Domain-Adaptive Automated Essay Scoring with Topic Relevance Learning

Sungjin Nam

ACT Education Corp., Iowa City IA 52243, USA

Abstract

This study investigates how capturing semantic relationships between essay prompts and responses can enhance the prediction performance in a domain-adaptive automated essay scoring (AES). Domain-adaptive AES models offer a balanced solution between cross-prompt models that are more generalizable but less accurate, and prompt-specific models that are more accurate but require training models for individual prompts. Our findings show that jointly training a model's *Relevance* block, which aims to learn the topical relevancy from prompt-response pairs using contrastive learning or classification methods, and *Scoring* block, which minimizes the regression loss, can significantly improve scoring performance in domain-adaptive AES tasks. Additionally, our models effectively mitigate the central tendency of predicted results, providing more reliable score predictions with substantially higher accuracies for low- and high-scored essays. Qualitative analysis results further demonstrate how our models capture the topical relevance between essay prompts and responses and improve the score predictions.

Keywords

Automated Essay Scoring, Multitask Learning, Contrastive Learning, Natural Language Processing

1. Introduction

Automated essay scoring (AES) systems are widely used in educational settings to provide efficient and consistent assessment results for a large volume of essay writings [1]. The models have evolved from traditional machine-learning models to deep neural network models, leveraging pre-trained models like GloVe [2] or BERT [3], to capture the nuanced semantic representations of essay responses [4]. Early AES models focused on capturing lexical quality indicators, such as grammatical error counts, linguistic complexity metrics, and other syntactic features [1]. While the models performed well on high-stakes tests, they were found to be susceptible to adversarial or off-topic responses [5, 6]. Recent advances in AES studies show that neural network models can more accurately predict human-annotated scores by capturing the semantic relationships between the essay prompts and responses [7, 8, 9, 10]. AES models that score essay responses with respect to the provided writing instructions could better reflect students' actual writing behaviors and improve the models' prediction performance [7]. However, existing studies tend to miss the balance between the model's generalizability and accuracy, which is crucial for real-world applications, and more detailed evaluations that can capture the model's performance across different score ranges.

Many AES studies were conducted using either prompt-specific or cross-prompt settings. Prompt-specific models train and evaluate on a single prompt, which provides higher accuracy but requires training multiple models [11, 12]. Cross-prompt models train on multiple prompts and evaluate on held-out prompts, offering generalizability but often sacrificing accuracy [13, 14]. Domain-adaptive models, on the other hand, balance the two approaches (Figure 3). They are advantageous in large-scale assessments by training a single model with essays from existing prompts and predicting scores across multiple prompts. They also maintain competitive accuracy without the need for managing numerous prompt-specific models [9, 10].

Moreover, evaluating AES models' performance often depends on overall scores, such as Quadratic Weighted Kappa (QWK) or correlation scores. Solely relying on these metrics may not provide detailed

EvalLAC'25: 2nd Workshop on Automatic Evaluation of Learning and Assessment Content, July 26, 2025, Palermo, Italy
sungjin.nam@act.org (S. Nam)
0000-0002-1893-4878 (S. Nam)

illustrations of the model's reliability across various score levels. Machine learning models may exhibit lower sensitivity to lower- or higher-scored examples, which are often underrepresented in training sets [15]. In the context of AES, models with high central tendencies may inaccurately reward poorly written essays or unduly penalize well-written ones, and misalign the output from the rating schema [16, 17]. Investigating the model's reliability across different score ranges can offer deeper understandings of how models perform on diverse essay responses. Qualitative analyses that take a closer look at what models have learned can also inform potential areas for improvement.

This study investigates whether jointly training the AES model with a non-scoring task, such as determining the topical relevance between essay prompts and responses, can enhance the AES model's overall and score-wise performance. Our contributions to the research community include:

1. We show that the domain-adaptive models balance the strengths and weaknesses of both cross-prompt and prompt-specific models, providing generalizability across multiple prompts while maintaining competitive accuracy.
2. We compare various methods to capture the topical relevance between essay prompts and responses to improve the overall score prediction performance.
3. We demonstrate that our multitask models also reduce the central tendency and provide more accurate predictions for both low- and high-scored essays.
4. Our qualitative analysis of essay prompts and responses provides a deeper understanding of how our approach enhanced the model's ability for AES.

2. Related Works

2.1. Domain-Adaptive Setting in AES

Prompt-specific AES models generally outperform cross-prompt models in terms of prediction accuracy. But training multiple models for each prompt can be costly and limit the creation of new writing prompts. Cross-prompt models can generalize across multiple prompts. However, using less accurate AES models in high-stakes tests can compromise the assessment program's reliability.

Domain-adaptive models [9, 10] can be an efficient solution that balances the two approaches for real-world applications. These models are trained to predict essay scores from multiple writing prompts. However, unlike the cross-prompt setting, evaluating essays may come from known or similar writing prompts that exist in the training dataset. While the model works on multiple prompts, the models can also achieve scoring accuracies that are comparable to those of prompt-specific models. In this study, we demonstrate that our domain-adaptive models significantly outperform cross-prompt AES models and achieve comparable performance to prompt-specific models.

2.2. Incorporating Essay Prompts and Other Responses in AES

Some AES studies aimed to capture the general quality of essays for accurate score predictions. This goal could be further sophisticated by comparing the essays' relative quality with other responses [12, 18], or by extracting common latent features for high quality essay writing across different prompts [14]. AES studies also have used essay prompts as input features [7, 8, 9, 19, 10]. Studies have employed attention mechanisms to capture the weighted vector values of the essay prompt and response texts [7] or sentence-level similarity scores between the prompts and responses [8] as input features for the downstream scoring task. In other machine learning applications, learning representations based on target labels, such as sentiment types or review ratings, has been demonstrated to enhance a model's generalizability and prediction performance [20, 21].

