is the resulting essay representation.

2.5. Ordered Ordinal Scoring

Following discourse fusion, the model obtains an essay representation that includes prompt semantics and paragraph information. Since essay scores follow a natural order, an ordered-threshold scoring head is adopted to predict cumulative probabilities over score levels. A linear transformation first produces a latent scoring value s, while strictly increasing thresholds ensure the correct ordering of the cumulative probabilities:

$$\begin{aligned} s &= \mathrm{w}_s^T \mathrm{f} + b_s \\ c_k &= c_1 + \sum_{j=1}^{k-1} \left(\ln(1 + e^{g_j}) + 10^{-4} \right) \\ P(y > k) &= \sigma(s - c_k) \\ \hat{y}_c &= 1 + \sum_{k=1}^{5} P(y > k) \end{aligned} \tag{6}$$

Here, c_1 is the initial threshold, g_j denotes a learnable parameter, and the recursive threshold expression applies to $k = 2, \ldots, 5$. The function σ denotes the sigmoid activation, and $\hat{y}_c$ is the continuous expected score. Because the thresholds are strictly increasing, the predicted cumulative probabilities are guaranteed to be non-increasing as the score threshold rises. The continuous estimate is subsequently rounded to the nearest integer and clipped to the range from 1 to 6 to obtain the discrete score.

2.6. Ordinal Contrastive Learning

Although ordered scoring captures the ranking of score levels, prediction losses alone may not sufficiently separate textual representations associated with different writing-quality levels. An ordinal contrastive objective is therefore introduced. The fused essay representation is projected into a normalized, lower-dimensional embedding space, and sample pairs are selected according to their score differences. Two different essays responding to the same prompt and receiving the same score form a positive pair. Essays separated by at least two score levels form negative pairs, whereas adjacent-score pairs do not contribute to this objective.

For normalized essay embeddings $\mathbf{z}_i$, a scaled cosine distance and score-dependent margin are defined as

$$\begin{aligned} d_{ij} &= \frac{1 - clip(\mathrm{z}_i^T \mathrm{z}_j, -1, 1)}{2} \\ \mu_{ij} &= \frac{0.6(|\, y_i - y_j \,| - 1)}{5} \end{aligned} \tag{7}$$

where d_{ij} is the distance between essay embeddings and μ_{ij} is the margin determined by the score gap. The ordinal contrastive loss is then defined as

$$\mathrm{L}_{OCL} = \frac{1}{2\,|\,\mathrm{P}\,|} \sum_{(i,j)\in \mathrm{P}} d_{ij}^2 + \frac{1}{2\,|\,\mathrm{N}\,|} \sum_{(i,j)\in \mathrm{N}} \left[\max(\mu_{ij} - d_{ij}, 0) \right]^2 \tag{8}$$

Here, $\mathcal{P}$ and $\mathcal{N}$ denote the sets of valid positive and negative pairs, respectively. If either set is empty, the corresponding loss term is defined as zero. This objective encourages essays sharing the same prompt and score to have more similar representations, while enforcing a minimum separation

between essays with substantially different scores. Only pairs from the current training batch are used; no historical embedding queue is introduced.

2.7. Joint Optimization and Inference

The scoring and representation-learning objectives are optimized jointly. The ordinal classification term employs binary cross-entropy across five score thresholds, while the regression term uses smooth L_1 loss to penalize differences between the continuous predicted and reference scores. The total objective is

$$\begin{aligned} \mathrm{L} &= \mathrm{L}_{ord} + 0.1\mathrm{L}_{reg} + \lambda_t \mathrm{L}_{OCL} \\ \lambda_t &= 0.02 \min(1, \frac{t}{T_e}) \end{aligned} \tag{9}$$

Here, $\mathcal{L}_{\mathrm{ord}}$ is the ordinal classification loss, $\mathcal{L}_{\mathrm{reg}}$ is the smooth L_1 regression loss, and λ_t is the time-dependent weight for ordinal contrastive learning; t denotes the current training-step index, and T_e is the number of batches in the first epoch. Gradually increasing the contrastive weight allows the model to establish an initial scoring relationship before introducing score-aware embedding constraints. During inference, only the prompt and essay are required: the ordered scoring head produces the final prediction without using reference scores or additional sample pairs.

