Prompt- and Trait Relation-aware Cross-prompt Essay Trait Scoring

Heejin Do[⋆], **Yunsu Kim**[⋆†], **Gary Geunbae Lee**[⋆†]
[⋆]Graduate School of AI, POSTECH
[†]Department of Computer Science and Engineering, POSTECH
{heejindo, yunsu.kim, gblee}@postech.ac.kr

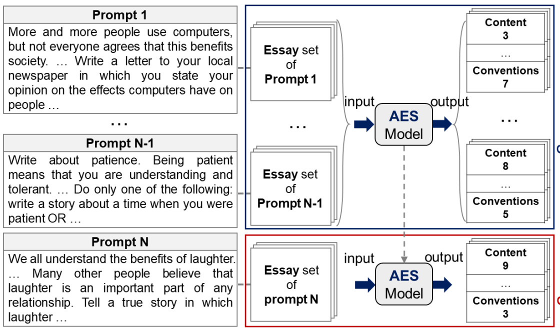
Abstract

Automated essay scoring (AES) aims to score essays written for a given prompt, which defines the writing topic. Most existing AES systems assume to grade essays of the same prompt as used in training and assign only a holistic score. However, such settings conflict with real-education situations; pre-graded essays for a particular prompt are lacking, and detailed trait scores of sub-rubrics are required. Thus, predicting various trait scores of unseen-prompt essays (called cross-prompt essay trait scoring) is a remaining challenge of AES. In this paper, we propose a robust model: prompt- and trait relation-aware cross-prompt essay trait scorer. We encode prompt-aware essay representation by essay-prompt attention and utilizing the topic-coherence feature extracted by the topic-modeling mechanism without access to labeled data; therefore, our model considers the prompt adherence of an essay, even in a cross-prompt setting. To facilitate multi-trait scoring, we design trait-similarity loss that encapsulates the correlations of traits. Experiments prove the efficacy of our model, showing state-of-the-art results for all prompts and traits. Significant improvements in low-resource-prompt and inferior traits further indicate our model's strength.

1 Introduction

Automated essay scoring (AES) aims to score essays written for a specific prompt, which defines the writing instructions and topic. As a subordinate or alternative to human scorers, it has the advantages of fairness and low costs. Thus far, most AES systems have been built on the assumptions of grading essays on the same prompt used for training and only assigning an overall score, achieving noticeable growth (Taghipour and Ng, 2016; Dong et al., 2017; Yang et al., 2020; Wang et al., 2022).

However, such settings conflict with real-education systems, where pre-labeled essays for a specific prompt are not given, and in-depth feedback requires multiple trait scores. Acknowledging this, recent works have suggested cross-prompt models (Jin et al., 2018; Li et al., 2020; Ridley et al., 2020) that are tested using essays of unseen prompt, like zero-shot learning, and trait-scoring models (Mathias and Bhattacharyya, 2020; Hussein et al., 2020; Kumar et al., 2021; He et al., 2022) that output multiple trait scores. Handling both settings (Figure 1) is a direction for practical AES and yet has rarely been studied (Ridley et al., 2021).



More and more people use computers, but not everyone agrees that this benefits society. … Write a letter to your local newspaper in which you state your opinion on the effects computers have on people …

Write about patience. Being patient means that you are understanding and tolerant. … Do only one of the following: write a story about a time when you were patient OR …

We all understand the benefits of laughter. … Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter …

Figure 1: Cross-prompt essay trait scoring task

For a cross-prompt setting, using non-prompt-specific features that capture the general essay qualities such as length and readability is emphasized (Ridley et al., 2020; Uto, 2021). This is to avoid the model biased toward the prompts of trained essays, but the model fails to reflect any prompt-relevant information (e.g., whether the essay fits the prompt topic), inhibiting accurate scoring. For trait scoring, most methods extend holistic scoring models without particular consideration of trait properties. Both settings leave huge room for improvement.

In this paper, we propose a robust model, prompt- and trait relation-aware cross-prompt trait scorer (ProTACT), with the ultimate goal of improving AES for practical use. Attending to the prompt-relevant aspects and trait similarities leads to overcoming both cross-prompt and multi-trait settings.

To ensure that the model reflects prompt-relevant information, we introduce a novel architecture to obtain prompt-aware essay representation. Rather

Findings of the Association for Computational Linguistics: ACL 2023, pages 1538–1551
July 9-14, 2023 ©2023 Association for Computational Linguistics

than only encoding the essay, we directly encode the prompt instruction and apply attention. This provides hints for scoring in cross-prompt settings since prompt content is always-given information, even for ungraded essays of new prompts. Furthermore, we suggest extracting the topic-coherence feature by applying the topic modeling mechanism latent Dirichlet allocation (LDA) (Blei et al., 2003). This feature notifies essay coherence on a specific topic to the model without accessing labels.

To facilitate multi-trait scoring, we designed a trait-similarity loss that incorporates correlations between different trait scores. Practically, trait scores are not independent of one another; for example, both *Prompt Adherence* and *Content* traits evaluate prompt-relevant aspects of an essay. Finding strong correlations between trait scores, we mirror this for model training. Specifically, we penalize when the similarity of actual trait scores is over a threshold but that of predicted trait scores is low. This enhances the advantages of multi-trait learning by mutually assisting in different tasks.

We evaluate ProTACT with the widely used ASAP and ASAP++ datasets. ProTACT achieves state-of-the-art results, outperforming the baseline system (Ridley et al., 2021) for all QWK scores of traits and prompts. Significant improvements of 6.4% on average and 10.3% for the *Content* trait are observed for a low-resource prompt, which performed poorly due to lacking similar-type training essays. This highlights the strength of ProTACT in the cross-prompt setting, overcoming the absence of pre-graded essays. Remarkably improved assessments for previously inferior traits further prove the effectiveness of multi-trait scoring. Codes and datasets are available on Github[1].

