Graph-Based Multi-Trait Essay Scoring

Shengjie Li and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75080-0688
sxl180006@utdallas.edu, vince@hlt.utdallas.edu

Abstract

While virtually all existing work on Automated Essay Scoring (AES) models an essay as a word sequence, we put forward the novel view that an essay can be modeled as a graph and subsequently propose GAT-AES[1], a graph-attention network approach to AES. GAT-AES models the interactions among essay traits in a principled manner by (1) representing each essay trait as a trait node in the graph and connecting each pair of trait nodes with directed edges, and (2) allowing neighboring nodes to influence each other by using a convolutional operator to update node representations. Unlike competing approaches, which can only model one-hop dependencies, GAT-AES allows us to easily model multi-hop dependencies. Experimental results demonstrate that GAT-AES achieves the best multi-trait scoring results to date on the ASAP++ dataset. Further analysis shows that GAT-AES outperforms not only alternative graph neural networks but also approaches that use trait-attention mechanisms to model trait dependencies.

1 Introduction

While the majority of traditional work on Automated Essay Scoring (AES) has focused on *holistic* scoring (the task of assigning a single score to an essay that summarizes its overall quality), AES researchers have begun work on the relatively new task of *multi-trait* essay scoring, which is a natural extension of the holistic scoring task that involves scoring a given essay not only holistically but also along different dimensions of essay quality (a.k.a. *traits*), such as ORGANIZATION and COHERENCE. The surge of interest in multi-trait scoring was in part propelled by the release of ASAP++ (Mathias and Bhattacharyya, 2018a), the first publicly available corpus in which an essay is annotated with both its holistic score and its trait scores.

Multi-task essay scoring opens exciting opportunities for both AES researchers and users. Specifically, ASAP++ has made it possible for researchers to develop joint models that permit different traits to influence each other when they are being scored simultaneously. AES users could gain more detailed feedback on their own essays. For example, if they receive a low holistic score, they can, through the trait scores, better understand which aspects of their essay need improvement.

Early work on multi-trait scoring has arguably not been very successful, producing holistic and trait scoring results that are at best mediocre (Kumar et al., 2022). In particular, not only does the holistic score fail to benefit from the incorporation of trait scoring into the model, but the trait scores fail to benefit each other. One reason could be that these multi-trait scoring architectures cannot adequately capture the interactions among different traits, as in these architectures the traits can only interact with each other *indirectly* via a shared representation layer. Another reason could be that traits simply cannot benefit each other: after all, they are *different* dimensions of essay quality.

To better understand whether the scoring of different traits can benefit each other, we compute the Pearson Correlation Coefficient (PCC) for each pair of traits in ASAP++. As can be seen in Table 1, *all* pairs of traits are positively correlated with a PCC of at least 0.6, indicating a strong correlation, and 18 trait pairs even have a PCC of at least 0.8, indicating a *very* strong correlation.[2] A closer examination of the rubrics reveals the reason for these high PCC scores. Consider, for example,

[1] Our code and trained checkpoint are available at https://github.com/samlee946/GAT-AES/

[2] All correlations are statistically significant at the $p < 0.001$ level. Note that not all traits are evaluated for all prompts. For example, LANGUAGE appears in prompts 3–6, while ORGANIZATION appears in prompts 1, 2, 7, and 8. Thus, no PCC is available for such trait pairs, and we use an em dash in the table to indicate this. Similar trends can also be observed on the ICLE++ essay corpus (Li and Ng, 2024d), where certain pairs of traits are strongly correlated.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33325–33351
November 4-9, 2025 ©2025 Association for Computational Linguistics

	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice
Overall	.697	.706	.643	.684	.685	.600	.725	.687	.883	.831
Content		.913	.759	.833	.854	.713	.852	.834	.661	.861
PA			.780	.848	–	–	–	–	–	–
Lang				.862	–	–	–	–	–	–
Nar					–	–	–	–	–	–
Org						.802	.833	.823	.758	.828
Conv							.852	.893	.748	.683
WC								.872	–	.839
SF									–	.762
Style										–

Table 1: Pearson Correlation between traits in the ASAP++ dataset.

Sentence Fluency and Conventions, which have a PCC of 0.893. Sentence Fluency concerns whether the writing has an easy flow, and whether sentences have varied structures that make oral reading easy, whereas Word Choice concerns proper use of conventions (e.g., spelling, capitalization, grammar). Intuitively, failure to use conventions properly could result in sentences that are difficult to parse and understand, implying a positive correlation between the two traits.

Given that trait interdependencies exist, there have been several attempts at modeling trait dependencies. For instance, some multi-trait scoring models are designed to output the trait scores in a pre-specified order (Do et al., 2024a,b; Chu et al., 2025), but the implication is that only the traits that are predicted later in the output sequence can leverage information from earlier predictions. Another line of work captures trait interaction using a trait-attention mechanism (Ridley et al., 2021; Do et al., 2023), but they can only capture one-hop dependencies between traits.

In light of these weaknesses, we propose GAT-AES, a graph attention network (GAT) approach to multi-trait essay scoring. The key advantages of a graph-based representation of an essay are that we can easily (1) allow the traits to *directly* interact with each other and, in addition, (2) capture *multi-hop* dependencies, by using a graph node to represent each trait and fully connecting these nodes to capture their interactions. To our knowledge, we are the first to propose modeling an essay as a graph, as virtually all existing work on AES models an essay as a word sequence.

Following existing work on multi-trait essay scoring (Kumar et al., 2022; Do et al., 2024a,b), we evaluate GAT-AES in a *within-prompt* setting, in which AES systems are evaluated on essays written for prompts that are seen during training. GAT-AES achieves the best multi-trait essay scoring results to date on the ASAP++ dataset.