3. Experimental Setup

3.1. Experimental Dataset and Scoring Criteria

To evaluate the scoring performance of the proposed method, the experiments use 2,160 English essays covering six writing prompts, with scores ranging from 1 to 6. The dataset comprises 1,152 training essays, 288 validation essays, and 720 test essays. The essay texts were generated programmatically using predefined linguistic templates, prompt-specific situations, and expression rules. Sentence patterns, paragraph arrangements, supporting arguments, grammatical errors, and vocabulary use were varied to produce texts of differing quality. Reference scores were computed by programmatic rules rather than assigned by human raters. The scoring criteria comprise relevance (30%), organization (20%), support (20%), grammar (20%), and vocabulary (10%). The final reference score is calculated as

$$w=0.30R+0.20O+0.20S+0.20G+0.10V \tag{10}$$

$$y = \min(6, \max(1, \mathrm{floor}(w + 0.5 + 10^{-9}))) \tag{11}$$

where w is the weighted criterion score, and R, O, S, G, and V denote the five criterion scores, each ranging from 1 to 6. The floor operation rounds the weighted result down to an integer. These reference scores provide a consistent basis for examining how the model learns quality levels and their relative differences.

3.2. Data Split and Leakage Audit

To support an objective evaluation, the samples were partitioned into training, validation, and independent test sets. The training set was used to learn model parameters, whereas the validation set guided checkpoint selection and training adjustments. The test set was evaluated only after model selection. Distinct scenario and opening-template settings were used across the splits, and normalized texts were checked for exact duplication. No identical essay bodies were detected across the three

sets. Nevertheless, some sentence structures and topic-specific expressions were shared. The mean nearest-neighbor text similarity between test and training essays, measured using term frequency–inverse document frequency (TF–IDF) features, was 0.8932. Accordingly, the reported findings characterize performance under the present scoring rules and text distribution; they do not establish generalization to authentic student writing.

3.3. Baselines and Ablation Configurations

To examine the effects of the proposed components, the experiments included a basic DeBERTa model, prompt-aware scoring models, and complete fusion models. Additional configurations combined prompt awareness with discourse enhancement or ordinal contrastive learning. For ordinal contrastive learning, the complete model was evaluated both with a historical-sample queue and with pairs drawn only from the current training batch. Eight model configurations were considered in total. Some models served as independently trained reference systems, while others were fine-tuned from the same prompt-aware checkpoint to limit initialization differences. Checkpoints were selected using quadratic weighted kappa (QWK) on the validation set, with the initial checkpoint retained as an eligible candidate. If additional training did not improve validation performance, the predefined selection rule was followed rather than interpreting the new component as beneficial.

3.4. Evaluation Metrics

Three metrics were used to assess scoring performance: quadratic weighted kappa (QWK), mean absolute error (MAE), and root mean squared error (RMSE). QWK measures agreement between predicted and reference scores, as defined in Eq. (1). MAE measures the average magnitude of score deviations, while RMSE places greater weight on larger errors. The latter two metrics are defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{12}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{13}$$

where N is the number of test essays, and y_i and $\hat{y}_i$ are the reference and predicted scores for essay i, respectively. All three metrics were calculated using discrete predicted scores rounded and constrained to the 1–6 scale. A QWK value closer to 1 indicates stronger scoring agreement, whereas lower MAE and RMSE values indicate smaller prediction errors.

