Improving Prompt Generalization for Cross-prompt Essay Trait Scoring from the Scoring-invariance Perspective

Jiong Wang[1] **and Shengquan Yu**[2]*
[1] School of Artificial Intelligence, Beijing Normal University, China
[2] Advanced Innovation Center for Future Education, Beijing Normal University, China
jjoinw@mail.bnu.edu.cn, yusq@bnu.edu.cn

Abstract

Cross-prompt trait scoring task aims to learn generalizable scoring capabilities from source-prompt data, enabling automatic scoring across multiple dimensions on unseen essays. Existing research on cross-prompt trait essay scoring primarily focuses on improving model generalization by obtaining prompt-invariant representations. In this paper, we approach the research problem from a different perspective on invariance learning and propose a scoring-invariant learning objective. This objective encourages the model to focus on intrinsic information within the essay that reflects its quality during training, thereby learning generic scoring features. To further enhance the model's ability to score across multiple dimensions, we introduce a trait feature extraction network based on routing gates into the scoring architecture and propose a trait consistency scoring objective to encourage the model to balance the diversity of trait-specific features with scoring consistency across traits when learning trait-specific essay features. Extensive experiments demonstrate the effectiveness of our approach, showing advantages in multi-trait scoring performance and achieving significant improvements with low-resource prompts.

1 Introduction

Automated essay scoring (AES) is a common and important application of artificial intelligence (AI) in the field of education. Compared to the human grading process, an effective AES system can provide language learners with timely feedback on multiple perspectives, such as overall essay quality, organization, prompt adherence, narrativity, etc.

In previous AES research, most studies focused on a prompt[1]-specific setting, where both the training and unrated test essays belong to the same prompt (Taghipour and Ng, 2016; Dong et al., 2017; Wang et al., 2018; Kumar et al., 2022; Xie et al., 2022). However, real-world AES systems often lack enough target-prompt essays, making it crucial to develop methods that can reliably score essays for prompts not included in the training data.

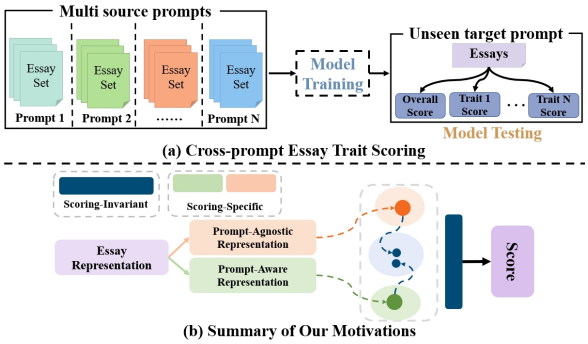


Figure 1: The illustration of cross-prompt essay trait scoring and a summary of our main motivations.

To address the aforementioned limitation, researchers have begun focusing on cross-prompt essay scoring (Ridley et al., 2020, 2021). Similar to the objective of domain generalization (DG) (Wang et al., 2022a), this task aims to enable AES models trained on source prompts to be effectively applied to "unseen" prompts, as shown in Figure 1 (a). In our work, we also focus on scoring essays while considering their trait-specific attributes.

Although existing studies have made progress in enhancing prompt generalization (Do et al., 2023; Chen and Li, 2023), most of them approach the problem from the perspective of *prompt-invariance*, aiming to improve model generalization by aligning features across essays from multiple source prompts. Jiang et al. 2023 proposed a disentangled learning approach to extract prompt-invariant representations from source prompts. Chen and Li 2024 proposed a meta-learning-based approach that simulates prompt distribution shifts to enhance the generalization of essay representations. These existing studies require constructing a large amount of coun-

*Corresponding author
[1]Prompt represents the writing theme and genre of essay.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2633–2646
November 4-9, 2025

terfactual data (Jiang et al., 2023) or meta-learning tasks (Chen and Li, 2023) to learn prompt-invariant representations effectively, which significantly increases the training burden.

In this work, we take a different perspective on invariance learning and propose a **scoring-invariant** learning approach. This approach encourages the model to focus on features relevant to scoring, which arise from the internal structure of the essays rather than prompt differences, thereby capturing the intrinsic scoring information. The motivation is illustrated in Figure 1 (b).

For an essay, whether obtaining the essay representation from its grammatical structure or using semantic information to extract the representation, there should be certain dimensions in both representations that can jointly reflect the essay's quality. Since these dimensions remain unchanged regardless of the method used to extract essay features, we refer to them as scoring-invariant dimensions. In this work, we first extract essay representations from two dimensions (prompt-aware and prompt-agnostic). Then, through the scoring-invariant objective we designed, we optimize the model's training process to encourage the model to focus on the intrinsic scoring features within the essay.

For multi-trait scoring tasks, learning diverse trait-specific scoring features and consistency information across traits is beneficial in promoting the model's ability to score across multiple dimensions (Do et al., 2023; Wang and Liu, 2025). Therefore, to address trait diversity, we introduce a sentence-level routing gate network based on the existing scoring architecture, leveraging the modeling approach of a mixture-of-experts (MoE) to enhance the model's perception of target trait features (Zhang et al., 2024; Liu et al., 2024). Regarding trait consistency information, we further propose a trait consistency learning objective, aiming to enable the model to further capture the relationships between traits. Finally, we named our method as Scoring-invariance Enhanced Cross-prompt Trait Scoring (**SICTS**). To summarize, the main contributions of our work can be summarized as follows:

1) From the perspective of scoring invariance, we propose a simple yet effective method to enhance the out-of-distribution transferability of cross-prompt scoring models by facilitating the learning of generalizable essay representations.

2) To enhance the diversity of trait-specific scoring features, we introduce a routing gate network at the shared encoding layer. Additionally, to achieve the optimization goal of multi-trait scoring, we introduce a trait consistency optimization objective at both the trait-representation and score levels.

3) Extensive experiments on the public datasets demonstrate that our approach outperforms the baseline method. And our method achieves higher scoring performance with fewer model parameters.