2 Related Work

In this section, we discuss related work on holistic scoring and trait scoring.[3]

2.1 Holistic Scoring

The vast majority of work on holistic scoring has focused on *within-prompt* holistic scoring, which involves scoring essays written for prompts that have already been seen during training. Early approaches to holistic scoring are *rule-based* (Attali and Burstein, 2006) or are built using traditional machine learning algorithms that focus on feature engineering (Larkey, 1998; Burstein et al., 1998; Miltsakaki and Kukich, 2004; Yannakoudakis et al., 2011), while recent approaches are deep learning-based (Uto et al., 2020; Wang et al., 2022b; Xie et al., 2022; Boquio and Naval, 2024; Das et al., 2024). Some recent work has focused on cross-prompt holistic scoring, where models are trained on essays from existing prompts and evaluated on essays from an unseen prompt (Phandi et al., 2015; Cummins et al., 2016; Jin et al., 2018; Li et al., 2020; Ridley et al., 2020; Jiang et al., 2023; Chen and Li, 2023; Zhang et al., 2025a; Wang et al., 2025).

2.2 Trait Scoring

Early approaches to trait scoring have focused on *single-trait* scoring, where heuristics or features are hand-crafted to optimize performance for a specific trait, such as Coherence (Higgins et al., 2004; Somasundaran et al., 2014; Wu et al., 2023), Organization (Persing et al., 2010), Thesis Clarity (Persing and Ng, 2013), Prompt Adherence (Persing and Ng, 2014; Zhuang et al., 2024), Argument Persuasiveness (Persing and Ng,

[3] For a comprehensive overview of AES research, we refer the reader to the books published by Shermis and Burstein (2003), Shermis et al. (2010) and Beigman Klebanov and Madnani (2021), as well as the surveys published by our group (Ke and Ng, 2019; Li and Ng, 2024a,b).

2015; Carlile et al., 2018), STYLE (Mathias and Bhattacharyya, 2018b), and THESIS STRENGTH (Ke et al., 2019). Recent efforts have shifted toward multi-task learning models that jointly predict trait scores and holistic scores using multiple linear heads on top of shared essay representations, which are obtained automatically (Kumar et al., 2022; Shibata and Uto, 2022; He et al., 2022; Do et al., 2023; Chen and Li, 2024; Wang and Liu, 2025).

To account for interdependencies between traits, some multi-trait scoring models are designed to output the trait scores in a pre-specified order (Do et al., 2024a,b; Chu et al., 2025), but, as mentioned before, only the traits that are predicted later in the output sequence can leverage information from earlier predictions, while other models capture trait interaction using a trait-attention mechanism (Ridley et al., 2021; Do et al., 2023), which can only capture *one-hop* dependencies between traits. While some improvements have been made to these multi-trait scoring models that involve (1) enriching the input to the model with LLM-generated rationales for each trait that provide explanations of how essays align with specific trait rubrics (Chu et al., 2025) and (2) using reinforcement learning to directly optimize for score-related metrics such as Quadratic Weighted Kappa (QWK)[4] (Do et al., 2024b), these improvements have nothing to do with improving the way trait interdependencies are captured.

3 GAT for Multi-Trait Essay Scoring

In this section, we describe our GAT-AES framework, which leverages a graph attention network (GAT) (Veličković et al., 2018) to model the interactions among traits, between traits and essays, and between essays and hand-crafted features.

3.1 Graph Construction

GATs are typically employed for downstream tasks with graph-structured inputs. Since essay scoring lacks an inherent graph structure, the design of an appropriate graph structure becomes crucial for the successful application of GATs to AES. Figure 1 illustrates how we construct the graph from an input essay. Details of this process are described below.

3.1.1 Node Construction

The graph consists of three types of nodes. Two types of nodes correspond to the two types of input features that are commonly used to represent an essay, namely the essay text and the statistical features computed from the essay. The remaining type of nodes corresponds to the traits.

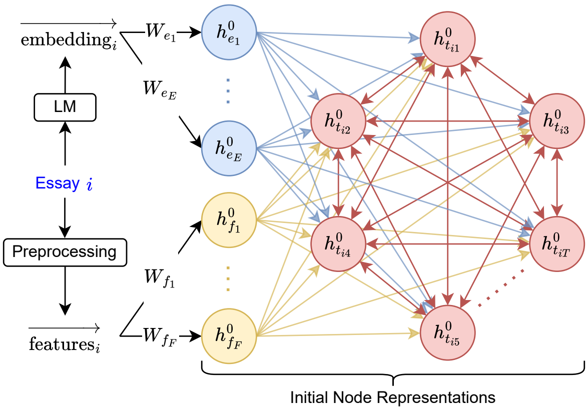

Figure 1: Constructing the graph from an input essay.

Text embedding nodes. We construct E text embedding nodes to encode the semantic information derived from the text of the input essay, where E is a hyperparameter. Using multiple nodes allows the model to encode semantic information from different portions of the essay text. The representations for these nodes are obtained by first extracting the essay embedding using a pre-trained language model LM and then transforming it via a learnable linear layer. Specifically, for each input essay i, its embedding is obtained as follows:

$$\overrightarrow{\text{embedding}_i} = \texttt{LM}(\text{essay}_i)$$

The initial representation $\vec{h}^0_{e_{ix}}$ for an embedding node x is computed using a learnable transformation:

$$\vec{h}^0_{e_{ix}} = \mathbf{W}_{e_x} \cdot \overrightarrow{\text{embedding}_i}$$

Hand-crafted feature nodes. We construct F hand-crafted feature nodes to encode prompt-independent information for the input essay, where F is a hyperparameter. Using multiple hand-crafted feature nodes allows the model to focus on different subsets of the input features. The representations for these nodes are obtained by transforming hand-crafted features via a learnable linear layer. Specifically, for each input essay i, the initial representation $\vec{h}^0_{f_{iy}}$ for a hand-crafted feature node y is obtained as follows:

$$\vec{h}^0_{f_{iy}} = \mathbf{W}_{f_y} \cdot \overrightarrow{\text{features}_i}$$

where $\overrightarrow{\text{features}_i}$ denotes the subset of the 1535 hand-crafted features for essay i used by Li and Ng

[4] QWK is the standard evaluation metric for AES. See Section 4.1.2 for details.