3.5. Implementation Details

Experiments were conducted on a computer equipped with an NVIDIA GeForce RTX 3060 GPU (12 GB) using Python 3.9, PyTorch 2.0.0, and CUDA 11.8. DeBERTa-v3-base served as the semantic encoder. The batch size was 4, with gradient accumulation over 2 steps. The maximum input length was 320 tokens for training and 352 tokens for standardized evaluation. AdamW was used for optimization with a weight decay of 0.01. During continued fine-tuning, the shared parameters used a learning rate of 5×10^{-6}, while newly introduced modules used 1×10^{-4}. The maximum ordinal contrastive-loss weight was 0.02, increased gradually during training. Training continued for at most six epochs, with early stopping and checkpoint selection based on validation performance. The random seed was fixed at 2026. Metrics were computed from actual predictions, with training logs and checkpoint-selection records retained for traceability.

4. Results and Discussion

4.1. Overall Performance Comparison

Eight model configurations were evaluated on the same held-out set of 720 essays. The independently trained reference systems are denoted by the B group, whereas the configurations fine-tuned from a shared prompt-aware checkpoint are denoted by the P group. Table 1 reports their QWK, MAE, and RMSE values, and Fig. 2 highlights differences in scoring agreement.

Table 1: Performance comparison of different essay-scoring configurations.

Model	QWK	MAE	RMSE
B1 Baseline DeBERTa	0.965133	0.176389	0.419987
B2 Prompt-aware reference	0.964160	0.175000	0.418330
B3 Full-architecture reference	0.961133	0.197222	0.444097
P0 Prompt-aware control	0.963395	0.186111	0.431406
P1 Prompt + discourse	0.963796	0.176389	0.419987
P2 Prompt + contrastive	0.964160	0.175000	0.418330
P3 Full model (queue)	0.963796	0.176389	0.419987
P4 Full model (no queue)	0.965397	0.173611	0.416667

As shown in Table 1, P4 attained the best observed values on all three metrics, with a QWK of 0.965397. Relative to P0, its QWK increased by $\Delta\text{QWK} = 0.002002$, while MAE fell from 0.186111 to 0.173611, a reduction of approximately 6.72%. Compared with B1, however, the QWK gain was only 0.000264 and the MAE reduction was 0.002778. B3 yielded a QWK of 0.961133, lower than that of B1, suggesting that adding architectural components did not consistently improve scoring. As Fig. 2 indicates, all QWK values lay between approximately 0.961 and 0.966; the small observed differences should not be interpreted as evidence of stable superiority.

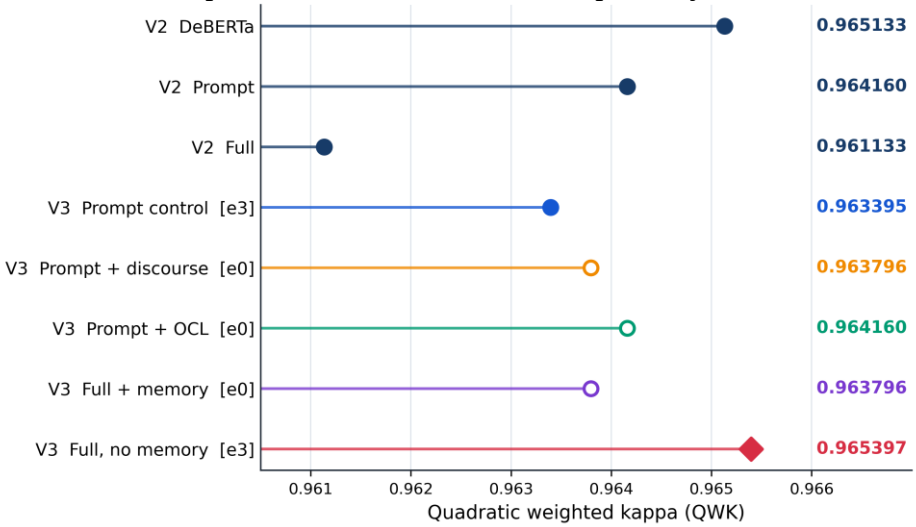


Fig. 2: QWK comparison of eight essay-scoring configurations.