Our study tests contrastive learning objectives to leverage intricate relationships between essay prompts and student responses, training a model to represent positive examples in proximity, while pushing away negative examples. Contrastive learning can improve data-efficient training and maximize the utility of the training data. In NLP research, it has been shown to improve model representations [22]

and performance in classification tasks [21]. Previous AES studies have utilized contrastive pair-wise ranking to leverage relative scores between responses [12] or compared essay responses with similar scores to identify common quality features [14]. These studies relied on annotated scores to identify positive and negative responses for contrastive learning. Our study explores whether two methods of capturing the topical relevance between essay prompts and responses can enhance the accuracy of score predictions.

2.3. Multitask Learning in AES

Multitask learning models refer to machine learning models that can learn to solve multiple tasks simultaneously. These models can achieve better performance on individual tasks by leveraging shared information across the tasks [23]. Recent studies in AES have also explored the multitask learning approaches. For instance, combining different objectives, such as regression, ranking [8], or similarity loss functions [11] could significantly enhance the performance of prompt-specific AES models. Predicting multiple trait scores simultaneously, by leveraging inter-trait [13] or hierarchical structures [24], could also improve the AES models' prediction performance.

Capturing the relationships between writing prompts and responses can improve AES performance. [19] utilized prompt label prediction and sentence coherency classification tasks as multitask learning objectives to train AES models. Other studies have explored various learning tasks, such as measuring the distance between responses and topical clusters, next sentence classification [9], or calibrated regression loss with topical relevance probability [10], to improve score prediction accuracy. In this study, we specifically investigate whether multitask learning objectives that extensively capture the semantic relationships between essay prompts and responses, such as contrastive learning or classification methods, can simultaneously improve domain-adaptive score prediction performance.

3. Model Design

This study explores whether the multitask model design, which learns both topical relevance and scoring simultaneously, can improve overall predictions and score bin-wise reliability results (Figure 1). The Relevance block learns contrastive representations of essay prompts and responses (Section 3.1). For example, we assume that the embeddings of an assigned essay prompt and response pair from the dataset should be similar to each other, whereas the similarity score between an essay prompt and a response from a different prompt should be low. Simultaneously, the Scoring block learns to score using essay prompts and responses (Section 3.2). During training, the encoder weights in each block are softly shared and updated with L2 regularization (Section 3.3).

3.1. Relevance Block

The Relevance block requires positive and negative prompt-response pairs for training (Figure 1 (right)). For example, if there is an essay response to score (anchor response), a prompt originally assigned to a student is viewed as a positive example, while other randomly selected prompts are seen as negative examples. When considering a writing prompt as an anchor point, a response written for that prompt is considered positive, and responses from other prompts are regarded as negative. For this study, we used one negative example for an anchor pair; however, it can be easily expanded by using multiple negative pairs, as the computational resources permit.

3.1.1. Contrastive Learning with Essay Prompts and Responses

In existing studies, learning representations of texts with respect to the target labels has been shown to improve the model's generalization and prediction performance [20]. Specifically in [21], the authors presented supervised contrastive learning methods, using multiple loss functions to learn representations of text passages and labels, such as sentiments from Yelp reviews or genres of DBPedia articles.

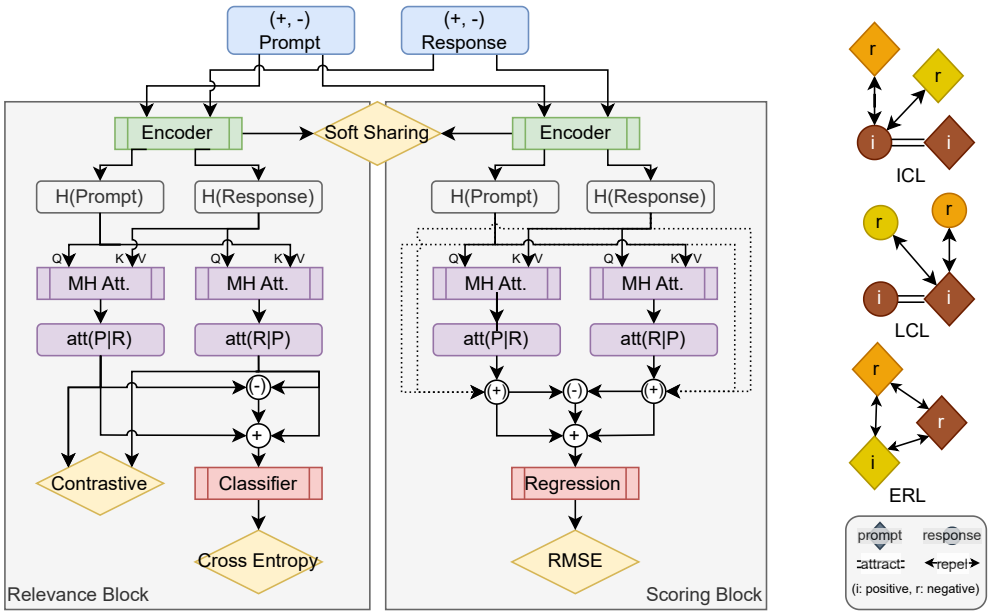

Figure 1: Left: This study used a multitask learning model with the Relevance and the Scoring blocks. The Relevance block learns contrastive representations of essay prompts and responses, while the Scoring block learns to score. The encoder weights are being softly shared. (·····›: residual connection, $+$: concatenation, $(+)$: element-wise sum, $(-)$: element-wise absolute difference.) **Right**: For the contrastive learning setting, we used a combination of loss functions. ICL attracts the embedding vector for an anchor response to the positive prompts (i) while repelling it from the negative prompts (r). LCL compares an anchor prompt with positive and negative responses. ERL ensures sufficient spacing between input prompts and output responses. [21].