(2024c) that survive a feature selection process.[5]

Essay trait nodes. We construct T essay trait nodes that represent the essay traits to be predicted, where T is the number of traits in the dataset. While some researchers do not consider the OVERALL (i.e., holistic) score a trait score, we follow existing work on multi-trait scoring (e.g., Do et al. (2024a,b)) and view the OVERALL score as one of the trait scores in this paper. Note that while both the embedding nodes and the hand-crafted feature nodes correspond to inputs, the essay trait nodes correspond to outputs. For each trait z, its corresponding trait node is initialized using a standard normal distribution as a trainable vector $\vec{h}^0_{t_z}$ and is shared across all input essays.

3.1.2 Edge Construction

For each input essay i, we construct three types of edges.

Embedding-Trait edges. These edges connect each embedding node $\vec{h}_{e_{ix}}$ to each trait node $\vec{h}_{t_{iz}}$. This enables the model to capture the interactions between the traits and the input essay embedding.

Feature-Trait edges. These edges connect each hand-crafted feature node $\vec{h}_{f_{iy}}$ to each trait node $\vec{h}_{t_{iz}}$. This enables the model to capture the interactions between the traits and the hand-crafted features.

Trait-Trait edges. These edges connect each trait node $\vec{h}_{t_{iz}}$ to each of the other trait nodes. This enables the model to capture the interdependencies between different traits.

3.2 Trait Scoring

Using the notation introduced in the previous subsection, we can denote the graph constructed by the aforementioned graph construction process for input essay i as $\{\vec{h}^0_{e_{ix}}, \vec{h}^0_{f_{iy}}, \vec{h}^0_{t_{iz}} | 1 \leq x \leq E, 1 \leq y \leq F, 1 \leq z \leq T\}$, where $\vec{h}^0$ denotes the initial node representations. This graph will pass through L GAT layers, where L is a hyperparameter. At GAT layer l ($0 \leq l \leq L-1$), the convolutional operator operates on the graph $\{\vec{h}^l_{e_{ix}}, \vec{h}^l_{f_{iy}}, \vec{h}^l_{t_{iz}}\}$, and computes the updated node representations $\{\vec{h}^{l+1}_{e_{ix}}, \vec{h}^{l+1}_{f_{iy}}, \vec{h}^{l+1}_{t_{iz}}\}$ by dynamically aggregating features from neighboring nodes using multi-head self-attention. Each node is also a neighbor of itself. For a node u, GAT layer l aggregates messages from its neighbors $v \in \mathcal{N}(u)$ regardless of their types as follows:

$$\vec{h}^{l+1}_u = \bigoplus_{k=1}^{K} \sigma\left(\sum_{v\in\mathcal{N}(u)} \alpha^k_{uv} \cdot \mathbf{W}^k \vec{h}^l_v\right)$$

where K is a hyperparameter for the number of heads in l, $\bigoplus$ denotes concatenation across k attention heads, σ is an activation function, $\mathbf{W}^k$ is a learnable weight matrix that transforms the node representations into higher-level node representations, and α^k_{vu} is the attention weight between nodes v and u in head k:

$$\alpha^k_{vu} = \text{softmax}\left(\text{LeakyReLU}\left(\mathbf{a}^k\left[\mathbf{W}^k\vec{h}^l_u \| \mathbf{W}^k\vec{h}^l_v\right]\right)\right)$$

where $\mathbf{a}^k$ is a shared attentional vector obtained from a linear layer.

Since each node aggregates messages from its neighbors, the three types of edges allow the model to learn the interdependencies among traits, the dependencies between traits and embeddings, and the dependencies between traits and features via the self-attention mechanism and by dynamically adjusting node representations.

After passing through L GAT layers, the resulting graph $\{\vec{h}^L_{e_{ix}}, \vec{h}^L_{f_{iy}}, \vec{h}^L_{t_{iz}}\}$ will be used for trait scoring of essay i. Specifically, for each trait node z, a regression head (single fully connected layer) maps the final node representation $\vec{h}^L_{t_{iz}}$ to a scalar $\hat{y}_{iz}$, which is the predicted score for trait z:

$$\hat{y}_{iz} = \sigma(\mathbf{W}_{t_z} \cdot \vec{h}^L_{t_{iz}})$$

where σ is the sigmoid activation function.

3.3 Training

The model is trained end-to-end using the mean squared error (MSE) loss:

$$\mathcal{L}_{MSE} = \frac{1}{N}\frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{iz} - \hat{y}_{iz})^2$$

where N is the number of samples, T is the number of traits, and y_{iz} is the ground truth score scaled to range [0, 1]. During the training process, the transformations $\mathbf{W}$, the node representations $\vec{h}$, the attentional vectors $\mathbf{a}$, and the parameters from the pre-trained language model `LM` are optimized.

4 Evaluation

4.1 Experimental Setup

4.1.1 Datasets

For model training and evaluation, we employ the ASAP[6] corpus and its extension, ASAP++.

[5] Details about the features can be found in Appendix A. The feature selection process is described in Section 4.1.4.

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

Prompt	# of Essays	Traits
1	1783	Overall,Cont,WC,Org,SF,Conv
2	1800	Overall,Cont,WC,Org,SF,Conv
3	1726	Overall,Cont,PA,Nar,Lang
4	1772	Overall,Cont,PA,Nar,Lang
5	1805	Overall,Cont,PA,Nar,Lang
6	1800	Overall,Cont,PA,Nar,Lang
7	1569	Overall,Cont,Org,Conv,Style
8	723	Overall,Cont,WC,Org,SF,Conv,Voice

Table 2: Statistics on the combined ASAP and ASAP++ dataset. The trait names are abbreviated as follows: Cont: CONTENT, Org: ORGANIZATION, WC: WORD CHOICE, SF: SENTENCE FLUENCY, Conv: CONVENTIONS, PA: PROMPT ADHERENCE, Lang: LANGUAGE, Nar: NARRATIVITY.