4.2. Component Ablation and Checkpoint Selection

To isolate the effects of the components, the continuation experiments were compared under a shared initialization. Table 2 summarizes the number of completed fine-tuning epochs, the checkpoint selected by validation performance, and the highest QWK reached after training. Fig. 3 presents the test-set QWK differences relative to the prompt-aware continuation control.

Table 2. Validation results and selected checkpoints for the continuation experiments.

Model	Epochs run	Selected epoch	Selected QWK	Best trained QWK
P0 Prompt-aware control	5	3	0.965084	0.965084
P1 Prompt + discourse	3	0	0.964394	0.959972
P2 Prompt + contrastive	3	0	0.964394	0.963729
P3 Full (queue)	3	0	0.964394	0.964265
P4 Full (no queue)	5	3	0.966062	0.966062

P1, P2, and P3 did not exceed their initialization checkpoints on validation data and therefore selected epoch 0. In particular, the best post-training validation QWK of the discourse-enhanced variant was 0.959972, which was 0.004422 below its initialization value. By contrast, P4 reached a validation QWK of 0.966062 at epoch 3, exceeding the P0 result by 0.000978. Fig. 3 shows a test-set QWK gain of 0.002002 for P4 over P0. Although the remaining configurations exhibit small numerical differences, their selected predictions came from initialization checkpoints rather than successfully trained added branches. The findings therefore characterize the interaction between model configuration and checkpoint selection, not independent improvements attributable to every component.

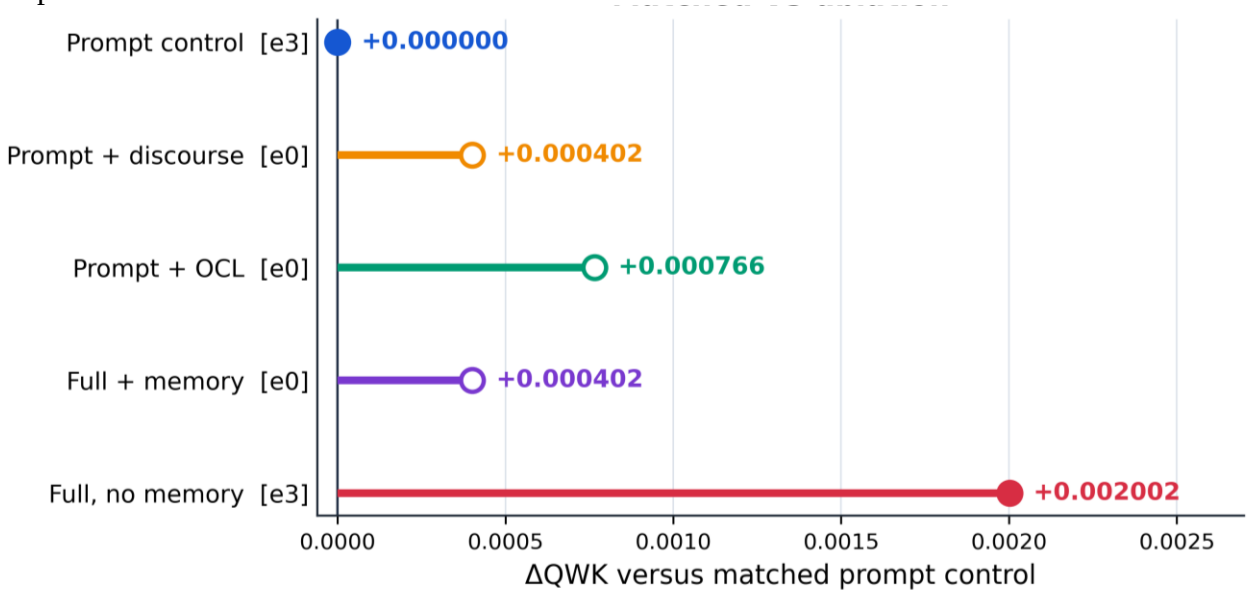


Fig. 3: QWK changes relative to the prompt-aware control (P0). Open markers denote initialization checkpoints.