In our study, we adapted this approach to essay responses and prompt identifiers, capturing the topical relevance between essay prompts and responses. However, unlike the original study [21], which trained new embeddings for the annotated labels, we utilized the essay prompts' [CLS] embeddings from transformer encoder models to leverage the rich content of essay prompts and simplify the model design. All similarity functions used cosine similarity.

We used three loss functions to learn the representations of essay prompts and responses. The Instance-Centered Loss (ICL) uses the InfoNCE loss [25] to increase the cosine similarity between the positive pairs (e.g., positive response x_i (anchor) and prompt y_i pair) (Eq. 1), and decrease the similarity between an anchor response x_i and negative prompts y_r $(i \neq r)$ (Eq. 2). τ is a scaling factor for the exponential function, and γ is a lower bound for the similarity score. Following [21], we used $\tau = 1/16$ and $\gamma = 0.1$. H denotes embeddings from the encoder. i and $r \in R$ are indexes of positive and negative examples from the batch. N is the number of responses in a batch.

$$S_{pr} = exp((sim(H_{x_i}, H_{y_i}) - \gamma)/\tau) \tag{1}$$

$$S_{\neg pr} = exp((sim(H_{x_i}, H_{y_r}) - \gamma)/\tau) \tag{2}$$

$$L_R^{ICL} = -\frac{1}{N} \log \frac{S_{pr}}{S_{pr} + \sum_{r \in R} S_{\neg pr}} \tag{3}$$

The Label-Centered Loss (LCL) function is similar to ICL, but it aims to increase the similarity between the positive prompt (anchor) and the response, and decrease the similarity between the anchor prompt and negative responses.

The previous study [21] used the the Embedding Regularizer Loss (ERL) function to regularize the label embeddings to be well dispersed in the embedding space. Instead, we used an average of two

ERL functions for responses ($L_R^{ERL_R}$) and prompts ($L_R^{ERL_P}$) to make both essay response and prompt embeddings are more separable. ($q, k \in \cup(R, i), q \neq k$).

$$L_R^{ERL_R} = \frac{1}{N}(\sum exp(1.0 + sim(H_{x_k}, H_{x_q})) - 1.0) \tag{4}$$

The contrastive loss (L_R^{CN}) comprises a weighted average of ICL, LCL, and ERLs. Following the previous study [21], we used $\lambda_R^{ERL} = 0.5$ [21].

$$L_R^{CN} = (L_R^{ICL} + L_R^{LCL} + \lambda_R^{ERL} \cdot L_R^{ERL})/3 \tag{5}$$

3.1.2. Positive and Negative Pair Classification

An alternative approach to contrastive learning is to train a classifier that directly predicts the essay prompt-response assignments. We created synthetic positive and negative pairs, similar to methods used in off-topic essay detection studies [26]. We selected one response and its assigned prompt as a positive pair. For negative examples, we paired the positive prompt with a randomly selected response from a different prompt, and the positive response with a randomly selected prompt.

For each input pair, we concatenated three vectors: the attention-weighted response vector, the prompt vector, and absolute differences between the two [27]. The single linear layer classifier was trained by using cross-entropy loss, measuring how well the sigmoid output aligns with the synthetic off-topic labels.

3.2. Scoring Block

For the Scoring block, we used a similar concatenated input from the classification module, but with additional residual connections (Figure 1). We used RMSE instead of MSE [12, 11, 10, 9, 14], for the predicted and annotated scores of positive responses to keep in scale with contrastive or cross-entropy loss values. The regression head was composed of a fully connected layer, dropout, ReLU, and a linear layer. The dropout probability was 0.3.

3.3. Input Representations

We tested two pre-trained encoders as backbones. We used BERT [3] to compare our results with the existing AES studies [9, 10, 11, 12]. The results from smaller-sized DistilBERT [28] also help us to examine the generalizability of our methods.

Following other AES and multitask learning studies, we used the embedding vector for the `[CLS]` token to represent an essay response or prompt [11, 29, 30]. For inputs exceeding the pre-trained encoder's context length (e.g., 512 tokens for BERT), we fetched multiple `[CLS]` vectors with a 128-token window and calculated their average. We used multi-head attention units with 8 heads to capture the semantic relationship between the essay prompts and responses. For example, to calculate the attention-weighted *response* vectors, the original response vector was used as the *query*, while the prompt vector served as the *key* and *value* inputs to the multi-head attention units. The process is reversed for the *prompt* vectors.

For the multitask learning settings, we used individual pre-trained encoders for the Relevance and Scoring blocks. All encoder layers were fine-tuned with their respective tasks. Each layer's weights were softly shared and regularized to minimize the L2-norms of layer-wise differences. Compared to other multitask AES models, where all tasks shared the same encoder module [14], the soft-sharing design may adapt better to multiple tasks [31, 32].

3.4. Summing Up the Loss Values

For the comprehensive model, we calculated the weighted sum of relevance (L_R^{CN}, L_R^{CL}), scoring (L_S), and encoder soft-share (L_E) losses. Based on a preliminary study, we used $\lambda_R^{CN} = 0.1$ and $\lambda_R^{CL} = 1.0$

Table 1
ASAP dataset by prompt. The response and prompt lengths are counted in the number of words and tokens from the BERT tokenizer.