ASAP (Automated Student Assessment Prize) is composed of essays manually annotated with their holistic scores. The essays are written for eight prompts, including two for persuasive essays, two for narrative essays, and four for source-dependent essays. Since different rubrics are used for scoring prompts, the score ranges for different prompts can be different. The eight prompts and their statistics can be found in Appendix B.

ASAP++ is an extension of ASAP where each essay is additionally scored along different traits. Ten traits are scored, including CONTENT (how clear and focused the writing is and how well-developed the main ideas are), WORD CHOICE (how well the words convey the intended message), ORGANIZATION (how well-organized the essay is), PROMPT ADHERENCE (how adherent the essay is to the prompt), SENTENCE FLUENCY (whether the sentences in the essay are of high quality), CONVENTIONS (how well the essay demonstrates standard writing conventions), NARRATIVITY (how coherent and cohesive the response is), LANGUAGE (how good grammar and spelling are), STYLE (how proficient and crisp the word choice is and how fluent the sentences are), and VOICE (how well the writer's commitment, expressiveness, and sense of audience enhance the writing's engagement and authenticity). Including the holistic score from ASAP, which reflects the overall quality of an essay, a total of eleven traits are manually annotated in the combined ASAP/ASAP++ dataset. However, the set of traits varies across prompts, as shown in Table 2.[7]

Since we perform within-prompt scoring, we follow the five-fold cross-validation setup from Taghipour and Ng (2016) with the same data partitions, as it has become the standard setup for evaluating multi-trait AES systems (Chu et al., 2025).

4.1.2 Evaluation Metric

We employ Quadratic Weighted Kappa[8] (QWK), which measures the agreement between model predictions and ground truth labels, as our evaluation metric. Higher values indicate better performance.

4.1.3 Baseline Systems

We employ six systems as our baselines. The first six systems are state-of-the-art within-prompt systems, namely HISK (Cozma et al., 2018), STL-LSTM (Dong et al., 2017), MTL-BiLSTM (Kumar et al., 2022), ArTS (Do et al., 2024a), SaMRL (Do et al., 2024b), and RMTS (Chu et al., 2025). The first two systems use separate models to score traits. The latter four systems perform multi-trait scoring, where the traits are scored using the same model.[9]

4.1.4 Implementation Details

Loss function. We employ the MSE loss. More specifically, since not all traits are applicable to all prompts, we use the *masked* MSE loss, where predictions for inapplicable traits do not contribute to the loss.

Feature selection. We employ feature selection on the set of hand-crafted features before feeding them into the feature nodes. For each feature f, we take the minimum of two correlation coefficients computed between f's values and the holistic scores for each prompt in the training set, Pearson and Spearman. We retain only those features whose minimum coefficient is greater than or equal to 0.2 when macro-averaged over the eight prompts.

Score rescaling. Since traits may have different score ranges across prompts, all trait scores are scaled to [0, 1] using the ranges provided in ASAP's official grading rubrics during training and re-scaled to their original ranges during evaluation.

Hyperparameters. GAT-AES is trained for 15 epochs using AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as the optimizer, and Devlin et al.'s (2019) $\texttt{BERT}_{\texttt{Large}}$[10] as the language model for obtaining embeddings and fine-tuning. The embedding of the

[7] This is because different prompts correspond to different types of essays: prompts 1–2 correspond to persuasive essays, prompts 3–6 correspond to source-dependent essays, and prompts 7–8 correspond to narrative essays.

[8] See https://www.kaggle.com/competitions/asap-aes/overview/evaluation for details.

[9] The results for the first four baselines are taken from Do et al. (2024a). A detailed description of each of these systems can be found in Appendix C.

[10] https://huggingface.co/google-bert/bert-large-cased

	Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG (SD)
1	HISK	.718	.679	.697	.605	.659	.610	.527	.579	.553	.609	.489	.611 (-)
2	STL-LSTM	.750	.707	.731	.640	.699	.649	.605	.621	.612	.659	.544	.656 (-)
3	MTL-BiLSTM	.764	.685	.701	.604	.668	.615	.560	.615	.598	.632	.582	.638 (-)
4	ArTS	.754	.730	.751	.698	.725	.672	.668	.679	.678	**.721**	.570	.695 (±.018)
5	SaMRL	.754	.735	.751	.703	.728	.682	.685	.688	.691	.710	.627	.705 (±.013)
6	RMTS	.755	.737	**.752**	**.713**	**.744**	.682	**.690**	.705	**.694**	.702	.612	.708 (±.043)
7	RMTS (rerunning code)	.729	.720	.739	.695	.730	.665	.656	.684	.672	.685	.609	.689 (±.018)
8	GAT-AES (w/ BERT$_{\text{Large}}$)	**.771**	**.742**	.749	.687	.726	**.694**	.686	**.709**	.692	.699	**.649**	**.710** (±.011)

Table 3: Trait scoring results of GAT-AES and the six baselines. The best result in each column is boldfaced.

[CLS] token is selected as the embedding of the essay. We set K, the number of attention heads in GAT, to 4; L, the number of GAT layers, to 2; the random seed to 11; the batch size to 32; and the dropout rate for both the LM and the GAT layers to 0.1. We perform a grid search to determine the remaining hyperparameters, including the learning rate, the hidden dimension of the node representations, the number of text embedding nodes E, and the number of feature nodes F. The model that achieves the highest QWK on the development data is selected for evaluation on the test set.[11]

4.2 Results and Discussion

4.2.1 Comparison with Baseline Systems

Results of trait scoring for each trait, when averaged over five folds and eight prompts, along with macro-averaged results over the traits and five-fold standard deviations, are shown in Table 3.[12] Rows 1–7 present the results of the six baselines. Note that rows 6 and 7 show the RMTS results obtained in two ways: in row 6 the results are taken verbatim from Chu et al. (2025), while in row 7 the results are obtained by our re-running the code provided by Chu et al., following their instructions. Row 8 shows the GAT-AES results.