2 Related Work

In the early research on AES, most studies focused on prompt-specific scoring tasks, where the training and testing data originated from the same prompt. Researchers initially relied on handcrafted features and applied machine learning methods to score essays (Rudner and Liang, 2002; Attali and Burstein, 2006; Phandi et al., 2015). With the development of deep learning, more studies have started using deep neural networks to automatically extract quality features from essays, enabling an end-to-end scoring process (Taghipour and Ng, 2016; Dong et al., 2017; Wang et al., 2018; Uto et al., 2020; Wang et al., 2022b; Wang and Liu, 2025). Although existing prompt-specific essay scoring methods have been widely studied and have made some progress, exploring cross-prompt scoring methods can reduce the need for annotations on target prompts, thereby lowering annotation costs and facilitating real-world applications.

Cross-prompt AES aims to train AES models that learn generalizable scoring abilities from source prompts to score essays from unseen target prompt(s) (Jin et al., 2018; Ridley et al., 2021; Do et al., 2023; Chen and Li, 2024). In existing cross-prompt AES research, one class of approaches focuses on evaluating the overall quality of essays. Ridley et al. 2020 apply a neural-network-based method to learn general essay features. Jiang et al. 2023 propose a prompt-aware neural AES model to extract essay quality features. Another line of research focuses on cross-prompt essay trait scoring, Chen and Li 2023 utilize contrastive learning to learn consistent prompt-agnostic representations across different prompts. Do et al. 2023 introduce a cross-prompt trait scoring method that can extract prompt-aware essay representations and topic-coherence features.

Unlike existing cross-prompt AES methods, we improve the prompt generalizability from the perspective of scoring-invariance learning.

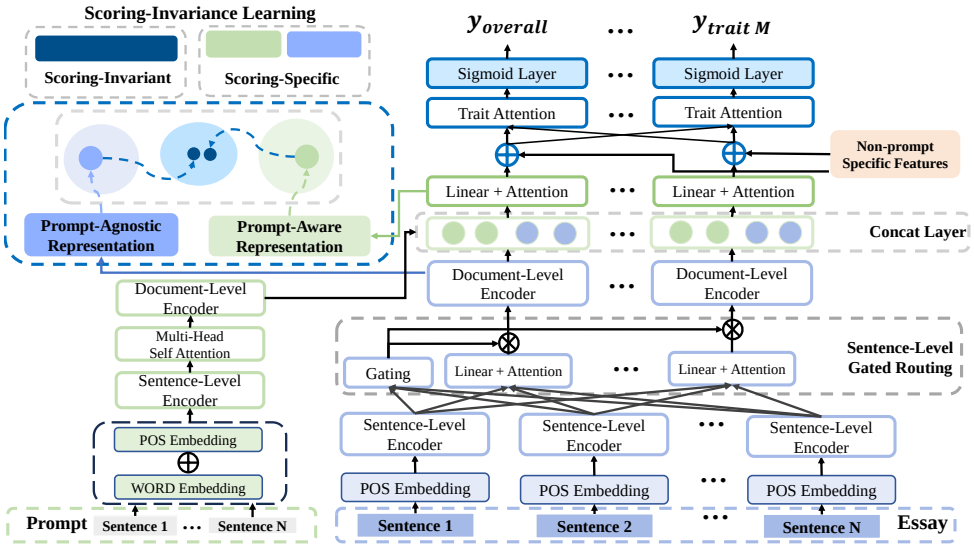


Figure 2: The scoring model architecture with the proposed sentence-level routing gating mechanism and illustrations of our proposed scoring-invariant learning strategies.

3 Task Definition and Preliminary

3.1 Task Definition

Under the prompt generalization setting, the cross-prompt essay trait scoring task can be defined as follows: given essay data from N source prompts $SP = \{SP_1, SP_2, \ldots, SP_N\}$, where each source prompt contains N_i labeled essay instances with M scoring dimensions $\{x_j^S, (y_j^{S_1}, \ldots, y_j^{S_M})\}_{j=1}^{N_i}$, where $y_j^{S_1}$ is the overall score. The objective of our approach is to learn a model from SP that can score essays from an unseen target prompt TP based on both overall quality and trait dimensions.

3.2 Invariant Learning in Cross-prompt AES

Invariant Learning (IL) has long been a popular solution for addressing out-of-distribution generalization problems (Lu et al., 2022; haoxin liu et al., 2024). In domain generalization (DG) tasks, learning domain-invariant representations is crucial since DG primarily focuses on achieving invariance across domains. In cross-prompt AES, previous studies have employed methods such as contrastive learning (Chen and Li, 2023), disentangled representation learning (Jiang et al., 2023), and meta-learning (Chen and Li, 2024; Wang et al., 2025) to achieve domain invariance for prompt generalization tasks. The optimization objective of the above methods is as follows:

$$\min_{\theta} \mathcal{L}(f_\theta(x), y) \quad \text{s.t. } f_\theta(x) \approx f_\theta(x') \qquad (1)$$

where $x \sim P$ and $x' \sim Q$, P and Q represent different prompts, $f_\theta(x)$ typically represents the domain-invariant features extracted from the input data. This objective emphasizes the invariance of the model across different prompt inputs, meaning that even if there is a domain shift in the input data, the model should produce consistent representations or predictions.

Unlike previous methods for obtaining prompt-invariant essay features, we revisit the concept of invariance in cross-prompt AES from a perspective more aligned with the nature of the AES task itself. *We assume that when evaluating the same essay, the scoring model, regardless of whether it extracts general essay features or prompt-specific essay features, will always capture certain features that consistently reflect the essay's quality*. Therefore, these features can be referred to as ***scoring-invariance***. In our work, our aim is to set the optimization objective to explore the invariant features. The objective function can be formulated as:

$$\min_{\theta} \mathcal{L}(f_\theta(x), y), \text{s.t. } \|g_\theta(x) - g_\theta(x')\|_2 \approx 0 \quad (2)$$

where $x \sim P$ and $x' \sim P'$, P and P' represent different types of features that reflect the quality of the same essay. $g_\theta(x)$ represents the scoring-invariant features extracted from the input data.