Prompt	Essays	Avg. Response Len. (words/tokens)	Prompt Len. (words/tokens)	Score Range	Score Bins
1	1783	365.68 / 452.53	111 / 128	2-12	2-6, 7-8, 9-10, 11-12
2	1800	380.75 / 459.07	126 / 153	1-6	1-2, 3, 4, 5-6
3	1726	108.69 / 130.05	1087 / 1411	0-3	0, 1, 2, 3
4	1772	94.51 / 113.40	1411 / 2007	0-3	0, 1, 2, 3
5	1805	122.13 / 154.22	825 / 1050	0-4	0-1, 2, 3, 4
6	1800	153.30 / 195.23	1611 / 2028	0-4	0-1, 2, 3, 4
7	1569	168.20 / 213.36	58 / 65	0-30	0-6, 7-12, 13-18, 19-30
8	723	604.87 / 752.24	45 / 55	0-60	0-30, 31-35, 36-40, 41-60

for better scaling between the losses.

$$L = \lambda_R^{CN} \cdot L_R^{CN} + \lambda_R^{CL} \cdot L_R^{CL} + L_S + L_E \tag{6}$$

4. Experiment Setup

4.1. Dataset and Evaluation

In this study, we used the ASAP dataset [1] (Table 1). The dataset includes essay responses from eight prompts covering argumentative (Prompts 1, 2), source-dependent (3-6), and narrative (7, 8) writing tasks. Following the previous studies [9, 10], we used a domain-adaptive setting that splits the entire dataset into a 5-fold cross-validation setting (60:20:20). The domain-adaptive setting differs from the prompt-specific setting, where each train and test set is derived from a single prompt set, or the cross-prompt setting, which uses essays from multiple prompts and a held-out prompt for the test set (Figure 3). We used the same data splits from [33], following other prompt-specific AES studies [11, 18]. A single domain-adaptive model is trained with essays from all prompts and predicts the test set essays from all prompts. As we used the same data splits, the results are comparable to a collection of prompt-specific model evaluations (Figure 3 (Left)). All scores were scaled to a 0-1 range using min-max scaling per prompt to score essay responses from multiple prompts within the Scoring block.

The source-dependent essays (Prompts 3–6) include background stories in the writing prompts, making these prompts significantly longer than the argumentative or narrative essay prompts (825–1611 words; Table 1). To control the length effect across different writing prompts as inputs, we summarized the background stories for Prompts 3–6 using GPT-4o (`version:2024-08-06`). The summaries were manually reviewed to ensure quality and relevance to the original prompts, resulting in shorter summaries that are 96–156 words long. Instructions and excerpts for the summarized prompts are shown in Figure 2.

The QWK results show an overall model performance across multiple essay prompts. However, they do not describe the model's performance at varying score levels, such as low, mid, and high scores. To measure this, we categorized all the essay scores into the same number of bins and calculated the score bin-wise accuracy and coefficient of variation (CoV) scores. We binned all essay scores into four bins with hand-picked cut scores to minimize modifications (e.g., Prompts 3-6 already had four or five score bins) and ensure that there are enough examples assigned for each score bin (Table 1). The distribution of binned scores closely matched the original scores upon visual inspection (Figure 3). CoV was calculated as the ratio of the standard deviation to the mean ($\frac{s}{\mu}$) [15]. A lower CoV value indicates that the scores deviate less from the mean, or a higher central tendency in the model's predicted scores. For the accuracy scores, we counted the number of correct predictions and divided it by the actual

[1] https://www.kaggle.com/competitions/asap-aes/data

Instruction:
Summarize the story below in 50-100 words. The summarization should include useful information to answer the question.\n\n#Question: {{question}}\n\n#Story: {{context}}

Summarized Story:
P3: A cyclist in Lodi follows misguided advice from old-timers, leading him on a challenging shortcut to Yosemite. He faces ghost towns, rough roads, dehydration, and unhelpful landmarks like an old Welch's factory in the high desert. ... (108 words / 152 tokens)
P4: The author concludes the story with Saeng's vow to retake her driver's test in spring, symbolizing hope and resilience. After failing the test, Saeng finds solace in the familiar hibiscus, reminiscent of Vietnam, and shares a bonding moment ... (151 words / 195 tokens)
P5: Narciso Rodriguez's memoir conveys a warm, nostalgic mood, celebrating the essence of family, love, and community. Raised in Newark's Ironbound, his Cuban immigrant parents exemplified sacrifice and hospitality, turning their modest home into a vibrant hub ... (99 words / 132 tokens)
P6: The Empire State Building's design aimed to surpass the Chrysler Building by incorporating a dirigible mooring mast, inspired by aviation pioneers. Despite consultations with experts and tests, the mast faced obstacles such as the Hindenburg disaster highlighting hydrogen's ... (96 words / 135 tokens)

Figure 2: Instructions and excerpts of LLM-summarized stories for the source-dependent essays (Prompts 3–6). The length of summarized prompts is also noted in the parentheses.

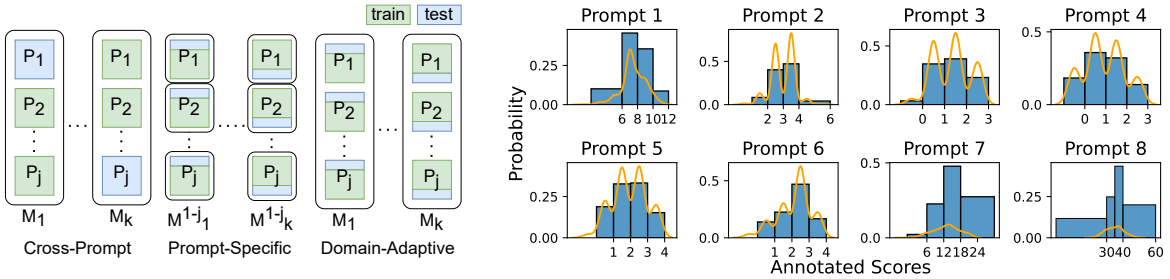

Figure 3: Left: A domain-adaptive model is trained and evaluated using responses from all prompts. It corresponds to a collection of training and test sets of the prompt-specific setting. This process is repeated across k cross-validation folds. This setting is distinguished from the cross-prompt setting, where a model is trained on essays from $j-1$ prompts and is evaluated with the j^{th} held-out prompt. **Right**: The distributions of the binned scores (blue) resemble the original scores (orange). The widths of the bars represent each score bin's range.

number of essays for each score bin. Metrics like score bin-wise accuracy and CoV are useful for evaluating the consistency and fairness of AES models [34, 35].