Several observations can be made. First, GAT-AES achieves state-of-the-art performance for OVERALL, CONTENT, ORGANIZATION, WORD CHOICE, VOICE, and the AVG trait score. Moreover, GAT-AES is significantly better[13] than SaMRL and our re-run of RMTS.[14]

Second, GAT-AES consistently outperforms MTL-BiLSTM across all traits. Notice that ArTS, SaMRL, and RMTS are all generative models, whereas MTL-BiLSTM is the current best-performing regression model. This suggests that GAT-AES establishes a new performance baseline for regression-based approaches.

Finally, GAT-AES achieves the lowest standard deviation in average trait QWK scores across five folds, indicating consistent and robust performance across different folds.

4.2.2 Comparison with Other GNNs

As noted before, we are the first to propose a graph-based approach to AES. Hence, none of our baselines is graph-based. To get a better sense of how good GAT-AES is in comparison with other graph-based approaches (developed for non-AES tasks), we show in Table 4 the results of three alternative approaches: (1) GCN (Kipf and Welling, 2017), which serves as the foundational model in graph neural networks and is widely popular among researchers; (2) GraphSAGE (Hamilton et al., 2017), which is widely used in industry for its scalability and efficiency; and (3) GIN (Xu et al., 2019), which has high expressive power and is known for distinguishing non-isomorphic graphs, often outperforming GCN and GraphSAGE on several benchmarks. For comparison purposes, we include in the last row of Table 4 the results of GAT-AES.

As can be seen, the results show that GAT performs significantly better than the other GNNs, while GraphSAGE is significantly worse than the other GNNs. However, the results of GCN and those of GIN are statistically indistinguishable. The relative performances of these models should not

[11] Additional training details and the best-found hyperparameter values are reported in Appendix D.

[12] Prompt-wise results can be found in Appendix E.

[13] All statistical significance tests in this paper are one-tailed paired t-tests, with $p < 0.05$. To determine if model B is significantly better than model A on trait T, we collect the average trait QWK scores of both models across random seeds and all applicable prompts for T (e.g., for CONTENT, model A's population would consist of eight QWK scores, i.e., one QWK score for each prompt). We then perform a one-tailed paired t-test on the trait-wise results. Let $D = \text{QWK}_A - \text{QWK}_B$. The null hypothesis is $\mu_D \leq 0$, and the alternative hypothesis is $\mu_D > 0$.

[14] In order to conduct statistical significance tests with state-of-the-art models, we contacted the authors of SaMRL and RMTS but obtained detailed results for only SaMRL. Our re-run of the RMTS code only achieved a trait-wise average performance of .689, which is considerably below their reported score of .708 and significantly underperforms GAT-AES.

	Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG (SD)
1	GCN	.767	.740	**.754**	.686	.720	.690	.671	.698	.685	**.708**	.628	.704 (±.010)
2	GraphSAGE	.642	.573	.572	.538	.587	.492	.431	.485	.498	.401	.324	.504 (±.103)
3	GIN	.762	.736	.748	.682	.722	.696	.668	.691	.687	.689	.628	.701 (±.008)
4	GAT-AES (Ours)	**.771**	**.742**	.749	**.687**	**.726**	**.694**	**.686**	**.709**	**.692**	.699	**.649**	**.710** (±.011)

Table 4: Trait-scoring results of GAT-AES and alternative GNNs.

	Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG (SD)
1	RMTS (rerunning code)	.729	.720	.739	**.695**	**.730**	.665	.656	.684	.672	.685	.609	.689 (±.018)
2	T5$_{\text{Base}}$	.748	.718	.732	.673	.717	.674	.645	.675	.676	**.707**	.589	.687 (±.014)
3	BERT$_{\text{Large}}$	.749	.731	.745	.673	.718	.676	.668	.692	.681	.680	.602	.692 (±.013)
4	T5$_{\text{Base}}$ w/ GAT	.757	.729	.747	.691	.724	.683	.660	.687	.684	.698	.621	.698 (±.008)
5	BERT$_{\text{Large}}$ w/ Trait Att.	.766	.735	**.751**	.689	.718	.676	.676	.699	.684	.672	.630	.700 (±.010)
6	GAT-AES (Ours)	**.771**	**.742**	.749	.687	.726	**.694**	**.686**	**.709**	**.692**	.699	**.649**	**.710** (±.011)

Table 5: Trait-scoring results of GAT-AES and different backbones.

be surprising for the following reasons. First, while GAT may not outperform GCN, GraphSAGE, or GIN on benchmarks specifically designed for evaluating GNNs, it is particularly well-suited to our task. Specifically, the attention mechanism in GAT enables different trait nodes to prioritize information from the most relevant neighboring trait nodes. In contrast, although other GNNs may exhibit high expressive power and perform better on inputs with dynamic graph structures, this advantage does not apply to our task, as the graph structure remains fixed across all essays. Second, GCN treats all edges equally, which is not ideal because one trait might not depend on another (e.g., Prompt Adherence does not depend on Language). Third, GraphSAGE underperforms because it samples a different subset of neighbors each time (e.g., creating an edge between Prompt Adherence and Language for one essay and between Prompt Adherence and Content for another). This is problematic because trait interdependencies are not random. Lastly, GIN uses a sum operation to combine features from neighboring nodes without assigning explicit weights to individual edges, meaning it treats all edges equally, similar to GCN.