4 Our Method

4.1 Scoring Model Architecture

In our method, to fully capture the textual representation of essays, we use a hierarchical structure (Dong et al., 2017) to encode them. This hierarchical model architecture has been proven effective for

AES and is widely used in cross-prompt AES tasks (Ridley et al., 2021; Do et al., 2023; Chen and Li, 2023, 2024). The hierarchical encoding framework first utilizes CNN and attention pooling to extract **Sentence-level** features from essay words, and then employs LSTM and another attention pooling to obtain **Document-level** essay features. Additionally, to score essays for multiple traits, similar to existing methods (Ridley et al., 2021; Do et al., 2023), we added trait-specific layers to the base model on top of the shared layer. The shared layer and trait-specific layers are used for sentence-level and document-level essay representations, respectively.

Unlike existing cross-prompt AES models that rely on LSTM (Ridley et al., 2021; Chen and Li, 2024) or multi-head self-attention (Do et al., 2023) to capture document-level features from trait-shared sentence features, we argue that in multi-trait scoring tasks, the document-level features extracted by different trait scoring modules should be distinct. In other words, the shared features emphasized by different trait-scoring models should exhibit some variability. To better capture trait-specific features, we propose a *Sentence-level Routing Gating* network, inspired by the MoE model (Zhang et al., 2024), which enhances feature diversity through selective routing. By using this routing network, we increase the features variability of different traits. The whole illustration of our approach is shown in Figure 2.

4.1.1 Prompt-Agnostic Representation

Shared Encoder Following previous work (Ridley et al., 2021), we utilize part-of-speech (POS) embeddings to represent the generalized essay features. For POS tagging, we employ the NLTK[2] toolkit. Each tagged word is then mapped to a dense vector. To obtain sentence-level representation, a convolutional layer followed by attention pooling (*Sentence-Level Encoder*) is applied. The equations are as follows:

$$c_i = \text{CNN}([w_i : w_{i+k-1}]) \quad (3)$$

$$s = \text{Attention}([c_i : c_m]) \quad (4)$$

where w_i represents the POS embeddings ($i = 1, 2, \ldots, m$), k is the kernel size of the CNN, c_i is the feature representation after the convolutional layer, and s is the final sentence representation.

Trait-specific Encoder Since the scoring dimensions for different traits are not the same, to obtain essay representations that capture diverse trait-specific features, we propose using a sentence-level gated routing network to enhance the diversity of trait-representation. This approach allows different trait scoring network modules to focus more effectively on capturing essay features that are beneficial for the specific trait being evaluated.

Gate weights w are computed by passing sentence representations $S = \{s_1, \ldots, s_n\}$ through a dense layer and applying softmax to derive routing weights for traits:

$$w = \text{softmax}(\text{Dense}(S)) \quad (5)$$

for each trait $j = 1, \ldots, m$, a trait-specific routing weight w^j is extracted and applied element-wise to S, generating a weighted sentence representation $S^j = w^j \cdot S$. Then, the generated trait-specific sentence representations S^j are passed through self-attention mechanism to further extract trait-relevant features. We hypothesize that *combining gated routing with the attention mechanism allows different trait scoring modules not only to capture the corresponding essay features but also to easily grasp the relationships between different dimensions within the essay*. Unlike existing studies (Wang and Liu, 2025), where the mixture-of-experts network was primarily applied at the token level within pre-trained language models, our work focuses on constructing essay representations from sentence-level features. We leverage attention mechanisms to extract relevant features and enhance document-level representation learning.

Next, the document-level essay representation can be extracted by the LSTM and another attention pooling (Ridley et al., 2021) from all trait-specific sentence representations:

$$th_i^j = \text{LSTM}(ts_{i-1}^j, ts_i^j) \quad (6)$$

where th_i^j represents the output of LSTM for the j-th task at the i-th time step, ts^j is the concatenated output of the previous layer, then it is followed by the attention layer (*Document-Level Encoder*).

4.1.2 Prompt-Aware Representation

To achieve the goal of invariance learning in the cross-prompt scoring scenario proposed in this paper, after obtaining prompt-agnostic essay features, we extract scoring features that reflect essay quality from another scoring perspective. Similar to Do et al. 2023, we also encode both the prompt and the essay together to produce representations that more effectively highlight variations in quality (Prompt-Aware Representation).

To obtain prompt-aware essay representations,

[2]http://www.nltk.org

we first extract the corresponding prompt representation for each essay. This involves embedding the prompt using pre-trained GloVe (Pennington et al., 2014) combined with POS embeddings, followed by a convolutional layer with attention pooling, multi-head self-attention with LSTM, and another attention pooling layer. Unlike essay representation extraction, this process focuses on capturing the specific descriptive content of the prompt. Specifically, after obtaining the POS embeddings and word embeddings of the prompt, we first add them together to obtain the prompt embedding encoding with semantic information. Then, through the process of obtaining the textual representation of the essay, we use a sentence-level feature extraction network and a document-level feature extraction network to obtain the final prompt representation. It is important to note that, in order to further capture the dependencies between multiple sentences, we use a multi-head attention mechanism to extract sentence-level features. Consider the task of predicting the j-th trait score. The output is a matrix of sentence representations that serve as the query, key, and value, which denoted as S, :

$$H_j^i = \text{Att}(W_j^{i1}, W_j^{i2}, W_j^{i3}) \tag{7}$$

$$MH(S)_j = \text{Concat}(H_j^1, \ldots, H_j^h)W_j^O \tag{8}$$

here, Att represents the scaled dot-product attention, and H_j^i is the output of the i-th attention head. W_j^{i1}, W_j^{i2}, and W_j^{i3} are the parameter matrices.

After obtaining the prompt representation, unlike Do et al. 2023 using multi-head attention to obtain prompt-aware representation, we combine the prompt representation with the extracted prompt-agnostic essay representation. Then, a linear layer with attention pooling is applied to obtain the prompt-aware essay representation.

By acquiring representations from two different scoring perspectives, we further extract shared features that reflect essay quality from both representations to achieve the optimization goal of scoring-invariance. The specific consistency learning strategy is detailed in Section 4.2.