2 Related Work

AES studies mostly focus on the **prompt-specific holistic scoring** task. Aside from early machine learning-based regression or classification approaches (Landauer, 2003; Attali and Burstein, 2006; Larkey, 1998; Rudner and Liang, 2002), recent deep-learning-based methods for automatically learning essay representation are dominant. Notably, approaches that hierarchically represent essays from word- or sentence- to essay-level show competitive accuracy (Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017). Late attempts to fine-tune pre-trained models to develop more successful AES include Yang et al. (2020), who fine-tune BERT by combining regression and ranking loss, and Wang et al. (2022), who suggest a multi-scale representation for BERT. Zhang and Litman (2019) additionally encode source excerpts of source-dependent essays and suggest a co-attention. Our essay-prompt attention is distinct from theirs, as we encode the prompt rather than the source excerpt and apply attention differently.

Pointing out that previous successes in AES are far from real-world systems, few studies of the **cross-prompt** setting suggest methods of not examining target-prompt essays (Jin et al., 2018; Li et al., 2020; Ridley et al., 2020). Considering the essay's semantic disparity by different prompts, the use of non-prompt-specific features of general essay qualities is highlighted in cross-prompt settings; Ridley et al. (2020) crafted the features of essay qualities, categorized as *length-based*, *readability*, *text complexity*, *text variation*, and *sentiment*. However, they disregard the topic-coherence of the essay, which is an important consideration for grading (Miltsakaki and Kukich, 2004). To consider coherence during rating, we suggest a way of extracting the topic-coherence feature.

To provide several trait scores that fit the subrubrics, a few **trait-scoring** studies have been proposed; however, they simply extend the existing holistic scoring methods by adding multi-output linear layers (Hussein et al., 2020) or using multiple trait-specific models (Mathias and Bhattacharyya, 2020; Kumar et al., 2021). Emphasizing both the cross-prompt and trait scoring task, Ridley et al. (2021) suggest a leading model for the **cross-prompt trait scoring** task. They extend the Dong et al. (2017) model by setting multiple trait-specific layers and concatenating the features of Ridley et al. (2021). Despite achieving the best results on the task, the performance still lags far behind the prompt-specific holistic scoring. In addition, the performance gaps between traits and between target prompts are remarkable. We propose a novel architecture to improve cross-prompt trait scoring and thereby reduce the performance gap.

3 Model Description: ProTACT

To benefit from both automatically learning essay representations and precisely designed essay features, we combine both approaches. Therefore, ProTACT comprises two main parts: obtaining the prompt-aware essay representation and extracting

[1] https://github.com/doheejin/ProTACT

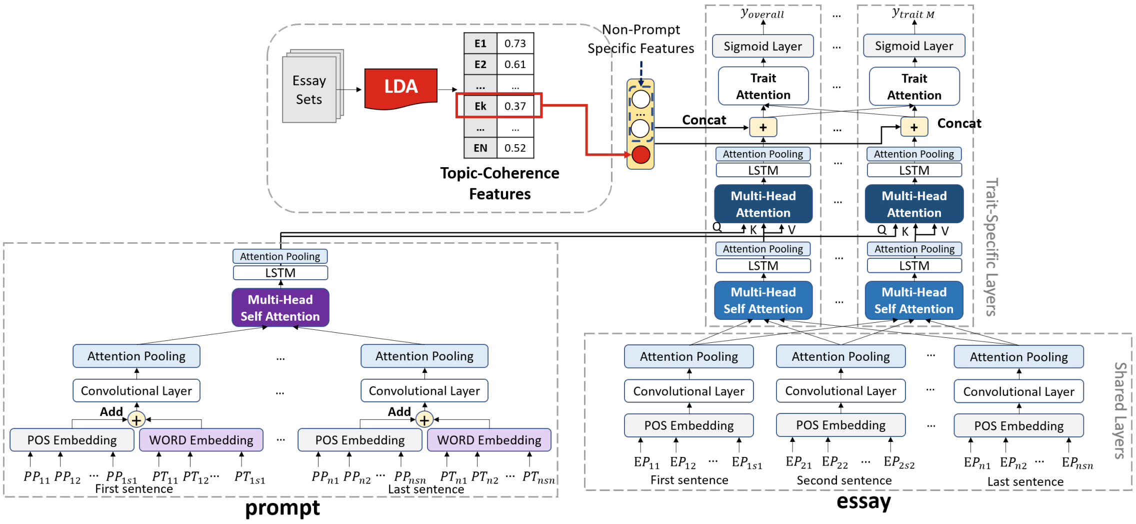


Figure 2: ProTACT model architecture.

the essay features (Figure 2). The learned prompt-aware essay representation is concatenated with the pre-extracted essay features, constructing the final essay representation to score. The model is trained with the loss function that interpolates our trait-similarity loss and the mean squared error (MSE).

3.1 Prompt-aware Essay Representation

We apply the hierarchical structure to encode the essay, first obtaining sentence-level representations and then a document-level representation. Hierarchically learning the document representation has proven effective for AES models, as it mirrors the essay structure that comprises sentences (Dong et al., 2017; Ridley et al., 2020, 2021).

To score multiple traits, we set separate trait-specific layers on top of the shared layers as the baseline model (Ridley et al., 2021), but with different layer constructions. Shared layers and trait-specific layers are for sentence- and essay-level representations, respectively. To obtain *M* essay representations for *M* traits (including the overall score), *M* trait-specific modules exist. Sharing low-level layers enables information interchange between different traits, alleviating the data shortage caused by partial trait coverage.

To obtain prompt-aware essay representation for each trait, we introduce essay-prompt attention. Unlike existing methods that only encode the essay, we encode the prompt information in parallel and apply attention to the essay representation.