4.2.3 Comparison with Additional Models

To gain further insights into the effectiveness of GAT-AES, we conduct experiments involving T5$_{\text{Base}}$ and trait attention, as described below.

Recall that GAT-AES uses the word embeddings derived from BERT$_{\text{Large}}$, whereas RMTS uses T5$_{\text{Base}}$ as its backbone. The question, then, is: did GAT-AES outperform RMTS simply because BERT$_{\text{Large}}$ offers an advantage over T5$_{\text{Base}}$?

To answer this question, we first conduct an experiment in which we separately fine-tune T5$_{\text{Base}}$ and BERT$_{\text{Large}}$ using the same inputs as GAT-AES. However, in neither case do we use the GAT network to update node representations; rather, we use a linear head on top of the essay representation and features to score each trait. Doing so allows us to make a head-to-head comparison between T5$_{\text{Base}}$ and BERT$_{\text{Large}}$ on our multi-trait scoring task. Results are shown in rows 2 and 3 of Table 5. It turns out that the performance difference between the two models is statistically indistinguishable from each other, meaning that using BERT$_{\text{Large}}$ does not offer any advantage over using T5$_{\text{Base}}$ on our task. Moreover, comparing the RMTS results we obtained by re-running their code (row 1) and the fine-tuned T5$_{\text{Base}}$ results (row 2), we see that their performance difference is also statistically indistinguishable. In other words, the extensions to T5$_{\text{Base}}$ made by RMTS do not yield better results according to our experiments.

Next, we conduct an experiment to determine whether GAT can improve T5$_{\text{Base}}$ on multi-trait scoring by using T5$_{\text{Base}}$ in combination with GAT. The results, which are shown in row 4 of Table 5, are significantly better than the fine-tuned T5$_{\text{Base}}$ results in row 2, suggesting that GAT contributes substantial performance improvements over the T5$_{\text{Base}}$ backbone. These results also show that GAT is more effective than RMTS in improving T5$_{\text{Base}}$. Furthermore, GAT-AES (row 6), which combines GAT with BERT$_{\text{Large}}$, yields significantly better results than using GAT with T5$_{\text{Base}}$ (row 4). These results suggest that GAT is more effective in improving BERT$_{\text{Large}}$ than T5$_{\text{Base}}$: as shown earlier, these two models perform statistically indistinguishably on our task without GAT.

Finally, we conduct an experiment to determine whether the trait-attention mechanism introduced by Ridley et al. (2021) can improve BERT$_{\text{Large}}$ on

	Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG (SD)
1	GAT-AES (w/ $\mathtt{BERT}_{\mathtt{Large}}$)	**.771**	**.742**	.749	.687	**.726**	**.694**	**.686**	**.709**	**.692**	**.699**	**.649**	**.710** (±.011)
2	– w/ $\mathtt{MXBAI}_{\mathtt{Large}}$	.755	.730	.744	.673	.712	.680	.647	.681	.670	.696	.641	.694 (±.012)
3	– w/ $\mathtt{UAE}_{\mathtt{Large}}$	.738	.728	**.751**	.687	.718	.666	.637	.682	.664	.669	.624	.688 (±.011)
4	– w/ $\mathtt{BGE}_{\mathtt{Large}}$	.753	.722	.740	.669	.713	.661	.629	.676	.649	.691	.597	.682 (±.011)
5	– w/ $\mathtt{E5}_{\mathtt{Large}}$	.754	.731	.745	.667	.713	.668	.633	.667	.658	.686	.599	.684 (±.003)
6	– w/ $\mathtt{BERT}_{\mathtt{Base}}$	.746	.728	.746	**.691**	.716	.674	.655	.673	.671	.674	.619	.690 (±.009)

Table 6: Trait scoring results of GAT-AES when used in combination with embeddings provided by newer LMs.

	Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG (SD)
1	GAT-AES	**.771**	**.742**	.749	.687	.726	**.694**	**.686**	**.709**	**.692**	.699	**.649**	**.710** (±.011)
2	– w/o trait-trait edges	.754	.739	.750	.690	**.730**	.680	.653	.697	.669	.698	.593	.696 (±.010)
3	– w/ ArTS-style edges	.767	.734	.746	.692	.721	.688	.663	.699	.685	**.704**	.611	.701 (±.014)
4	– w/o embedding nodes	.655	.609	.615	.564	.636	.568	.490	.584	.594	.635	.541	.590 (±.015)
5	– w/o feature nodes	.763	.735	.745	.688	.720	.676	.661	.691	**.692**	.677	.617	.697 (±.011)
6	– w/ single embedding node and feature node	.761	.737	**.752**	**.694**	.721	.684	.668	.686	.683	.674	.632	.699 (±.017)

Table 7: Trait scoring results of GAT-AES when certain nodes or edges are removed from the graph.

our task. The results, which are shown in row 5 of Table 5, are significantly better than those of $\mathtt{BERT}_{\mathtt{Large}}$ (row 3), suggesting that trait attention can effectively improve multi-trait scoring results. However, these results are significantly worse than the GAT-AES results (row 6), meaning that GAT is more effective in improving $\mathtt{BERT}_{\mathtt{Large}}$ than Ridley et al.'s trait-attention mechanism for our task.

4.3 Ablation Studies

4.3.1 Effect of Embedding Models

Since we use a relatively older model ($\mathtt{BERT}_{\mathtt{Large}}$) as the LM for obtaining essay embeddings, a natural question is: can the overall multi-trait scoring performance be improved by using a more recent (and potentially stronger) LM? To investigate this question, we select four top-performing pre-trained LMs from the English Massive Multilingual Text Embedding Benchmark leaderboard v2[15] that are ranked 8th, 9th, 11th, and 24th among the 139 LMs with fewer than 5B parameters[16]. These LMs are Lee et al.'s (2024) $\mathtt{MXBAI}_{\mathtt{Large}}$, Li and Li's (2024) $\mathtt{UAE}_{\mathtt{Large}}$, Xiao et al.'s (2024) $\mathtt{BGE}_{\mathtt{Large}}$, and Wang et al.'s (2022a) $\mathtt{E5}_{\mathtt{Large}}$. We also include $\mathtt{BERT}_{\mathtt{Base}}$ to see whether there is a performance difference between variants of BERT.