4.2. Baselines and Training Settings

Our baseline models were simple BERT-based regression models, fine-tuned solely with essay responses, without incorporating any writing prompts or multitask learning objectives. `Baseline-PS` models were trained and evaluated in a prompt-specific manner. `Baseline-DA` models used essays from all of the multiple writing prompts.

We included results from other domain-adaptive (DA) AES models for comparison. First, we considered domain-adaptive AES studies based on BERT [3] and multitask learning. `SST+DAT` [9] combined multiple objectives, such as next-sentence classification, noise detection, and distance to the topical cluster's centroid. `AOES` [10] used the off-topic probability of essay responses with respect to the writing prompt and calculated weighted scores. Although these studies' results are not directly comparable to ours, as they used different cross-validation splits for their evaluation, they serve as references for domain-adaptive models.

Additionally, we included the results from cross-prompt (CP) and prompt-specific (PS) AES studies. `PMAES` [14] is a state-of-the-art cross-prompt model that used contrastive learning with similarly

Table 2
Our domain-adaptive (DA) models demonstrated better prediction performance than the cross-prompt (CP) models, but they did not perform as well as the state-of-the-art prompt-specific (PS) models. The multitask learning models[‡] achieved better average QWK scores than the single-task models[†]. Numbers in **bold** and *italic* represent the best and second-best scores from the same pre-trained backbone. Statistically significantly different results (one-tailed wilcoxon test, $p < 0.05$) from the baseline models are marked (DA: *, PS: $^\circ$). Results from other studies were collected from each paper.

PLM	+Features	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
BERT	‡+P+CL+CN	*.816**	*.702*$^{*\circ}$	**.703**$^{*\circ}$	.810*	**.807**	*.808**	**.830***	.733*	*.776*$^{*\circ}$
	‡+P+CL	.812*	**.705**$^{*\circ}$	*.698*$^\circ$	**.820**$^{*\circ}$	**.807**	.807°	**.830***	**.737***	**.777**$^{*\circ}$
	‡+P+CN	.805*	.692$^{*\circ}$	.691°	*.818**	.800*	.807*	.829*	*.736**	.772$^{*\circ}$
	†+P	.794*	.674	.695	.803	.793	.803°	.816	.702*	.760*
	†(Baseline-DA)	.771°	.645	.685	.788	.791	.799°	.801	.670	.744°
	†(Baseline-PS)	**.818**	.645	.680	.802	.801	**.820***	.825	.716	.763*
Distil-BERT	‡+P+CL+CN	.803*	**.691**	*.698**	**.815**	.798	.803	**.823**$^{*\circ}$	.724*	*.769**
	‡+P+CL	**.813***	*.688*	.693*	.813	**.802**	.798$^{*\circ}$	.820*	**.738**$^{*\circ}$	**.771**$^{*\circ}$
	‡+P+CN	*.812**	.681*	.680*	**.815**	.793	*.811**	**.823***	*.725**	.768*
	†+P	.795	.656	**.704**$^{*\circ}$	.803	.800	.800°	.804	.702*	.758*
	†(Baseline-DA)	.788°	.653	.706$^{*\circ}$	.808	.794	.796°	.805	.638	.748
	†(Baseline-PS)	.810*	.671	.670	.814	**.802**	**.814***	.810	.666	.757
(CP)	PMAES [14]	.758	.674	.658	.625	.735	.578	.749	.718	.687
(PS)	BERT-LS [29]	.815	.672	.693	.816	.809	.814	.829	.717	.771
	Tr.-BERT [11]	.834	.716	.714	.812	.813	.836	.839	.766	.791
	NPCR [12]	.856	.750	.756	.851	.847	.858	.838	.779	.817
(DA)	AOES [10]	.793	.661	.667	.789	.782	.787	.845	.754	.760
	SST+DAT [9]	.824	.699	.726	.859	.822	.828	.840	.726	.792

scored essays across multiple prompts to capture common quality features of essay writing. All prompt-specific AES study results used BERT [3] as a backbone, and used the same test sets as our study [33], making the results directly comparable to ours. BERT-LS [29] utilized the logical relationship between sentences to enhance the performance of transformer-based AES models. Trans-BERT [11] pre-trained the model using all non-target prompts first, and then fine-tuned it as a prompt-specific scorer for the target prompt. Comparably, our approach is a single-stage training that is trained and evaluated with multiple prompts simultaneously. NPCR [12] used contrastive ranking with a scoring task, achieving state-of-the-art performance for a prompt-specific approach.

All models were trained for ten epochs with a 0.1 warm-up epoch. We used a batch size of 64, a gradient accumulation of two, and a cosine learning rate scheduler with learning rates of 2.5e-5 (BERT) or 5e-5 (DistilBERT).

5. Results

5.1. Overall Prediction Performance

Our models significantly outperformed the baseline models' overall QWK scores (Table 2). The single-task model with essay prompt input (+P) achieved significant improvements over the domain-adaptive baselines (Baseline-DA) and similar performance to the prompt-specific (Baseline-PS) baselines.