Essay Representation Instead of directly using word embedding, we use part-of-speech (POS) embedding for generalized representation, since doing so prevents overfitting to training data in cross-prompt settings (Jin et al., 2018; Ridley et al., 2020, 2021). Each sentence is POS-tagged with the Python NLTK[2] package, and the tagged words of each sentence are mapped to dense vectors. Then, to obtain **sentence-level representation**, the 1D convolutional layer followed by attention pooling (Dong et al., 2017) is applied for each sentence. The following equations explain the convolutional (Eq. 1) and attention-pooling layers (Eqs. 2, 3, 4):

$$\mathbf{c}_i = f(\mathbf{W}_c \cdot [\mathbf{x}_i : \mathbf{x}_{i+h_w-1}] + \mathbf{b}_c) \qquad (1)$$

$$\mathbf{a}_i = \tanh(\mathbf{W}_a \cdot \mathbf{c}_i + \mathbf{b}_a) \qquad (2)$$

$$u_i = \frac{\exp(\mathbf{w}_u \cdot \mathbf{a}_i)}{\sum \exp(\mathbf{w}_u \cdot \mathbf{a}_j)} \qquad (3)$$

$$\mathbf{s} = \sum u_i \mathbf{c}_i \qquad (4)$$

where $\mathbf{c}_i$ is the feature representation after the convolutional layer, $\mathbf{W}_c$ is the weight matrix, $\mathbf{b}_c$ is the bias vector, and $\mathbf{h}_w$ is the window size of the convolutional layer. The final sentence representation s is obtained by the weighted sum where $\mathbf{u}_i$ is the attention weight, $\mathbf{a}_i$ is the attention vector, and $\mathbf{w}_u$ is the weight vector. $\mathbf{W}_a$ and $\mathbf{b}_a$ are the attention matrix and bias vector, respectively.

To examine each point of the long-range essays effectively, we first apply the multi-head self-attention (Vaswani et al., 2017) mechanism for the **essay-level representation**. Each trait-specific module takes the generated sentence-level representations as input and applies the multi-head self-attention. Consider the j-th trait score prediction

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

task; the output of the previous layer, S, which is the matrix of sentence representations set as a query, key, and value:

$$\mathrm{H}_{\mathrm{i}}^{\mathrm{j}} = \mathrm{Att}(SW_i^{j1}, SW_i^{j2}, SW_i^{j3}) \quad (5)$$

$$\mathrm{MH}(S)^j = \mathrm{Concat}(\mathrm{H}_1^{\mathrm{j}}, ..., \mathrm{H}_{\mathrm{h}}^{\mathrm{j}})W^{jO} \quad (6)$$

where Att and $\mathrm{H_i}$ denote scaled-dot product attention and the i-th head, respectively, and W_i^{j1}, W_i^{j2}, and W_i^{j3} are the parameter matrices. To the best of our knowledge, we are the first to apply the multi-head self-attention mechanism in both cross-prompt and trait-scoring settings. We hypothesize that this better models the structural aspect of the essay with the use of POS embedding and easily captures the relationship between different points of the essay from various perspectives.

Next, the recurrent layer of LSTM (Hochreiter and Schmidhuber, 1997) is applied to the output:

$$\mathbf{h}_t^j \quad = \quad \mathrm{LSTM}(m_{t-1}^j, m_t^j) \quad (7)$$

where j is the j-th trait score prediction task, m^j is the concatenated output of the previous layer, and $\mathbf{h}_t^j$ denotes the hidden representation for the j-th task at time-step t. As LSTM captures sequential connections, directly applying it to a relation-encoded representation can lead to the sequential interpretation of relations (Li et al., 2018). This is followed by the attention pooling layer (Eqs. 2, 3, 4).

Prompt Representation In practical education situations, grades are scored based on prompt instructions. Inspired by this, we encode the prompt instruction corresponding to each essay and make the model attend to it. Prompt representation is also obtained in the same order as the essay representation: embedding layer, convolutional layer with attention pooling, multi-head self-attention with LSTM, and attention pooling layer. However, to contain the contents of the prompt, we add the POS embedding with the pre-trained GloVe (Pennington et al., 2014) word embedding.

Essay-Prompt Attention For the next step, essay-prompt attention is performed using a multi-head self-attention mechanism. The queries are set as the obtained prompt representation and the keys and values are set as the obtained essay representation. This allows every position of the essay to view sub-parts of the prompt; hence, essay-prompt attention captures the relationship between the essay and the prompt. Finally, the LSTM with attention pooling layer is applied to obtain the prompt-aware essay representation, $\mathbf{pa}^j$, for each j-th task.

Essay ID	Topic Distribution [(Topic, Prob)]	**TC**
1	[(0, **0.8337**), (5, 0.16295)]	0.8337
2	[(0, **0.7541**), (1, 0.0472), (5, 0.1472), ...]	0.7541
...	...	...
11194	[(2, 0.0477), (5, **0.8701**), (6, 0.0727)]	0.8701
11195	[(0, 0.0705), (2, 0.0664), (5, **0.8405**), ...]	0.8405

Table 1: Example of the extracted features by LDA for each essay (**TC** denotes the Topic-coherence feature).

Final Prediction The essay representation is subsequently concatenated with pre-engineered features. As in the baseline model, we also use the non-prompt-specific features of PAES (Ridley et al., 2020) that are exquisitely engineered to represent general essay quality in various aspects. However, we additionally concatenate our own feature of topic coherence. The feature vectors, $\mathbf{f}$, are then concatenated with each trait prediction, $\mathbf{pa}^j$: $\mathbf{con}^j = [\mathbf{pa}^j; \mathbf{f}]$.

Then, the trait-attention defined in Ridley et al. (2021) is performed to attend to the representations of the other traits where $j = 1, 2, \ldots, M$:

$$\mathbf{A} \quad = \quad [\mathbf{con}^1, \ldots, \mathbf{con}^M] \quad (8)$$

$$v_i^j \quad = \quad \frac{\exp(\mathrm{score}(\mathbf{con}^j, \mathbf{A}_{-j,i}))}{\sum_l \exp(\mathrm{score}(\mathbf{con}^j, \mathbf{A}_{-j,l}))} \quad (9)$$

$$\mathbf{t}^j \quad = \quad \sum v_i^j \mathbf{A}_{-j,i} \quad (10)$$

$$\mathbf{final}^j \quad = \quad [\mathbf{con}^j; \mathbf{t}^j] \quad (11)$$

where $\mathbf{A}$ is a concatenation of the representations for each trait prediction; $\mathbf{A}_{-j}$ indicates the masking of the target trait's representation; v_i^j is the attention weight for the i-th trait; $\mathbf{t}^j$ is the attention vector. The final representation, $\mathbf{final}^j$, for each trait prediction is obtained by concatenating $\mathbf{con}^j$ and $\mathbf{t}^j$. Lastly, the final trait score, $\hat{y}^j$ is obtained by applying a linear layer with the sigmoid function $\hat{y}^j = \mathrm{sigmoid}(\mathbf{w}_y^j \cdot \mathbf{final}^j + b_y^j)$. Here, $\mathbf{w}_y^j$ is a weights vector and $\mathbf{b}_y^j$ is a bias.