Trait-wise results are reported in Table 6. As can be seen, GAT-AES w/ $\mathtt{BERT}_{\mathtt{Large}}$ consistently outperforms variants using other LMs in almost all traits. Specifically, $\mathtt{BERT}_{\mathtt{Large}}$ exhibits 3.1%-point and 2.7%-point improvements in QWK for CONVENTIONS and WORD CHOICE, respectively, in comparison to the second best-performing LM. Not surprisingly, $\mathtt{BERT}_{\mathtt{Base}}$ underperforms $\mathtt{BERT}_{\mathtt{Large}}$ on almost every trait. However, it outperforms all the newer LMs w.r.t. LANGUAGE, CONVENTIONS, and SENTENCE FLUENCY.

4.3.2 Effect of Edges and Nodes

Since message aggregation between nodes via edges is a core component of GAT, the configuration of edges and nodes may significantly impact performance. Below we present results for different edge and node configurations.

Effect of edge configuration. We experiment with two alternative edge configurations: (1) removing all trait-trait edges, and (2) an ArTS-style configuration that mimics ArTS's trait prediction order.[17] To implement the ArTS-style configuration, we connect each trait to all traits predicted earlier in the sequence. Results are shown in rows 2 and 3 of Table 7. Both configurations perform significantly worse than the fully connected configuration (row 1) in terms of trait-wise results, indicating that reducing the number of trait-trait edges significantly degrades performance. However, GAT-SAT with ArTS-style edges outperforms the variant without any trait-trait edges for most traits. This suggests that, although capturing interdependencies between only half of the pairs of traits is not optimal, it is still better than not modeling interdependencies between traits at all.[18]

Effect of node configuration. We experiment with three alternative node configurations: (1) re-

[15] `https://huggingface.co/spaces/mteb/leaderboard?benchmark_name=MTEB%28eng%2C+v2%29`

[16] Given the computing resources available to us, we are unable to employ models with more than 7B parameters, hence this restriction in our selection of models.

[17] ArTS predicts traits in the following sequence: VOICE → STYLE → SENTENCE FLUENCY → WORD CHOICE → CONVENTIONS → ORGANIZATION → NARRATIVITY → LANGUAGE → CONTENT → OVERALL.

[18] Additional analyses on how GAT-AES captures inter-trait dependencies can be found in Appendix F.

Trait	< -2 Diff	= -2 Diff	= -1 Diff	= 0 Diff	= 1 Diff	= 2 Diff	> 2 Diff
Overall	.04	.03	.15	.55	.18	.03	.03
Content	.00	.02	.18	.55	.23	.02	.00
PA	.00	.01	.16	.60	.22	.01	.00
Lang	.00	.01	.17	.59	.23	.01	.00
Nar	.00	.01	.17	.60	.22	.01	.00
Org	.00	.03	.20	.49	.24	.03	.00
Conv	.00	.02	.20	.51	.23	.03	.00
WC	.00	.02	.18	.53	.25	.02	.00
SF	.00	.01	.17	.53	.25	.03	.00
Style	.00	.03	.24	.52	.19	.01	.00
Voice	.01	.02	.19	.46	.27	.05	.00

Table 8: Characterization of the seriousness of the errors made by GAT-AES w.r.t. each trait. Each column shows the percentage of essays for which the predicted score and the gold score differ by a specific amount.

moving all embedding nodes (setting $E = 0$), (2) removing all feature nodes (setting $F = 0$), and (3) retaining only one embedding node and one feature node (setting $E = F = 1$). The first configuration forces GAT-SAT to rely solely on hand-crafted features, the second configuration forces it to rely exclusively on essay embeddings, and the third configuration reduces the expressiveness of GAT-SAT by reducing the amount of information captured from features and embeddings. Trait-wise results are shown in rows 4–6 of Table 7. As can be seen, all three configurations perform significantly worse with 1.1–12.0%-point drops in trait-wise average QWK scores. This suggests that reducing the number of nodes considerably degrades trait scoring performance for almost all traits, with embedding nodes being particularly important.[19]

4.4 Error Analysis

Next, we analyze the errors made by GAT-AES.[20]

To begin with, we show in the columns of Table 8 how different the scores predicted by GAT-AES are from the gold scores for each trait. As an example, consider PROMPT ADHERENCE (row 3): the predicted and gold scores are the same (i.e., the difference is 0) in 60% of the essays, and the predicted score is lower than the gold score by one point (i.e., a difference of -1) in 16% of the essays. As can be seen, across all traits, 46−60% of the essays are perfectly scored, and in 92−98% of the essays the predicted and gold scores differ by at most one point. Hence, most errors are near misses.

	Description
6	The writing has an effective flow and rhythm. Sentences show a high degree of craftsmanship, with consistently strong and varied structure that makes expressive oral reading easy and enjoyable.
5	The writing has an easy flow and rhythm. Sentences are carefully crafted, with strong and varied structure that makes expressive oral reading easy and enjoyable.

Table 9: Partial rubric for SENTENCE FLUENCY. The colors highlight the three key differences between the descriptions for these two scores.

The question, then, is: how much difference is there between two consecutive scores (e.g., 3 and 4) *semantically*? A closer look at the rubrics associated with the traits in ASAP++ reveals that in many cases the distinction is rather subtle. To exemplify, consider Table 9, which shows the (partial) rubric for scoring SENTENCE FLUENCY.