The multitask models provided additional significant improvements over all single-task models ($p < 0.05$). The models with the classification objective (+P+CL) performed better than the contrastive learning (+P+CN) or the combined (+P+CL+CN). These models performed similarly [29] or worse [12, 11] than other prompt-specific AES models. Although it was not directly comparable, our models also showed a higher overall QWK score than the state-of-the-art cross-prompt model [14] and were similar

Table 3
Nearly all multitask learning models[‡] and single-task models[†] demonstrated significant improvements in the accuracies of the lowest and highest score bins, overall accuracy, QWK, and CoV scores. Numbers in **bold** and *italic* represent the best and second-best scores from the same pre-trained backbone. Statistically significant different results from the Baseline models (DA: *, PS: $^\circ$) are also marked.

PLM	+Features	1	2	3	4	Acc.	QWK	CoV
BERT	‡+P+CL+CN	*+8.6*$^{*\circ}$	**+1.9**	-4.1$^{*\circ}$	*+7.9*$^{*\circ}$	*62.10**	**.735***	*.322*$^{*\circ}$
	‡+P+CL	+7.2*	**+1.9***	-4.3$^{*\circ}$	**+10.1**$^{*\circ}$	**62.22***	*.734**	.320$^{*\circ}$
	‡+P+CN	**+9.3**$^{*\circ}$	+0.5°	*-3.7*$^{*\circ}$	+7.8$^{*\circ}$	61.98*	.733*	**.323**$^{*\circ}$
	†+P	+2.8	+1.0	**-3.1***	+5.1*	59.96	.721*	.305$^{*\circ}$
	†(Baseline-DA)	46.94	64.61°	72.42	50.09	58.52	.702	.295°
	†(Baseline-PS)	+2.5	+2.8*	-0.9	+1.3	59.94	.727	.284*
Distil-BERT	‡+P+CL+CN	+3.6°	**+3.5***	*-2.4*$^\circ$	+2.9°	*61.55**	**.727***	**.322**$^{*\circ}$
	‡+P+CL	*+3.8*$^{*\circ}$	*+2.8*	-3.1$^{*\circ}$	*+4.4*$^{*\circ}$	61.32*	**.727***	*.320*$^{*\circ}$
	‡+P+CN	**+4.0**$^{*\circ}$	+1.8°	**-0.8**$^\circ$	+1.7°	**61.61***	.726*	.319$^{*\circ}$
	†+P	+2.6	+1.6	-2.5°	**+5.5**$^{*\circ}$	61.45	.722	.305$^{*\circ}$
	†(Baseline-DA)	50.76	62.11°	71.31°	54.39°	59.64	.707	.304°
	†(Baseline-PS)	-3.2	+3.6*	+3.0*	-4.9*	59.24	.722	.277*

to a domain-adaptive model [10].

The smaller DistilBERT-based `Baseline-DA` model performed better than BERT. However, the BERT-based models outperformed DistilBERT in multitask learning settings (e.g., +P+CL: 0.777 (BERT-based) vs. 0.771 (DistilBERT-based) ($p < 0.05$)). These results indicate that the larger backbone might gain greater advantages from the multitask learning design by capturing more nuanced semantic relationships between essay prompts and responses for scoring.

5.2. Score Bin-Wise Performance

The QWK results with the binned scores were lower than the original scores, but the overall patterns were similar (Table 3). Our models provided significant improvements in accuracies for the lower- or higher-end bins (e.g., 1, 4), while there were marginal improvements in the lower-mid scores (2) and significant decreases in the higher-mid scores (3). However, the overall benefits in accuracy scores over the baseline models were greater, especially with the multitask models.

We calculated CoV for min-max scaled scores across the prompts. Compared to the average of human-annotated scores (.369), our AES models' score predictions were still more centered around the mean, but significantly improved from both domain-adaptive and prompt-specific baselines ($p < 0.05$).

5.3. Visualizing Prompt and Response Embeddings

Lastly, we investigated how different model features impacted the attention-weighted essay prompt and response embeddings of the Scoring and Relevance blocks. We used t-SNE [36] to reduce the embeddings' dimensionality (Figure 4 (Top)).

For the Scoring block, the single-task baseline model (`Baseline-DA`) did not clearly grouped the essay responses by each prompt, especially for the source-dependent responses (Prompts 3-6). Adding essay prompts as input (+P) resulted in more distinct prompt-wise clusters of essay responses, although some essay prompts were still not clearly separated. The multitask models using the classification task (+P+CL or +P+CL+CN) showed similarly better clusters for both response and prompt texts.

The embeddings for Prompt 8 from the `Baseline-DA` did not clearly differentiate between the scores, whereas the multitask model with classification (+P+CL) demonstrated better clustering of low- and high-scored responses (Figure 4 (Bottom Left)).

The results from the Relevance block showed similar patterns because of the soft-sharing regularization between the two encoders (Figure 4 (Bottom Right)). The embeddings from the contrastive