3.2 Topic-Coherence Feature

To complement the existing non-prompt-specific features, in which prompt-related information is entirely excluded, we suggest using the LDA topic modeling mechanism. Looking at the document sets with the number of topics as a hyper-parameter, LDA identifies the topics and topic distributions for each document. Therefore, it can find out how an essay is focused on a particular topic, considering essays as documents. Since only essays are used without labels, features can be extracted even for new prompt essays in cross-prompt situations.

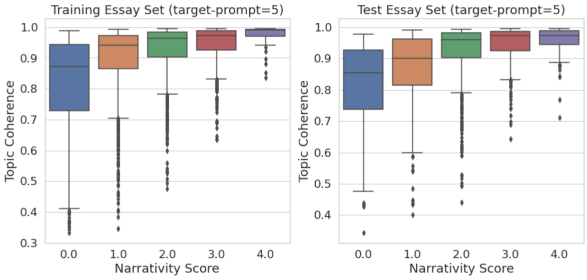


Figure 3: Box plot distributions of topic-coherence features of essays according to their *Narrativity* trait scores.

Specifically, given the essay sets written for N prompts, we apply LDA by setting the number of topics as N to obtain the topic distribution for each essay (Table 1); having multiple topics with low probabilities indicates lacked focus on a single topic, while the presence of a high-probability topic implies high focus on a certain topic. Then, the highest topic rate among the topic distributions for each essay is extracted as the topic-coherence feature. LDA is conducted separately for each training set since target-prompt essays should not be seen in training, eg., the training set of target-prompt 1 only includes prompts 2–8 essays. For testing, target-prompt essays are also used for extraction.

Does the Feature Imply Topic-coherence? To examine whether our feature connotes the topic-coherence, we investigate the extracted feature's distribution with labeled *Narrativity* trait score on our dataset, which is the attribute for evaluating the essay's coherence to the prompt, with integers 0–4. We plot the case when the target-prompt is 5, as it has the most essays among the prompts for which *Narrativity* trait is evaluated. Figure 3 shows box plot[3] distributions of extracted features according to essays' *Narrativity* trait scores. The plotted training set only includes prompts 3,4, and 6 essays because only prompts 3–6 have labeled *Narrativity* score. The greater distribution of high topic coherence at higher *Narrativity* scores indicates that our feature reflects the essay's actual topic coherence. It is noteworthy that the test set shows similar trends in that our feature can give direct hints about consistency when scoring unseen prompt essays.

Does the Topic Correspond to Each Prompt? We further investigate the probability of the same prompt's essays having the same highest topic, in each training set (Table 2). Each `Set` denotes the training essay set of the target-prompt n, where LDA is separately applied. The left index of `Pr1-7` denotes different prompts by the `Set` since each target prompt is excluded. For example, index `Pr1` of `Set1` denotes the probability of prompt 2 essays having the same highest topic. Overall high probabilities imply that topics extracted by LDA are strongly related to the actual prompts, further notifying that our feature allows the model to recognize prompt relevance even in the cross-prompt setting.

	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8
Pr1	0.996	1.000	0.685	0.680	1.000	0.979	0.885	0.996
Pr2	0.985	0.984	0.994	0.994	0.993	0.996	0.994	0.996
Pr3	0.976	0.987	0.980	0.981	0.982	0.978	0.987	0.981
Pr4	0.995	0.994	0.996	0.996	0.985	0.985	0.981	0.986
Pr5	0.999	0.998	0.998	0.999	0.998	0.994	0.997	0.996
Pr6	0.841	0.975	0.994	0.995	0.992	0.964	0.998	0.998
Pr7	0.936	0.535	0.996	0.994	0.994	0.965	0.996	0.978
Avg	0.961	0.925	0.949	0.948	0.992	0.980	0.977	0.990

Table 2: The probabilities of the essays of the same prompt have the same highest topic in each training set.

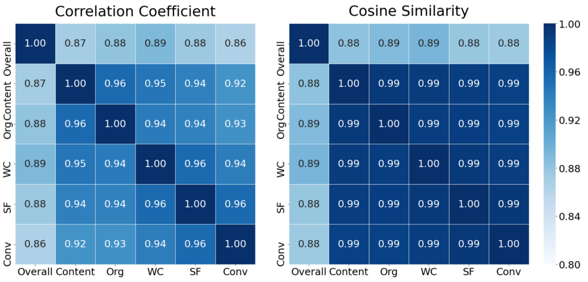


Figure 4: Correlation coefficients and cosine similarities between the ground-truth trait scores of prompt types 1, 2, and 8, which have the same trait composition.

3.3 Trait-Similarity Loss

As in most AES systems, the existing cross-prompt trait scoring system is trained with the MSE loss. However, the only use of MSE loss disregards the correlations between different trait scores (Figure 4). We integrate trait-relationship into the loss function, called the Trait-Similarity (TS) loss. In detail, when the similarity of the ground-truth trait score vectors is beyond the threshold, the model learning proceeds in the direction to increase the similarity of the predicted trait score vectors. The TS loss ($\mathrm{L_{ts}}$) is defined as follows:

$$\mathrm{L_{ts}}(y, \hat{y}) = \frac{1}{c} \sum_{j=2}^{M} \sum_{k=j+1}^{M} \mathrm{TS}(\hat{\mathbf{y}}_j, \hat{\mathbf{y}}_k, \mathbf{y}_j, \mathbf{y}_k) \quad (12)$$

$$\mathrm{TS} = \begin{cases} 1 - \cos(\hat{\mathbf{y}}_j, \hat{\mathbf{y}}_k) & \text{, if } \mathrm{r}(\mathbf{y}_j, \mathbf{y}_k) \geq \delta \\ 0 & \text{, otherwise} \end{cases} \quad (13)$$

where cos and r denote the cosine similarity and the Pearson correlation coefficient (PCC),

[3] seaborn (`https://seaborn.pydata.org/`) is used.