Trait-specific Scorer Finally, the obtained representations are combined with non-prompt-specific features (Ridley et al., 2021; Do et al., 2023). These features are meticulously engineered to represent various aspects of general essay quality. Following Ridley et al. 2021, we also apply trait-attention to capture relationships among traits. The predicted trait score, $\hat{y}_j$, is then calculated by applying a linear layer with the $sigmoid$ function:

$$\hat{y}_j = sigmoid(w_y^j \cdot z_j + b_y^j) \tag{9}$$

where w_y^j is the trainable weight matrice, b_y^j is the bias vector and z_j is the obtained final trait representation. Given the trait set Y under a prompt, the specific trait set Y_i for each i-th training sample varies depending on its prompt. Therefore, to handle missing gold scores for traits, a masking mechanism is applied (Ridley et al., 2021; Do et al., 2023). For the i-th sample and its j-th trait, the mask is defined as:

$$\text{mask}_{ij} = \begin{cases} 1, & \text{if } Y_j \in Y^i, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

This mask is used to adjust the true values (y_i) and predicted values ($\hat{y}_i$) through element-wise multiplication: $y_i = y_i \otimes \text{mask}_i, \quad \hat{y}_i = \hat{y}_i \otimes \text{mask}_i$. This ensures that only traits with valid gold scores contribute to calculations such as loss and evaluation metrics, avoiding errors from missing annotations and maintaining accuracy.

4.2 Scoring-Invariance Objective

After obtaining essay representations of different types (prompt-aware & prompt-agnostic), next, to enable the scoring model to capture the information that consistently reflects essay quality from both types of trait representations, Z_{agnostic} and Z_{aware}, we propose a novel invariance learning objective for cross-prompt AES. We assume that each perspective in the different types of representation follows a Gaussian distribution. By using this invariance as an optimization target, we encourage the trait representations to remain consistent across perspectives, thereby promoting the model to focus on the features that commonly reflect essay quality across the two distributions. Formally, the learning objectives (Inv) can be formulated as follows:

$$\min \sum_i ||Z_i^{aware} - Z_i^{agnostic}||_2^2,$$
$$s.t. ||Std(Z_i^{aware}) - Std(Z_i^{agnostic})||_2^2 \approx 0 \tag{11}$$

where $Z_i^{agnostic}$ and Z_i^{aware}denote the i-th dimension of two scoring perspective feature matrices. The first term enforces alignment of feature values while the second term constrains the consistency of standard deviations. The aim of this objective is to learn consistent scoring features between prompt-agnostic and prompt-aware essay representations.

By minimizing the consistency between different scoring-perspective features, the model is encour-

aged to focus on more general scoring features, leading to more generalizable trait representations. Considering multi-trait scoring as a scoring task under the multi-task learning paradigm, in the cross-prompt AES task, it is necessary to aim for scoring invariance across multiple traits. Therefore, the final optimization objective for cross-prompt multi-trait scoring invariance in this paper is as follows:

$$\mathcal{L}_{SIL} = \frac{1}{m}\sum_{t}^{m} \text{Inv}(Z_t^{aware}, Z_t^{agnostic}) \quad (12)$$

where m is the number of essay traits.

4.3 Loss Function

In the multi-trait scoring task, different essay traits often exhibit a certain intrinsic scoring consistency. Therefore, constructing such a scoring consistency objective helps improve the model's multi-dimensional scoring capability (Chen and Li, 2023). In existing research, Trait Similarity Loss (TSL) (Do et al., 2023) has been proposed to obtain such consistency from the predicted score level. However, using only target score predictions as the consistency objective makes it challenging to truly align traits in the hidden space, thereby hindering the promotion of writing proficiency across traits. Moreover, this method only learns the consistency among traits, while ignoring the consistency between individual traits and the overall score. In our work, to achieve alignment of essay traits in the latent state space, we further introduce Trait-representation Consistency Loss (TCL). As TSL, we also use the Pearson correlation coefficient (PCC, P) to calculate the consistency between different trait scores, in order to identify trait relationships with strong scoring consistency in the essays. The TCL is formally defined as follows:

$$\mathcal{L}_{TCL} = \log\left(\frac{1}{c}\sum_{j=1}^{M}\sum_{k=j+1}^{M} \text{Cons}(z_j, z_k, y_j, y_k)\right) \quad (13)$$

$$\text{Cons} = \begin{cases} \exp(\cos(z_i, z_j)/\tau), \text{if} P(y_j, y_k) \geq \delta, \\ 0, \quad \text{otherwise.} \end{cases} \quad (14)$$

where $z_j = [z_{1j}, z_{2j}, ...z_{Nj}]$ is extracted j-th trait representation, $y_j = [y_{1j}, y_{2j}, ..., y_{Nj}]$ is j-th ground-truth trait vector, cos denotes the cosine similarity, τ denotes the temperature parameter, δ is the threshold and c is the number of calculated $Cons$ that is not 0.

Finally, the overall loss function L_{Final} is the summation of the ground truth MSE loss $\mathcal{L}_{MSE}$, TSL loss $\mathcal{L}_{TSL}$, SIL loss $\mathcal{L}_{SIL}$, TCL loss $\mathcal{L}_{TCL}$:

$$\mathcal{L}_{Final} = \lambda\mathcal{L}_{MSE} + (1-\lambda)\mathcal{L}_{TSL} + \alpha\mathcal{L}_{SIL} + \beta\mathcal{L}_{TCL} \quad (15)$$

where λ, α,and β are hyperparameters.

5 Experiment

5.1 Datasets and Evaluation

Datasets. We conduct experiments on Automated Student Assessment Prize (ASAP[3]) and ASAP++ (Mathias and Bhattacharyya, 2018) dataset, which have been widely used for AES (Ridley et al., 2021; Chen and Li, 2024). The dataset contains eight sets of essays, each corresponding to a different essay prompt. More details are provided in Appendix A.

Evaluation. Following research on cross-prompt AES (Do et al., 2023; Chen and Li, 2024), to evaluate the performance of our model, we use the widely-adopted Quadratic Weighted Kappa (QWK) metric (Cohen, 1968).

5.2 Implementation Details

In our method, to ensure fairness in the experiment results, the model parameters in our model are set consistent with those in previous studies (Ridley et al., 2021; Do et al., 2023). The dropout rate was set to 0.5, with the CNN filter size and kernel size configured as 100 and 5, respectively. LSTM units were set to 100, while the POS embedding dimension was set to 50, and the batch size was 10. For multi-head attention layer, we used two heads with an embedding dimension of 100. Among the 50 training epochs, we selected the one with the highest average QWK score across all traits in the development set for testing. The number of parameters in our model is ***1.86M***. More experimental details are provided in Appendix B.