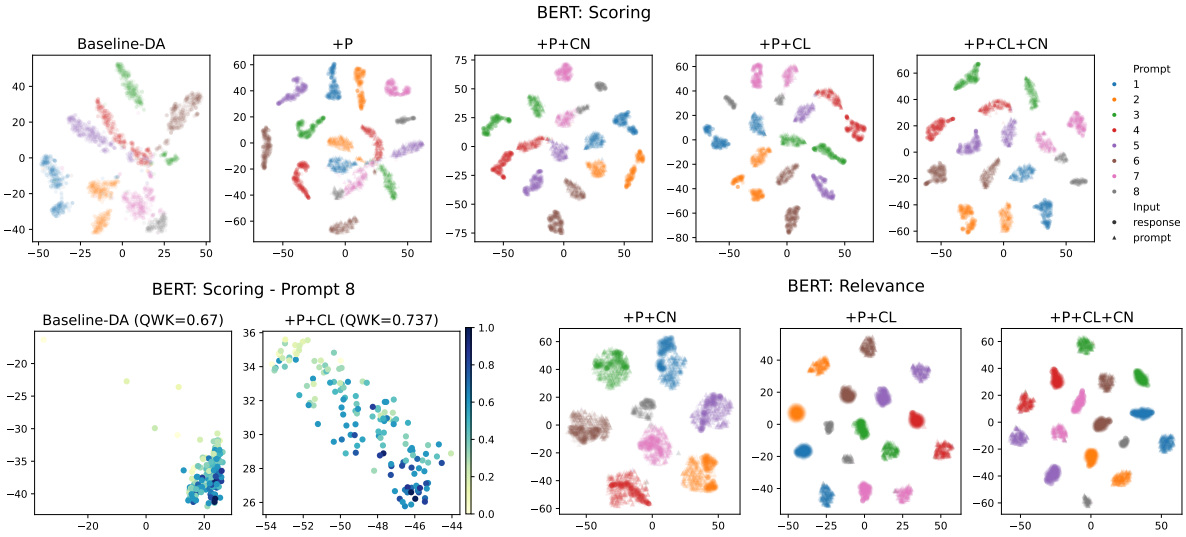

Figure 4: Scoring block embeddings (**Top**), example response embeddings with scores (Prompt 8; **Bottom left**), and Relevance block embeddings (**Bottom right**) from the BERT-based models. The visualization results show that adding essay prompt input (+P) or multitask learning objectives (+CN or +CL) provides clearer clusters of essay prompts and responses, and supports more accurate score predictions.

learning model (+P+CN) formed larger clusters that include both essay prompts and responses. The classification multitask models (+P+CL or +P+CL+CN) developed more distinctive clusters for essay prompts and responses. These results suggest that the multitask models can provide embeddings that are both topically distinct and better suited for the scoring task.

6. Conclusion and Discussion

This study investigated whether capturing the topical relevance between essay prompts and responses could provide more accurate and reliable AES predictions in domain-adaptive settings. Our multitask learning design significantly improved overall QWK scores and reliability across different score ranges. Also, the qualitative analysis showed that our methods can provide better representations of prompt and essay text relationships and distinctions between the scores. We believe our domain-adaptive AES models with the multitask learning design, offering generalizability across multiple prompts while maintaining competitive accuracy, can be useful for real-world essay assessments.

Although the results were promising, our models did not reach the state-of-the-art performance of the prompt-specific AES models [11, 12]. Adopting contrastive learning methods to extract general essay quality features [14] could enhance scoring performance by combining both domain-adaptive and domain-agnostic features. Testing the embedding quality of the Relevance block using a downstream task, such as off-topic detection [10], or evaluating the relevance of essay responses to their assigned prompts based on their writing quality, may also offer further insights. Additionally, t-SNE visualizations provided potential explanations for the improvements observed with our contrastive learning methods. A more formal and comprehensive investigation would further clarify why our methods are effective and how they can be improved.

Investigating hyperparameters or model architectures could lead to better results. For instance, our results indicated that the simple classification method provided marginally better scoring results than using the contrastive learning objectives. The visualization results also showed that the essay clusters from the classification (+P+CL) method were more distinctive than those from the contrastive learning method (+P+CN). This may indicate that the ERL loss, which encourages the model to maintain distance between the prompts (or responses) in the embedding space, was less effective than the simple classification method in distinguishing different essay prompt groups. More systematic investigations

into the combinations of contrastive loss functions or weight parameters would help enhance the model's performance. For this study, we employed a BERT-based model design to compare our results with other AES studies. Testing the method with more recent models, such as DeBERTa [37] or ModernBERT [38], would improve our ability to handle longer inputs with more accurate representations.

Training our multitask learning models took longer than the baselines or single-task models because they used an additional encoder unit and retrieved more essay response and prompt embeddings for contrasting examples. Investigations into more efficient model architectures, such as mixture-of-experts [30] or sparse sharing [32], may help address the resource issue and enhance the model's predictive capabilities in a parameter-efficient manner.

Lastly, our analyses were limited to the models trained with the ASAP dataset. In certain scenarios, the differences between the models were minimal. Conducting experiments with cross-prompt settings or additional datasets, such as PERSUADE [39], TOEFL writing [40], non-English essays, or other tasks like scoring constructive responses would provide more insights into the models' generalizability.

Declaration on Generative AI

During the preparation, the author(s) used GPT-4o-mini in order to: Grammar and spelling check.

References

[1] Y. Attali, J. Burstein, Automated essay scoring with e-rater® v. 2, The Journal of Technology, Learning and Assessment 4 (2006).

[2] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162/. doi:10.3115/v1/D14-1162.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[4] D. Ramesh, S. K. Sanampudi, An automated essay scoring systems: A systematic literature review, Artificial Intelligence Review 55 (2022) 2495–2527.

[5] L. Perelman, The babel generator and e-rater: 21st century writing constructs and automated essay scoring (aes), Journal of Writing Assessment 13 (2020).

[6] A. Kabra, M. Bhatia, Y. K. Singla, J. Jessy Li, R. Ratn Shah, Evaluation toolkit for robustness testing of automatic essay scoring systems, in: Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD), 2022, pp. 90–99.

[7] H. Do, Y. Kim, G. G. Lee, Prompt- and trait relation-aware cross-prompt essay trait scoring, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1538–1551. URL: https://aclanthology.org/2023.findings-acl.98/. doi:10.18653/v1/2023.findings-acl.98.