Prompt	Essay Type	Num of Essays	Avg Length	Traits
1	Argumentative	1785	350	Content, Word Choice, Organization, Sentence Fluency, Conventions
2	Argumentative	1800	350	Content, Word Choice, Organization, Sentence Fluency, Conventions
3	Source-Dependent	1726	150	Content, Prompt Adherence, Narrativity, Language
4	Source-Dependent	1772	150	Content, Prompt Adherence, Narrativity, Language
5	Source-Dependent	1805	150	Content, Prompt Adherence, Narrativity, Language
6	Source-Dependent	1800	150	Content, Prompt Adherence, Narrativity, Language
7	Narrative	1569	300	Content, Organization, Conventions
8	Narrative	723	650	Content, Word Choice, Organization, Sentence Fluency, Conventions

Table 3: Summarization of the ASAP and ASAP++ combined dataset (Mathias and Bhattacharyya, 2018).

respectively; δ is the threshold and c is the number of calculated TS that is not 0; $\mathbf{y}_j = [y_{1j}, y_{1j}, \cdots, y_{Nj}]$ is j-th ground-truth trait vector and $\hat{\mathbf{y}}_j = [\hat{y}_{1j}, \hat{y}_{2j}, \cdots, \hat{y}_{Nj}]$ is predicted trait vector. Note that *Overall* trait ($j = 1$) is excluded, as its score has relatively low correlations than other traits. The total loss, $\mathrm{L}_{\mathrm{total}}$, is calculated as the interpolation of $\mathrm{L}_{\mathrm{mse}}$ and L_{ts}:

$$\mathrm{L}_{\mathrm{total}}(y, \hat{y}) = \lambda \cdot \mathrm{L}_{\mathrm{mse}}(y, \hat{y}) + (1 - \lambda) \cdot \mathrm{L}_{\mathrm{ts}}(y, \hat{y})$$

where the MSE loss is defined as, $\mathrm{L}_{\mathrm{mse}}(y, \hat{y}) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{y}_{ij} - y_{ij})^2$, when predicting M trait scores for N essays and given the ground truth y and prediction $\hat{y}$. Note that TS loss of reflecting similarity between the traits is distinct from Wang et al. (2022)'s work of reflecting the similarity between the actual score and predicted score in a loss function for prompt-specific holistic scoring.

Given the entire trait set, Y, the specific trait set for each i-th training sample Y^i differs depending on its prompt. Thus, for accurate calculation, masking to handle traits without gold scores is applied as $\mathbf{y}_i = \mathbf{y}_i \otimes mask_i$ and $\hat{\mathbf{y}}_i = \hat{\mathbf{y}}_i \otimes mask_i$. On the i-th essay, $mask_{ij}$ is computed for the j-th trait with the following function (Ridley et al., 2021):

$$mask_{ij} = \begin{cases} 1, & \text{if } Y_j \in Y^i \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

4 Experiment

We experimented with the same dataset[4] as the baseline system, which is comprised of the publicly available Automated Student Assessment Prize (ASAP[5]) and ASAP++[6] datasets (Mathias and Bhattacharyya, 2018). The original ASAP dataset contains eight prompts and their corresponding English-written essay sets, without personal information. Essays are assigned human-graded scores for their overall quality, and only essays of prompts 7 and 8 are assigned additional scores for several traits of scoring rubrics. Thus, the ASAP++ dataset, which has the same essay sets as ASAP but additionally graded trait scores for Prompts 1–6, is also utilized. Therefore, trait scores for prompts 1–6 are from the ASAP++, whereas trait scores for prompts 7 and 8 and overall scores for all prompts are from the ASAP dataset (Table 3). For comparison, we exclude the *Style* and *Voice* attributes, which only appear in one prompt, as in the baseline model.

Validation and Evaluation For the overall training procedure, we applied the prompt-wise cross-validation that is used for the existing cross-prompt AES (Jin et al., 2018; Ridley et al., 2020, 2021). In detail, essays of one prompt are set as test data while essays of other prompts are set as training data, which is repeated for each prompt. The development set of each case comprises essays of the same prompts as the training set. For the evaluation, we used Quadratic Weighted Kappa (QWK), the official metric for ASAP competition and most frequently used for AES tasks, which measures the agreement between the human rater and the system.

Training Details For a fair comparison, we maintained training details of the baseline model, other than those required by ProTACT. Out of the total 50 epochs, the one with the highest average QWK score for all traits in the development set was selected for the test. We set the dropout rate as 0.5, CNN filter and kernel size as 100 and 5, respectively, LSTM units as 100, POS embedding dimension as 50, and batch size as 10. We set two heads and the embedding dimension to 100 for multi-head attention. The total number of parameters is $2.76M$. For TS loss, δ of 0.7, and λ of 0.7 are used. The RMSprop algorithm (Dauphin et al., 2015) is used for optimization. The code is implemented in Tensorflow 2.0.0 and Python 3.7.11, and a Geforce RTX 2080Ti GPU card is used. Running the model five times with different seeds, $\{12, 22, 32, 42, 52\}$,

[4] https://github.com/robert1ridley/cross-prompt-trait-scoring/tree/main/data

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

[6] https://lwsam.github.io/ASAP++/lrec2018.html

	Prompts									
Model	1	2	3	4	5	6	7	8	AVG	SD(↓)
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	-
*CTS-baseline	0.629	0.543	0.596	0.620	0.614	0.546	0.382	0.501	0.554	0.020
ProTACT	**0.647**	**0.587**	**0.623**	**0.632**	**0.674**	**0.584**	**0.446**	**0.541**	**0.592**	0.016

Table 4: Average QWK scores over all traits for each **prompt**; *SD* is the averaged standard deviation for five seeds, and **bold** text indicates the highest value.

	Traits										
Model	Overall	Content	Org	WC	SF	Conv	PA	Lang	Nar	AVG	SD(↓)
PAES (Ridley et al., 2020)	0.657	0.539	0.414	0.531	0.536	0.357	0.570	0.531	0.605	0.527	-
CTS (Ridley et al., 2021)	0.67	0.555	0.458	0.557	0.545	0.412	0.565	0.536	0.608	0.545	-
*CTS-baseline	0.670	0.551	0.459	0.562	0.556	0.413	0.568	0.533	0.610	0.547	0.012
ProTACT	**0.674**	**0.596**	**0.518**	**0.599**	**0.585**	**0.450**	**0.619**	**0.596**	**0.639**	**0.586**	0.009

Table 5: Average QWK scores over all prompts for each **trait** (WC: Word Choice; PA: Prompt Adherence; Nar: Narrativity; Org: Organization; SF: Sentence Fluency; Conv: Conventions; Lang: Language).

the average scores represent the final scores. LDA is applied using the Gensim[7] library, specifying the number of prompts as the number of topics. Considering that each training and test uses an essay set of seven and eight prompts for LDA, the passes are set to 12 and 15, respectively.