5.3 Baseline Methods

We compare our method with the following methods under the cross-prompt essay scoring setting. Fine-tuning **BERT** (Devlin et al., 2019) to evaluate the cross-prompt feature scoring capability of pre-trained models. **PAES** (Ridley et al., 2020): This model is based on the CNN-LSTM architecture (Dong et al., 2017) and models each attribute separately for multi-attribute scoring. **CTS** (Ridley et al., 2020): Building on PAES, this model proposed a trait-attention mechanism to establish interactions between different traits. **PMAES**

[3] https://www.kaggle.com/c/asap-aes

Model	P1	P2	P3	P4	P5	P6	P7	P8	AVG↑
BERT (Devlin et al., 2019)	0.542	0.546	0.574	0.603	0.630	0.459	0.256	0.235	0.481
PAES (Ridley et al., 2020)	0.605	0.522	0.575	0.606	0.634	0.545	0.356	0.447	0.536
CTS (Ridley et al., 2021)	0.623	0.540	0.592	0.623	0.613	0.548	0.384	0.504	0.553
PMAES (Chen and Li, 2023)	0.656	0.553	0.598	0.606	0.626	0.572	0.386	0.530	0.566
ProTACT (Do et al., 2023)	0.647	0.587	0.623	0.632	**0.674**	0.584	0.446	0.541	0.592
PLAES (Chen and Li, 2024)	0.648	0.563	0.604	0.623	0.634	**0.593**	0.403	0.533	0.575
SICTS (Ours)	**0.665**	**0.595**	**0.625**	**0.638**	0.669	0.586	**0.450**	**0.593**	**0.602**

Table 1: Average QWK scores across all traits for each prompt. Bold text indicates the highest value.

Model	**Overall**	**Content**	**Org**	**WC**	**SF**	**Conv**	**PA**	**Lang**	**Nar**	**AVG↑**
BERT	0.499	0.492	0.370	0.473	0.408	0.331	0.591	0.529	0.608	0.478
PAES	0.657	0.539	0.414	0.531	0.536	0.357	0.570	0.531	0.605	0.527
CTS	0.670	0.555	0.458	0.557	0.545	0.412	0.565	0.536	0.608	0.545
PMAES	0.671	0.567	0.481	0.584	0.582	0.421	0.584	0.545	0.614	0.561
ProTACT	0.674	0.596	0.518	0.599	0.585	0.450	**0.619**	**0.596**	**0.639**	0.586
PLAES	0.673	0.574	0.491	0.579	0.580	0.447	0.601	0.554	0.631	0.570
SICTS (Ours)	**0.677**	**0.606**	**0.548**	**0.612**	**0.604**	**0.485**	0.612	0.587	**0.639**	**0.597**

Table 2: Average QWK for each *trait* over all prompts.

(Chen and Li, 2023): This method leverages contrastive learning to facilitate feature transfer between source and target topics. **ProTACT** (Do et al., 2023): This model introduces an essay representation framework that integrates topic information and improves the model's understanding of target-topic essays by incorporating topic coherence features. **PLAES** (Chen and Li, 2024): This model uses meta-learning and a writing proficiency learning strategy to enhance the model's ability.

5.4 Main Results

Following previous works (Do et al., 2023; Chen and Li, 2024), we also present the comparison results of our method with the baseline method from two dimensions. Firstly, we report the scoring performance of the models on each prompt, with the specific results shown in Table 1. From the table, we can see that our method achieves the best results on most prompts and obtains the best average result. Secondly, we show the scoring performance of the models on each attribute in Table 2. It can be seen that in terms of scoring performance for each attribute, our method also achieves the best average performance. Especially for the *Org* and *Conv* attributes, our method directly improves the prediction accuracy by about 3%. We conducted a statistical experiment employing a paired t-test in comparison to ProTACT and PLAES. For the

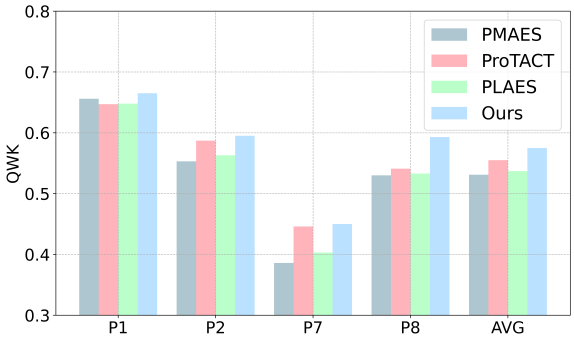


Figure 3: Comparison of scoring performance on low-resource prompts 1, 2, 7 and 8.

significance calculation of the average QWK for each trait across all prompts, the obtained p-value compared to PLAES is 0.0005, and compared to ProTACT, the p-value is 0.041. These results indicate that our method is statistically significant in comparing with the latest method. Besides, relying heavily on semantic representations (i.e., the outputs of BERT) as inputs to trait scorers in cross-prompt settings tends to cause underfitting, thereby impairing the model's generalization ability.

To further demonstrate the performance of our method, we further investigated the low-resource prompt scenario, where the training data provides a small amount of essays consistent with the target prompt. As shown in Table 6, when prompts 1, 2,

Method	P1	P2	P3	P4	P5	P6	P7	P8	AVG↑
MSE + TSL (Baseline)	0.659	0.582	0.602	0.600	0.665	0.582	0.410	0.560	0.583
w/ TCL	0.641	**0.598**	0.602	0.564	0.650	0.575	0.413	0.578	0.578
w/ SIL	0.642	0.579	0.608	0.626	0.667	**0.587**	0.445	**0.601**	0.595
w/ SIL & TCL (SICTS)	**0.665**	0.595	**0.625**	**0.638**	**0.669**	0.586	**0.450**	0.593	**0.602**

Table 3: Results of ablation studies. Average QWK scores across all traits for each prompt. **TCL** indicates the trait representation consistency loss. **SIL** indicates scoring-invariant learning strategy.