[8] Y. Yang, J. Zhong, Automated essay scoring via example-based learning, in: International Conference on Web Engineering, Springer, 2021, pp. 201–208.

[9] Y. Cao, H. Jin, X. Wan, Z. Yu, Domain-adaptive neural automated essay scoring, in: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 1011–1020.

[10] S. D. Das, Y. A. Vadi, K. Yadav, Transformer-based joint modelling for automatic essay scoring and off-topic detection, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings

of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 16751–16761. URL: https://aclanthology.org/2024.lrec-main.1457/.

[11] Y. Wang, C. Wang, R. Li, H. Lin, On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3416–3425. URL: https://aclanthology.org/2022.naacl-main.249/. doi:10.18653/v1/2022.naacl-main.249.

[12] J. Xie, K. Cai, L. Kong, J. Zhou, W. Qu, Automated essay scoring via pairwise contrastive regression, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 2724–2733. URL: https://aclanthology.org/2022.coling-1.240/.

[13] R. Ridley, L. He, X.-y. Dai, S. Huang, J. Chen, Automated cross-prompt scoring of essay traits, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 13745–13753. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17620. doi:10.1609/aaai.v35i15.17620.

[14] Y. Chen, X. Li, PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1489–1503. URL: https://aclanthology.org/2023.acl-long.83/. doi:10.18653/v1/2023.acl-long.83.

[15] J. M. Kernbach, V. E. Staartjes, Foundations of machine learning-based clinical prediction modeling: Part i—introduction and general principles, in: V. E. Staartjes, L. Regli, C. Serra (Eds.), Machine Learning in Clinical Neuroscience, Springer International Publishing, Cham, 2022, pp. 7–13.

[16] Y. Wang, Z. Wei, Y. Zhou, X. Huang, Automatic essay scoring incorporating rating schema via reinforcement learning, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 791–797. URL: https://aclanthology.org/D18-1090/. doi:10.18653/v1/D18-1090.

[17] H. Do, S. Ryu, G. Lee, Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 16427–16438. URL: https://aclanthology.org/2024.emnlp-main.917/. doi:10.18653/v1/2024.emnlp-main.917.

[18] R. Yang, J. Cao, Z. Wen, Y. Wu, X. He, Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1560–1569. URL: https://aclanthology.org/2020.findings-emnlp.141/. doi:10.18653/v1/2020.findings-emnlp.141.

[19] Y. Yang, J. Zhong, C. Wang, Q. Li, Exploring relevance and coherence for automated text scoring using multi-task learning., in: The 34th International Conference on Software Engineering and Knowledge Engineering, 2022, pp. 323–328.

[20] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, Z. Wu, Content attention model for aspect based sentiment analysis, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1023–1032.

[21] Z. Zhang, Y. Zhao, M. Chen, X. He, Label anchored contrastive learning for language understanding, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1437–1449. URL: https://aclanthology.org/2022.naacl-main.103/. doi:10.18653/v1/2022.naacl-main.103.

[22] J. Giorgi, O. Nitski, B. Wang, G. Bader, DeCLUTR: Deep contrastive learning for unsupervised textual representations, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 879–895. URL: https://aclanthology.org/2021.acl-long.72/. doi:10.18653/v1/2021.acl-long.72.

[23] R. Caruana, Multitask learning, Machine Learning 28 (1997) 41–75.

[24] R. Kumar, S. Mathias, S. Saha, P. Bhattacharyya, Many hands make light work: Using essay traits to automatically score essays, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1485–1495. URL: https://aclanthology.org/2022.naacl-main.106/. doi:10.18653/v1/2022.naacl-main.106.

[25] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[26] A. Louis, D. Higgins, Off-topic essay detection using short prompt texts, in: J. Tetreault, J. Burstein, C. Leacock (Eds.), Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Los Angeles, California, 2010, pp. 92–95. URL: https://aclanthology.org/W10-1013/.

[27] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.

[28] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[29] M. Yamaura, I. Fukuda, M. Uto, Neural automated essay scoring considering logical structure, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 267–278.

[30] J. Ma, Z. Zhao, J. Chen, A. Li, L. Hong, E. H. Chi, Snr: Sub-network routing for flexible parameter sharing in multi-task learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 216–223.

[31] S. Ruder, J. Bingel, I. Augenstein, A. Søgaard, Latent multi-task architecture learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 4822–4829.

[32] T. Sun, Y. Shao, X. Li, P. Liu, H. Yan, X. Qiu, X. Huang, Learning sparse sharing architectures for multiple tasks, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 8936–8943.

[33] K. Taghipour, H. T. Ng, A neural approach to automated essay scoring, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1882–1891. URL: https://aclanthology.org/D16-1193/. doi:10.18653/v1/D16-1193.

[34] K. P. Yancey, G. Laflair, A. Verardi, J. Burstein, Rating short L2 essays on the CEFR scale with GPT-4, in: E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, T. Zesch (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 576–584. URL: https://aclanthology.org/2023.bea-1.49/. doi:10.18653/v1/2023.bea-1.49.

[35] K. Yang, M. Raković, Y. Li, Q. Guan, D. Gašević, G. Chen, Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 22466–22474.

[36] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of Machine Learning Research 9 (2008).

[37] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with

gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, ????

[38] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, et al., Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, arXiv preprint arXiv:2412.13663 (2024).

[39] S. A. Crossley, P. Baffour, Y. Tian, A. Franklin, M. Benner, U. Boser, A large-scale corpus for assessing written argumentation: Persuade 2.0, Available at SSRN 4795747 (2023).

[40] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, M. Chodorow, Ets corpus of non-native written english ldc2014t06, Philadelphia: Linguistic Data Consortium (2014).