5 Results and Discussion

The results clearly show that ProTACT outperforms the baseline CTS model for all prompts (Table 4) and traits (Table 5). In Ridley et al. (2021), PAES of the cross-prompt holistic scoring model is separately used for each trait scoring as a comparison of CTS, which is their proposed model. The *CTS-baseline is our implementation, with which we mainly compared our model for a fair comparison.

For target-prompt predictions (Table 4), ProTACT achieved 3.8% improvements on average. Compared to prompts 1 and 4, which already had high-quality predictions of 0.629 and 0.620, the other six prompts' predictions achieved larger improvement, reducing gaps between different prompts. This indicates that our methods provide more aid when predicting essays of a target prompt vulnerable to cross-prompt settings.

Target	**Model**	Overall	Content	Org	Conv	**Avg**
1,2,8 (avg)	*CTS-baseline	**0.679**	0.523	0.535	0.490	0.557
	ProTACT	0.673	**0.585**	**0.585**	**0.523**	**0.592**
	Δ	-0.6%	6.2%	5.0%	3.3%	3.5%
7	*CTS-baseline	0.720	0.398	0.231	0.179	0.382
	ProTACT	**0.735**	**0.501**	**0.315**	**0.232**	**0.446**
	Δ	1.5%	10.3%	8.4%	5.3%	6.4%

Table 6: Comparison of QWK scores for Prompt 7 and Prompt 1,2,8 (averaged). *Target* means target-prompt.

We further investigated the low-resource prompt, which lacks similar-type essays in its training data (Table 6). When predicting target-prompt 7, only 723 essays are of the same *Narrative* type (prompt 8) in the training set (Table 3). We compare their results with prompts 1,2, and 8, which contain all traits of prompt 7. ProTACT for target-prompt 7 achieved an average 6.4% increment, and especially a 10.3% increment for the *Content* trait; the improvement rate is almost twice that of prompts 1, 2, and 8. This is noticeable given the severely inferior baseline target-prompt 7 predictions of all three trait scores, except *Overall*. Another point to note is that prompts 1, 2, 8, and 7 all deal with long essays (Table 3) that require strong encoding ability (Wang et al., 2022), but improved 4.2% on average, implying the efficacy of our encoding strategy.

For each trait scoring task (Table 5), ProTACT achieved an average 3.9% enhancement over the baseline system. In particular, noticeable improvements are shown in all traits except the *Overall*, which already had considerably higher performance than other traits. Multiple trait-scoring tasks share information between layers, so inferior tasks might benefit more than superior tasks. Thus, ProTACT alleviates the data shortages in specific trait-scoring tasks caused by partial-trait coverage.

5.1 Ablation Studies

Incremental Analysis To explore the impact of each model component, we conducted an incremental analysis. Starting from encoding essay representation with a multi-head self-attention mechanism and using general essay features, we gradually added essay-prompt attention, topic-coherence fea-

[7] https://radimrehurek.com/gensim/

	Prompts									
Model	1	2	3	4	5	6	7	8	AVG	SD(↓)
*CTS-baseline	0.629	0.543	0.596	0.620	0.614	0.546	0.382	0.501	0.554	0.020
MSA	0.635	0.561	0.594	0.617	0.617	0.557	0.404	0.533	0.565	0.017
+ Essay-Prompt Att	0.638	0.559	0.595	0.624	0.615	0.567	0.397	0.531	0.566	0.017
+ TC feature	0.639	0.581	0.618	**0.634**	0.657	0.580	0.436	0.525	0.584	0.015
+ TS loss (ProTACT)	**0.647**	**0.587**	**0.623**	0.632	**0.674**	**0.584**	**0.446**	**0.541**	**0.592**	0.016

Table 7: Results of ablation studies. The average QWK scores over all traits for each **prompt**. *MSA* denotes multi-head self-attention, *TC feature* denotes the Topic-coherence feature.

	Traits										
Model	Overall	Content	Org	WC	SF	Conv	PA	Lang	Nar	AVG	SD(↓)
*CTS-baseline	0.670	0.551	0.459	0.562	0.556	0.413	0.568	0.533	0.610	0.547	0.012
MSA	0.671	0.562	0.486	0.580	0.573	0.441	0.568	0.545	0.610	0.560	0.012
+ Essay-Prompt Att	0.671	0.565	0.477	0.582	0.574	0.435	0.573	0.550	0.618	0.561	0.012
+ TC feature	0.673	0.592	0.500	0.591	0.577	0.444	0.612	0.570	0.633	0.577	0.012
+ TS loss (ProTACT)	**0.674**	**0.596**	**0.518**	**0.599**	**0.585**	**0.450**	**0.619**	**0.596**	**0.639**	**0.586**	0.009

Table 8: Results of ablation studies. The average QWK scores over all prompts for each **trait**.

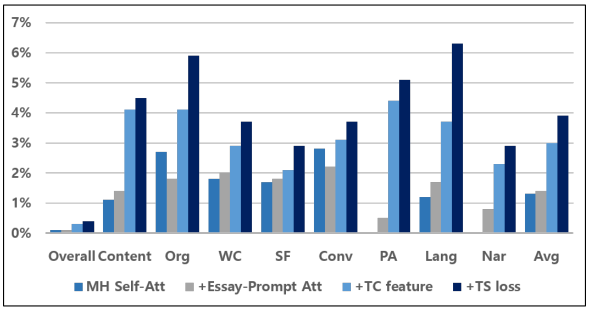


Figure 5: Improvement over the baseline model when incrementally applying each method of ProTACT.

ture, and TS loss. The results show both prompt- and trait-wise incremental advances (Tables 7,8).