Method	**Overall**	**Content**	**Org**	**WC**	**SF**	**Conv**	**PA**	**Lang**	**Nar**	**AVG↑**
MSE + TSL (Baseline)	0.647	0.586	0.534	0.611	0.600	0.468	0.593	0.577	0.628	0.583
w/ TCL	0.668	0.577	0.520	0.613	0.597	0.461	0.584	0.554	0.581	0.573
w/ SIL	**0.679**	0.598	0.540	0.604	0.603	0.468	0.605	**0.587**	0.634	0.590
w/ SIL & TCL (SICTS)	0.677	**0.606**	**0.548**	**0.612**	**0.604**	**0.485**	**0.612**	**0.587**	**0.639**	**0.597**

Table 4: Results of ablation studies. Average QWK for each *trait* over all prompts.

Method	Model Params	Scoring Performance↑
ProTACT	2,764,493	0.592
SICTS	1,859,084	0.602
Δ	-32.75 %	1.68%

Table 5: Comparison results with the SOTA method. Δ represents the magnitude of improvement.

7 and 8 are respectively used as the target prompts, only one type of prompt essay is consistent with the target prompt in the source prompt data. The final comparison results are shown in Figure 3. The results clearly show that our method demonstrates a significant improvement in scoring performance on low-resource prompts, especially on prompt 8, where our method improves by 5.2%. It is worth noting that prompts 1, 2, 7 and 8 all involve long essays that require strong encoding abilities (Table 6), but our method shows improvement on all four prompts, indicating the effectiveness of our model in learning essay scoring representations.

To further demonstrate the effectiveness of our proposed method, we compared it with current state-of-the-art (SOTA) method (Do et al., 2023) from the perspective of scoring efficiency. The comparison results are shown in Table 5. We evaluated both the model parameter and the scoring performance. As can be seen from the table, compared to the current SOTA method (ProTACT), our proposed method achieves better scoring performance while reducing the model parameter count by nearly 33%.

5.5 Ablation Study

To further investigate the effectiveness of our proposed learning strategies, we conducted ablation studies. The ablation results are presented from two dimensions as before. As shown in Table 3, we use MSE and TSL as baseline methods, and then gradually incorporate our proposed training strategies. The combined performance of the proposed learning strategies (w/ SIL & TCL) proposed in this paper outperforms the use of either the SIL strategy or the TCL strategy alone. This indicates that the integration of these two strategies can further enhance the generalization ability of the cross-prompt scoring model. Notably, our experimental results (w/ TCL) reveal that in the cross-prompt trait scoring task, solely focusing on the correlations between traits may reduce the model's generalization ability. However, further enhancing the model's generalization ability (w/ SIL) proves to be significantly beneficial for the cross-prompt trait scoring task. This experiment demonstrates that enhancing the model's general scoring ability is beneficial for improving its cross-prompt trait scoring performance, validating the rationale behind our designed optimization objective. More ablation experimental results can be found at Appendix C.

5.6 Visualization for Generalization Consistency

To verify that our approach can learn more generalizable essay representations, we further conducted a visual analysis on the prompt representations learned by the model using t-SNE (Van der

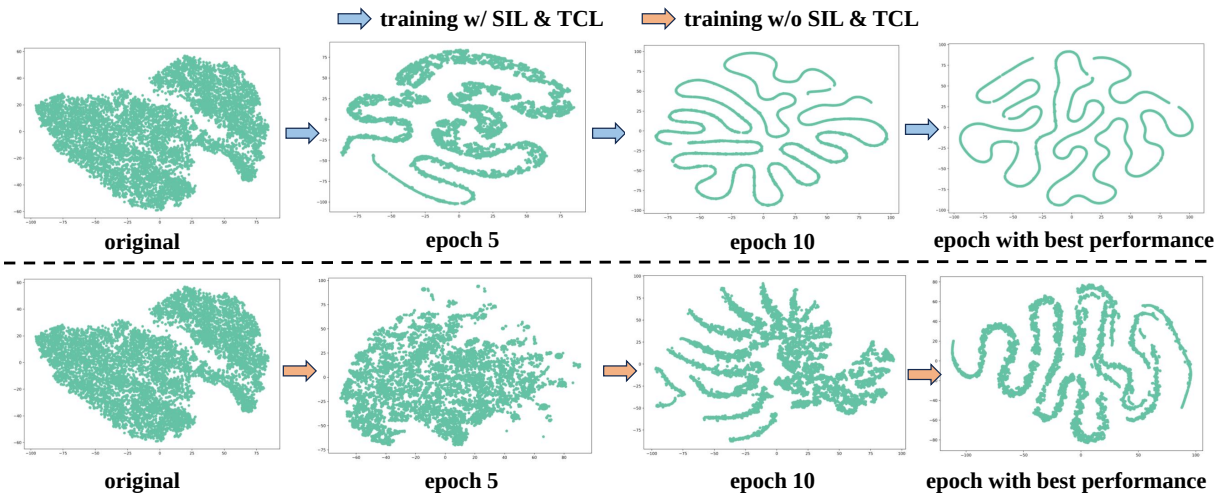


Figure 4: Visualization of changes in source prompt representations during training with our method w/o SIL & TCL (bottom) and w/ SIL & TCL (top), when P4 is the target prompt. The figures represent the visualization of source prompt essay representations at epoch 0 (original), 5, 10 and epoch with best performance, respectively.

Maaten and Hinton, 2008). Specifically, we visualized the representations of all essays on source prompts in the training data, which are generated by shared representations under random initialization (original), training with our method w/o SIL & TCL and w/ SIL & TCL, respectively.

Figure 4 further illustrates how learned source prompt representations change during training progresses, we visualize the essay representations generated at epoch 0 (original), 5, 10 and epoch with best performance during training with and without SIL & TCL, using P4 as the target prompt. The top row presents the results with SIL & TCL, while the bottom row shows the results without SIL & TC. The visualizations show that the representations produced by the two models are quite different at the start of training. As the training epochs increase, training with the proposed learning strategy can obtain more consistent prompt representations. The visualization analysis highlights that our method can significantly make source prompt representations more consistent to improve scoring performance.