4.3. Training and Validation Convergence

The training trajectories provide additional context for these results. Figure 4 shows validation QWK throughout continued fine-tuning, with epoch 0 representing the checkpoint before further updates. All five configurations experienced a pronounced decline after their first training epoch and subsequently recovered to varying degrees.

For P4, validation QWK rose from 0.937743 at epoch 1 to 0.966062 at epoch 3, before falling to 0.957158 at epoch 5. Over the same interval, its training loss continued to decline from 0.241083 to 0.231664, demonstrating that an improved training objective did not consistently translate into stronger scoring agreement. P0 likewise achieved its maximum validation QWK of 0.965084 at epoch 3 and subsequently deteriorated. The convergence patterns in Fig. 4, together with Table 2, support

selecting checkpoints by validation QWK rather than using the final epoch indiscriminately. They also indicate that the effectiveness of the additional branches remains sensitive to the available texts and optimization conditions.

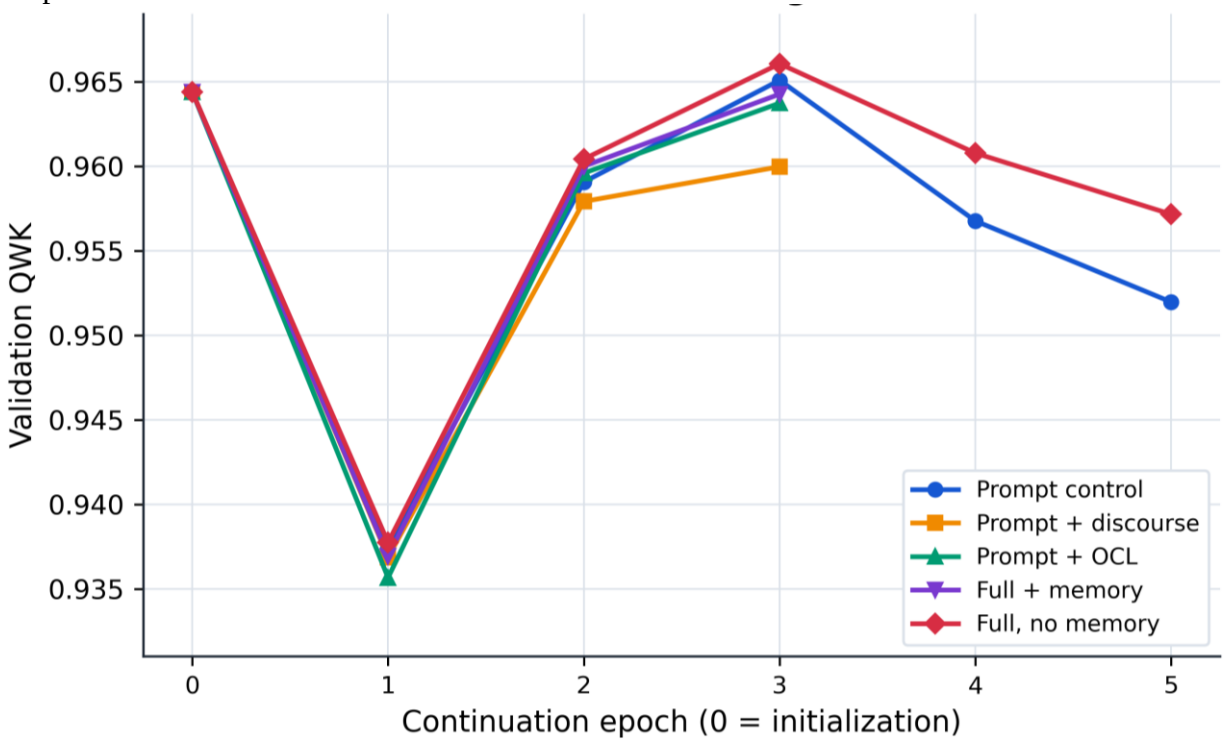


Fig. 4: Validation QWK trajectories during continued fine-tuning.