In particular, remarkable improvements on all prompts after applying +*TC feature* prove that informing prompt-related knowledge facilitates scoring in cross-prompt settings (Table 7). Figure 5 shows increases in trait scoring tasks over the baseline (Table 8). The simple use of self-attention improves overall trait-scoring tasks, especially for syntactic traits such as *Conventions* and *Organization*, which evaluate overall grammatical writing conventions and essay structure, respectively. This matches our goal of multi-head self-attention capturing the structural and syntactic aspects. In contrast, supplementary use of essay-prompt attention somewhat decreases the scoring quality for those syntactic traits, yet particularly increases prompt-relevant traits such as *Prompt-Adherence* and *Narrativity*. Using the topic-coherence feature remarkably enhances scoring for *Prompt Adherence* and *Content* traits, which evaluates the essay's adherence to the topic and quantity of prompt-relevant text in the essay, respectively (Mathias and Bhattacharyya, 2018). The results on typical coherence-related traits (Shin and Gierl, 2022) prove that our feature explicitly supports related-aspect scoring and grows interpretability. Lastly, TS loss enhances all trait-scoring tasks, which shows the reflection of trait correlations boosts multi-trait joint learning.

δ	0.5	0.6	0.7	0.8	0.9
PCC	0.583	0.585	**0.586**	**0.586**	**0.586**
Cosine Similarity	0.585	**0.586**	0.584	0.585	0.585

Table 9: Results of the TS loss with variations

TS Loss with Variations To further optimize the TS loss, we have changed the criterion for the loss from PCC to cosine similarity. In addition, we experimented with the different values of the hyper-parameter δ for both conditions. Different δ values greater than 0.6 and condition change have little influence (Table 9). Since the average correlation between trait scores is 0.87 and the cosine similarity is 0.97, no significant variation appeared when constraining the similarity over high values.

6 Conclusion

We proposed a prompt- and trait relation-aware cross-prompt essay trait scorer (ProTACT) to improve AES in practical settings. Experimental results prove that informing prompt-relevant knowledge to the model assists the scoring of unseen prompt essays, and capturing trait similarities facilitates joint learning of multiple traits. Significant improvements in low-resource-prompt and inferior traits indicate the capacity to overcome the lacked pre-rated essays and strength in multi-trait scoring.

Limitations

The limitations of our work can be summarized in three points. First, as mentioned in Section 5, although a direct consideration of prompt information is helpful for related trait-scoring tasks, it may not be for irrelevant traits. Therefore, selectively applying each method depending on which traits are to score might further improve the model. Second, although the use of pre-engineered features, such as our topic-coherence feature, has the advantage of interpretability (Uto et al., 2020), it requires additional engineering steps, as in other AES studies using hand-crafted features (Amorim et al., 2018; Dascalu et al., 2017; Nguyen and Litman, 2018; Ridley et al., 2021). Finally, despite the large improvements observed on the specific datasets ASAP and ASASP++, the model has not experimented on other datasets. Feedback Prize dataset[8] is well-designed for scoring English-written argumentative writings with multiple trait labels, but the prompts are not defined; thus, it does not fit for cross-prompt AES. Essay-BR dataset (Marinho et al., 2022) contains essays on multiple prompts with labeled multiple trait scores. Thus, in future work, our proposed methods can be extended to multilingual cases of AES using the dataset.

Ethics Statement

We adhere to the ACL Code of Ethics. This work did not use any private datasets and did not contain any personal confidential information.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01906, Artificial Intelligence Graduate School Program (POSTECH)) and MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

[8] https://www.kaggle.com/competitions/feedback-prize-2021/data

References

Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237.

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

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. 2017. Readerbench learns dutch: building a comprehensive automated essay scoring system for dutch language. In *International Conference on Artificial Intelligence in Education*, pages 52–63. Springer.

Yann Dauphin, Harm De Vries, and Yoshua Bengio. 2015. Equilibrated adaptive learning rates for non-convex optimization. *Advances in neural information processing systems*, 28.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring-an empirical study. In *EMNLP*, volume 435, pages 1072–1077.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *CoNLL*, pages 153–162.

Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2022. Automated chinese essay scoring from multiple traits. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3007–3016.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mohamed A Hussein, Hesham A Hassan, and Mohammad Nassef. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. *Int J Adv Comput Sci Appl*, 11(5):287–293.

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.

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

Thomas K Landauer. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated essay scoring: A cross-disciplinary perspective*.

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*, pages 90–95.

Xia Li, Minping Chen, and Jian-Yun Nie. 2020. Sednn: shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems*, 210:106491.

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. Springer.

Jeziel C Marinho, Rafael T Anchiêta, and Raimundo S Moura. 2022. Essay-br: a brazilian corpus to automatic essay scoring task. *Journal of Information and Data Management*, 13(1).

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)*.

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

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

Huy Nguyen and Diane Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

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

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 preprint arXiv:2008.01441*.

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

Jinnie Shin and Mark J Gierl. 2022. Evaluating coherence in writing: Comparing the capacity of automated essay scoring technologies. *Journal of Applied Testing Technology*.

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.

Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):459–484.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

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.

Haoran Zhang and Diane Litman. 2019. Co-attention based neural network for source-dependent essay scoring. *arXiv preprint arXiv:1908.01993*.

A Detailed Ablation Studies

In our main paper, we have conducted the incremental analysis in Section 5.1 to examine the effect of gradually adding each model component. The results have shown that adding the TC feature to the model, where multi-head self-attention and essay-prompt attention are applied, yields the greatest performance improvement. To closely investigate the individual contribution of the seemingly effectual TC feature, we now compare the results of separately adding the TC feature and essay-prompt attention (Table 10).

The noticeable point is that despite little overall improvements when separately applying essay-prompt attention and the TC feature, their simultaneous application leads to significantly increased performance. These results indicate our proposed methods can yield synergies when jointly applied.

	Traits									
Model	Overall	Content	Org	WC	SF	Conv	PA	Lang	Nar	AVG	SD(↓)
MSA	0.671	0.562	0.486	0.580	0.573	0.441	0.568	0.545	0.610	0.560	0.012
MSA + Essay-Prompt Att	0.671	0.565	0.477	0.582	0.574	0.435	0.573	0.550	0.618	0.561	0.012
MSA + TC feature	0.672	0.562	0.485	0.585	0.565	0.428	0.609	0.568	0.629	0.567	0.011
MSA + Essay-Prompt Att + TC feature	**0.673**	**0.592**	**0.500**	**0.591**	**0.577**	**0.444**	**0.612**	**0.570**	**0.633**	**0.577**	0.012

Table 10: Results of detailed ablation studies. The average QWK scores over all prompts for each **trait**.