6 Conclusion

In this paper, we focus on the cross-prompt essay trait scoring task. To enhance the model's ability to obtain more generalized essay representations, we adopt a scoring-invariance perspective and design a scoring-invariance optimization objective, which promotes the model's general scoring capability. To further improve the cross-prompt trait scoring ability, we propose a novel cross-prompt scoring model based on a sentence-level routing gate mechanism to obtain trait-sharable scoring features. Experimental results demonstrate the significant effectiveness of the proposed method in cross-prompt multi-trait scoring tasks.

Limitations

The limitations of our work can be summarized as follows. While the model demonstrated enhancements on the particular datasets ASAP and ASASP++, it has not been tested on other datasets and other languages. For cross-prompt essay scoring, learning generalizable essay representations helps enhance the model's ability to score across different prompts. Therefore, further exploration of methods that enhance the generalizability of essay representations could further improve the model's cross-prompt scoring capabilities.

Acknowledgements

This work was supported by National Key Research and Development Program of China (2022YFC3303600). We also thank the reviewers for their insightful comments.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

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

J. Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, 70 4:213–20.

Yann Dauphin, Harm de Vries, and Yoshua Bengio. 2015. Equilibrated adaptive learning rates for non-convex optimization. In *NIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

haoxin liu, Harshavardhan Kamarthi, Lingkai Kong, Zhiyuan Zhao, Chao Zhang, and B. Aditya Prakash. 2024. Time-series forecasting for out-of-distribution generalization using invariant learning. In *Forty-first International Conference on Machine Learning*.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097, Melbourne, Australia. Association for Computational Linguistics.

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

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

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.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1114.

Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie. 2022. Domain-invariant feature exploration for domain generalization. *Transactions on Machine Learning Research*.

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).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *ArXiv*, abs/2008.01441.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of AAAI Conference on Artificial Intelligence*.

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

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

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. 2022a. Generalizing to unseen domains: A survey on domain generalization. *Preprint*, arXiv:2103.03097.

Jiong Wang and Jie Liu. 2025. T-MES: Trait-aware mix-of-experts representation learning for multi-trait essay scoring. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1224–1236, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiong Wang, Qing Zhang, Jie Liu, Xiaoyi Wang, Mingying Xu, Liguang Yang, and Jianshe Zhou. 2025. Making meta-learning solve cross-prompt automatic essay scoring. *Expert Systems with Applications*, 272:126710.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022b. 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.

Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 791–797, Brussels, Belgium. Association for Computational Linguistics.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 293–305, New York, NY, USA. Association for Computing Machinery.

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.

Zijian Zhang, Shuchang Liu, Jiaao Yu, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Ziru Liu, Qidong Liu, Hongwei Zhao, Lantao Hu, et al. 2024. M3oe: Multi-domain multi-task mixture-of experts recommendation framework. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 893–902.

Prompt ID	No.of Essays	Essay Type	Attributes	Overall Score Range	Attribute Score Range
P1	1,783	Argumentative	Cont, Org, WC, SF, Conv	2-12	1-6
P2	1,800	Argumentative	Cont, Org, WC, SF, Conv	1-6	1-6
P3	1,726	Source-Dependent	Cont, PA, Lan, Nar	0-3	0-3
P4	1,772	Source-Dependent	Cont, PA, Lan, Nar	0-3	0-3
P5	1,805	Source-Dependent	Cont, PA, Lan, Nar	0-4	0-4
P6	1,800	Source-Dependent	Cont, PA, Lan, Nar	0-4	0-4
P7	1,569	Narrative	Cont, Org, Conv	0-30	0-6
P8	723	Narrative	Cont, Org, WC, SF, Conv	0-60	2-12

Table 6: Composition of the ASAP/ASAP++ combined dataset. The prompt is an instruction that defines the writing theme. Over: Overall, WC: Word Choice, Org: Organization, SF: Sentence Fluency, Conv: Conventions, PA: Prompt Adherence, Nar: Narrativity, Lang: Language.

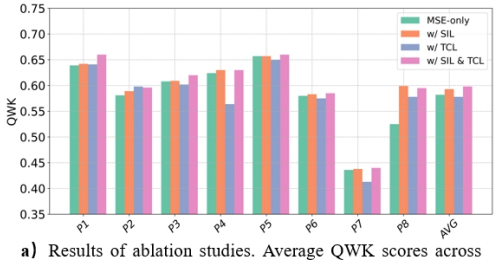
A Statistics on ASAP an ASAP++

In this section, we present additional statistics on the ASAP and ASAP++ dataset. Table 6 provides the detail for various relevant traits across Prompts 1–8. The total number of essays in ASAP as well as the scoring range are also shown in this table.

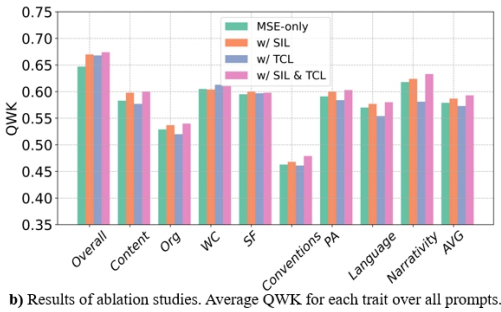
B Implementation Details

To demonstrate the effectiveness of our proposed approach, we apply the same data pre-processing steps as in (Ridley et al., 2021; Do et al., 2023; Chen and Li, 2024). We remove traits that appear in only one prompt, as this is necessary to ensure that no training samples (i.e., essays from non-target prompts) lack trait scores when the prompt is a target prompt. Specifically, we remove the trait *style* from Prompt 7 and *voice* from Prompt 8.

We implement all methods with Tensorflow and Python 3.8.18. We run the model five times with different seeds, the average scores represent the final scores. The handcrafted features are from Ridley et al. 2021, including features of Length-based, Readability, Text Complexity, Text Variation, Sentiment and Topic-coherence (Do et al., 2023). The RMSprop algorithm (Dauphin et al., 2015) is used for optimization. The learning rate is $1e-3$. Our model is trained on 1 NVIDIA GeForce RTX4090 GPU. Running the model five times with different seeds, $\{12, 22, 32, 42, 52\}$. For the final loss, the parameters λ is set to 0.7. And α and β are tune $\in \{0.5, 1.0\}$ accoring to develop set performance. For TCL loss, δ is set to 0.7 and τ is 0.1. For TSL loss, we set the same setting with Do et al. 2023. The average scores represent the final scores. We perform prompt-wise eight-fold cross-validation, where essays from the target prompt are used as test data, and essays from non-target prompts are used as training data.



a) Results of ablation studies. Average QWK scores across all traits for each prompt.



b) Results of ablation studies. Average QWK for each trait over all prompts.