4.4. Prediction and Error Analysis

Figure 5 displays the test-set confusion matrix for P4, which was selected in advance using validation performance. Of the 720 essays, 595 received the exact reference score, corresponding to an accuracy of $595/720 \approx 82.64\%$. The remaining 125 essays were all assigned an adjacent score: 56 were underestimated and 69 were overestimated. No prediction differed from its reference score by two or more levels; thus, for every misclassified essay, $|y_i - \hat{y}_i| = 1$.

In Fig. 5, all 62 essays assigned a reference score of 1 and all 83 essays assigned a reference score of 6 were correctly identified, whereas most confusions occurred in the intermediate levels. Specifically, 124 essays with a reference score of 2 were correctly classified and 31 were predicted as 1; among essays scored 4, 107 were correctly classified and 24 were predicted as 5; among those scored 5, 110 were correctly identified and 29 were predicted as 6. Across the six writing prompts, per-prompt MAE ranged from 0.1417 to 0.2000. The predominance of adjacent-level errors suggests that the model generally preserved the ordering imposed by the current reference-scoring rules, although distinctions between intermediate score categories remain imperfect.

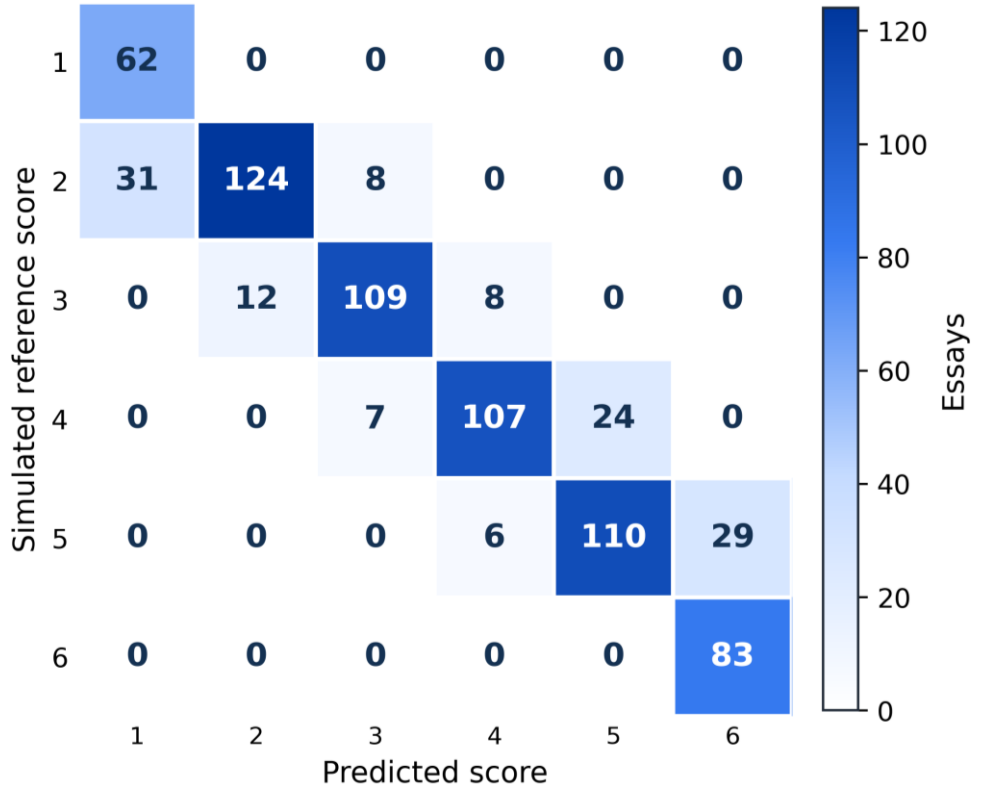


Fig. 5: Confusion matrix of the validation-selected model on 720 test essays.