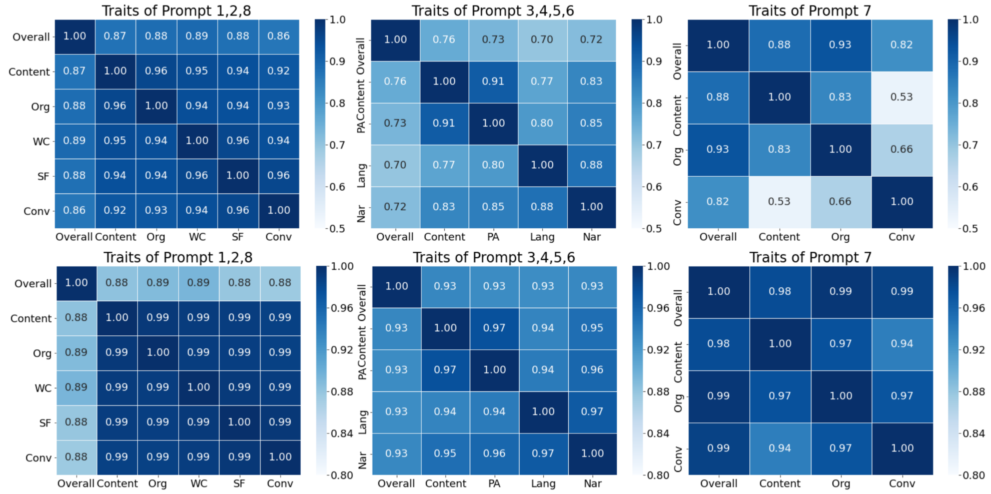


Figure 6: Correlation coefficient (1st row) and cosine similarity (2nd row) between ground-truth trait scores of all prompt types.

B Analysis of Trait Relationship

In Section 3.3, we showed the correlation coefficients and cosine similarities between the ground-truth trait scores of prompt types 1,2, and 8, which have the same trait composition. To further analyze relations between all different traits, we additionally examined trait scores of other prompts (Figure 6). Likewise, we investigated the relationship between trait scores within prompts that are evaluated of the same traits. The correlation and cosine similarity results within the same prompt sets show similar tendencies, although the specific values are different. This explains the construction of our TS loss, which has criteria of correlation between actual trait scores while reflecting cosine similarities of predicted trait scores. Moreover, we find out higher similarities between prompt-related traits such as *Prompt Adherence* and *Content*. However, a relatively low association is observed for traits with distinctive evaluation rubrics, such as *Conventions* and *Content* traits.

C Topic-coherence Feature and Related Traits

In the main paper, we examined the relationship of our topic-coherence feature with *Narrativity* trait score to see if extracted features using LDA truly reflect the coherence of the essay (Section 3.2). Since we subsequently found that the topic is highly related to the prompt, we additionally investigated the feature relationship with the *Prompt Adherence* trait score, which is another coherence-related trait (Shin and Gierl, 2022). We also examine the case of predicting target-prompt 5, where the training set contains essays of prompts except 5. Since only prompts 3–6 have *Prompt Adherence* trait for evaluation, plotted training set only contains essays of prompts 3,4 and 6. Figure 7 shows similar tendencies as the distribution with *Narrativity* trait, implying that the topic-coherence feature also conveys

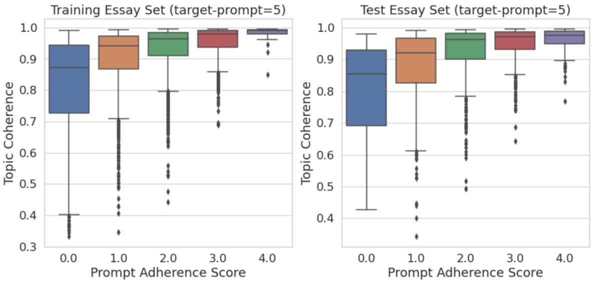


Figure 7: Box plot distributions of topic-coherence features of essays according to their *Prompt Adherence* trait scores.

whether the essay written adherent to the prompt. These findings further explain the observed significant improvements on *Prompt Adherence* trait scoring, in incremental analysis (Figure 5).

D Examples of the Prompt

Table 11 shows the specific examples of the prompt in the ASAP dataset, which we utilized. We encoded the corresponding prompt contents for each essay. Prompts 1–2 define argumentative essay writing, prompts 3–6 describe the writing of source-dependent essays, and prompts 7–8 define the narrative type of essays.

Prompt ID	Prompt
1	More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.
2	Censorship in the Libraries. "All of us can think of a book that we hope none of our children or any other children have taken off the shelf. But if I have the right to remove that book from the shelf – that work I abhor – then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us." –Katherine Paterson, Author. Write a persuasive essay to a newspaper reflecting your vies on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive? Support your position with convincing arguments from your own experience, observations, and/or reading.
3	Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.
4	Read the last paragraph of the story. "When they come back, Saeng vowed silently to herself, in the spring, when the snows melt and the geese return and this hibiscus is budding, then I will take that test again." Write a response that explains why the author concludes the story with this paragraph. In your response, include details and examples from the story that support your ideas.
5	Describe the mood created by the author in the memoir. Support your answer with relevant and specific information from the memoir.
6	Based on the excerpt, describe the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. Support your answer with relevant and specific information from the excerpt.
7	Write about patience. Being patient means that you are understanding and tolerant. A patient person experience difficulties without complaining. Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.
8	We all understand the benefits of laughter. For example, someone once said, "Laughter is the shortest distance between two people." Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.

Table 11: The eight prompts of the ASAP dataset.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section

☑ A2. Did you discuss any potential risks of your work?
Limitations Section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction (Section1)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section3.2, Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section3.2, Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 4

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4, Table 3

C ☑ Did you run computational experiments?

Section 4, 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4, 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.