Figure 5: Results of detailed ablation studies.

C More Experimental Results

C.1 Compared with Large Language Model

Recently, the application of large language models to automated essay scoring tasks has attracted increasing research attention (Xiao et al., 2025; Lee et al., 2024). To further demonstrate the effectiveness of our method, we conducted a comparison with ChatGPT-3.5-turbo under a zero-shot setting as Li and Pan 2025. The specific experimental results are shown in Table 7 and Table 8. It can be observed that the general large language model does not exhibit a significant advantage, indicating that for the task of cross-prompt essay scoring, domain-specific small models still hold considerable value in automated essay scoring research. We follow and adapt the prompt templates proposed by Lee et al. 2024 to guide the LLM in automated essay scoring. An example prompt template for

Model	P1	P2	P3	P4	P5	P6	P7	P8	AVG↑
GPT-3.5 (zero-shot)	0.231	0.501	0.382	0.419	0.532	0.491	0.153	0.297	0.376
SICTS (Ours)	**0.665**	**0.595**	**0.625**	**0.638**	**0.669**	**0.586**	**0.450**	**0.593**	**0.602**

Table 7: Average QWK scores across all traits for each prompt when compared with LLM.

Model	**Overall**	**Content**	**Org**	**WC**	**SF**	**Conv**	**PA**	**Lang**	**Nar**	**AVG↑**
GPT-3.5-turbo(0-shot)	0.390	0.423	0.281	0.335	0.301	0.285	0.430	0.471	0.410	0.369
SICTS (Ours)	**0.677**	**0.606**	**0.548**	**0.612**	**0.604**	**0.485**	**0.612**	**0.587**	**0.639**	**0.597**

Table 8: Average QWK for each *trait* over all prompts.

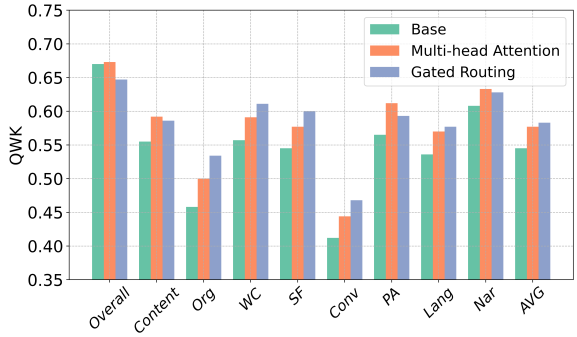


Figure 6: Performance of different shared layer essay feature extraction methods.

scoring an essay from prompt 1 is provided below:

You are a member of the English essay writing test evaluation committee. Your task is to perform automated essay scoring on five specific traits. Please carefully read the following essay and assign a score for each trait: Content, Word Choice, Conventions, Organization, Sentence Fluency, and Overall quality. Each trait should be scored on a scale from 1 to 6, and the Overall score ranges from 1 to 12. After assigning a score for each trait, provide a explanation to justify your score.

Essay Prompt: Prompt

Essay: Essay

Please use the following format for each trait: [Trait] Score: ..., Explain: ...

C.2 More Ablation Studies

In our main paper, we have conducted the ablation analysis in Section 5.5 to illustrate the effect of the proposed learning strategies. The experimental results show that adding the proposed learning strategies yield the greatest performance improvement. In the ablation experiments presented in the main text, we only report the results of ablating the SIL and TCL losses. To further demonstrate the advantages of the proposed method, we removed the Trait-Similarity Loss from the final loss function and used MSE, SIL, and TCL losses as optimization objectives to train the model. Experimental results can be found in Figure 5. The experimental results also indicate our proposed methods can yield synergies when jointly applied.

C.3 Effect of Sentence-Level Gated Routing

To further validate the effectiveness of the proposed Gated Routing method for extracting shared features at the sentence-level in essays, we designed an incremental analysis experiment. Starting from encoding essay representation with a basic shared feature extraction method as Ridley et al. 2021, we gradually incorporated the multi-head self-attention mechanism (Do et al., 2023) and the routing gate control mechanism. The experimental results are shown in Figure 6. It can be observed that, compared to the other two shared layer feature extraction methods, our proposed routing gating mechanism demonstrates a certain advantage in most trait score prediction capabilities and overall average performance. Although the routing mechanism shows a slight disadvantage in predicting the overall essay score compared to the other two methods, its performance in predicting overall score significantly improves when combined with our proposed scoring strategies (as shown in Table 4). This enhancement further demonstrates the effectiveness of our approach in boosting the model's prompt generalization capabilities.

C.4 Effect of Hyper-parameters

For the hyper-parameter search, we use grid search to search for the best values and select the value that performs the best on the validation set. We experimented with the different values of the hyper-parameter δ of TCL. The results of Table 9 show

Setting	$\delta = 0.6$	$\delta = 0.7$	$\delta = 0.8$	$\delta = 0.9$
QWK	0.599	0.602	0.602	0.601

Table 9: Effect of the thresholds of TCL.

Setting	$\alpha = 0.5$ $\beta = 0.5$	$\alpha = 0.5$ $\beta = 1$	$\alpha = 1$ $\beta = 0.5$	$\alpha = 1$ $\beta = 1$
QWK	0.591	0.583	0.593	0.589

Table 10: Effect of the parameters setting of α and β on Prompt 8.

that different δ values greater than 0.6 and condition change have little influence.

Furthermore, we investigated the impact of different parameter configurations on the model's generalization performance. As shown in the Table 10, we report the evaluation results on P8 under various parameter settings. The experimental results indicate that, in the overall optimization objective, an excessive focus on trait-level consistency may impair the model's generalization ability in scoring traits, which is consistent with the conclusions drawn from our ablation studies. In addition, improving the model's generalization ability contributes to learning more generalized representations. This suggests that, for cross-prompt trait scoring tasks, emphasizing the learning of universal representations can effectively enhance model generalization.