5. Conclusion

This paper addresses the challenge of jointly modeling the relationships among topical relevance, discourse organization, and scoring levels in automated essay scoring (AES). We propose a DeBERTa-based scoring method that integrates prompt awareness, lightweight discourse enhancement, and ordinal contrastive learning. Through model comparisons, component analysis, and an investigation of the training process, we verify the feasibility of the proposed framework under established experimental conditions. Our findings indicate that prompt information provides a valuable semantic reference for assessing essay quality, while the impact of discourse features and ordinal contrastive constraints is contingent upon the model's integration approach and training strategy. Increasing model complexity does not necessarily yield better scoring performance; rather, it is equally important to effectively manage the relationships between different features and to select appropriate model parameters. This work offers an empirical reference for automated scoring methods that incorporate multidimensional linguistic features, and it underscores that the design of a scoring model must balance semantic expressive capacity, training stability, and predictive accuracy.

References

[1] Li S, Ng V. Automated essay scoring: A reflection on the state of the art[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 17876-17888.

[2] Xu J, Liu J, Lin M, et al. EPCTS: Enhanced prompt-aware cross-prompt essay trait scoring[J]. Neurocomputing, 2025, 621: 129283.

[3] Boquio E N V, Naval Jr P C. Beyond canonical fine-tuning: Leveraging hybrid multi-layer pooled representations of BERT for automated essay scoring[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024: 2285-2295.

[4] Madnani N, Mulcaire P, Lo K L. Applications of large language models in automated writing evaluation[M]//The Routledge Handbook of Digital Language Assessment. Routledge, 2026: 153-170.
[5] Jiang Z, Gao T, Yin Y, et al. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 12456-12470.
[6] Do H, Kim Y, Lee G G. Prompt-and trait relation-aware cross-prompt essay trait scoring[C]//Findings of the association for computational linguistics: ACL 2023. 2023: 1538-1551.
[7] Chen Y, Li X. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring[C]//Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). 2023: 1489-1503.
[8] Chen Y, Li X. Plaes: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring[C]//Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024). 2024: 12775-12786.
[9] Bevilacqua M, Oketch K, Qin R, et al. When automated assessment meets automated content generation: Examining text quality in the era of gpts[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-36.
[10] Misgna H, On B W, Lee I, et al. A survey on deep learning-based automated essay scoring and feedback generation: H. Misgna et al[J]. Artificial Intelligence Review, 2024, 58(2): 36.
[11] Gunjal A, Wang A, Lau E, et al. Rubrics as rewards: Reinforcement learning beyond verifiable domains[C]//International Conference on Learning Representations. 2026, 2026: 127924-127945.
[12] Zhang E, Poellabauer C. Mitigating Interviewer Bias in Multimodal Depression Detection: An Approach with Adversarial Learning and Contextual Positional Encoding[C]//EMNLP (Findings). 2025: 12169-12188.
[13] Sampaio P R, Maxcici H. Unsupervised document and template clustering using multimodal embeddings[J]. International Journal on Document Analysis and Recognition (IJDAR), 2026: 1-29.
[14] Susladkar O K, Deshmukh G S, Gorade V, et al. GRIZAL: Generative prior-guided zero-shot temporal action localization[C]//Proceedings of the 2024 conference on empirical methods in natural language processing. 2024: 19046-19059.
[15] Barnfield N, Cui H, Lu Y. High-dimensional analysis of single-layer attention for sparse-token classification[C]//International Conference on Learning Representations. 2026, 2026: 40813-40871.