As can be seen, there is a subtle distinction between essays with a SENTENCE FLUENCY score of 5 and those with a score of 6. Keep in mind that while the gold scores are produced by human raters according to a rubric like this, GAT-AES does not have access to any of the rubrics: it has to infer these rubrics, or more precisely, identify the often subtle distinction between two score categories purely from the training data. How easy it would be to capture such fine-grained distinction is often affected by a number of factors, such as the presence/absence of features that can effectively encode this distinction (note that the features used by GAT-AES are far from being able to do so) and, more often, the score distribution in the training data (if a score category is under-represented, it would difficult to capture such fine distinctions).

Given this analysis, future work should explore incorporating information about the rubrics into scoring models, by designing rubric-aware features or encoding the score definitions in the input.

5 Conclusion

We presented GAT-AES, a novel graph-based approach for multi-trait essay scoring. By modeling input essays as graphs, GAT-AES provided a principled way of capturing the interdependencies among essay traits. Extensive experiments demonstrated its effectiveness in modeling trait interdependencies, which has in turn enabled it to achieve state-of-the-art trait scoring performance on the ASAP++ dataset and establish a strong baseline for regression-based essay scorers. In future work, we plan to further explore the potential of graph-based essay representations by using them for other AES tasks, such as cross-prompt scoring.

[19] Additional ablation results as well as an augmentation experiment can be found in Appendix G.

[20] A detailed error analysis can be found in Appendix H.

Limitations

We believe our work has several limitations. First, although we experimented with top-performing language models with fewer than 5B parameters, we did not explore larger models (e.g., Zhang et al.'s (2025b) Qwen3-Embedding-8B with 8B parameters) due to computational constraints. Second, we did not incorporate any information from the scoring rubrics into the models, but our error analysis revealed that such information may be useful.

References

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

Beata Beigman Klebanov and Nitin Madnani. 2021. *Automated Essay Scoring*. In Graeme Hirst, editor, *Synthesis Lectures in Human Language Technologies*. Morgan & Claypool Publishers.

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

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, Martin Chodorow, Lisa Braden-Harder, and Mary Dee Harris. 1998. Automated scoring using a hybrid feature identification technique. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 206–210, Montreal, Quebec, Canada. Association for Computational Linguistics.

Andrei M. Butnaru and Radu Tudor Ionescu. 2017. From image to text classification: A novel approach based on clustering word embeddings. *Procedia Computer Science*, 112:1783–1792. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

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

SeongYeub Chu, Jong Woo Kim, Bryan Wong, and Mun Yong Yi. 2025. Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5796–5814, Albuquerque, New Mexico. Association for Computational Linguistics.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia. Association for Computational Linguistics.

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

Sourya Dipta Das, Yash A. Vadi, and Kuldeep Yadav. 2024. Transformer-based joint modelling for automatic essay scoring and off-topic detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16751–16761, Torino, Italia. ELRA and ICCL.

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

Heejin Do, Yunsu Kim, and Gary Lee. 2024a. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian's, Malta. Association for Computational Linguistics.

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

Heejin Do, Sangwon Ryu, and Gary Lee. 2024b. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, USA. Curran Associates, Inc.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 185–192, Boston, Massachusetts, USA. Association for Computational Linguistics.

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2014. Can characters reveal your native language? a language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1363–1373, Doha, Qatar. Association for Computational Linguistics.

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

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

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give me more feedback II: Annotating thesis strength and related attributes in student essays. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3994–4004, Florence, Italy. Association for Computational Linguistics.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6300–6308, Macao, China.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France.

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

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

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024. Open source strikes bread - new fluffy embeddings model.

Shengjie Li and Vincent Ng. 2024a. Automated essay scoring: A reflection on the state of the art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024b. Automated essay scoring: Recent successes and future directions. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 6300–6308, Jeju, Republic of Korea.

Shengjie Li and Vincent Ng. 2024c. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024d. ICLE++: Modeling fine-grained traits for holistic essay scoring. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

(Volume 1: Long Papers), pages 8465–8486, Mexico City, Mexico. Association for Computational Linguistics.

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.

Xianming Li and Jing Li. 2024. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839, Bangkok, Thailand. Association for Computational Linguistics.

Sandeep Mathias and Pushpak Bhattacharyya. 2018a. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sandeep Mathias and Pushpak Bhattacharyya. 2018b. Thank “goodness”! a way to measure style in student essays. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 35–41, Melbourne, Australia. Association for Computational Linguistics.

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

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

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

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):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*, abs/2008.01441.

Mark D. Shermis, Jill Burstein, Derrick Higgins, and Klaus Zechner. 2010. Automated essay scoring: Writing assessment and instruction. In *International Encyclopedia of Education*, 3rd edition. Elsevier, Oxford, UK.

Mark D. Shermis and Jill C. Burstein. 2003. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Mahwah, NJ.

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

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 950–961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

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

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

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

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

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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

Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023. A multi-task dataset for assessing discourse coherence in Chinese essays: Structure, theme, and logic analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6673–6688, Singapore. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, USA.

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

Chunyun Zhang, Jiqin Deng, Xiaolin Dong, Hongyan Zhao, Kailin Liu, and Chaoran Cui. 2025a. Pairwise dual-level alignment for cross-prompt automated essay scoring. *Expert Systems with Applications*, 265:125924.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Xinlin Zhuang, Hongyi Wu, Xinshu Shen, Peimin Yu, Gaowei Yi, Xinhao Chen, Tu Hu, Yang Chen, Yupei Ren, Yadong Zhang, Youqi Song, Binxuan Liu, and Man Lan. 2024. TOREE: Evaluating topic relevance of student essays for Chinese primary and middle school education. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5749–5765, Bangkok, Thailand. Association for Computational Linguistics.

A Hand-Crafted Features

Table 10 enumerates the 1535 features used in our models alongside their detailed descriptions and the categories to which they belong. Specifically, features marked with superscript 1 are features derived using the textstat package[21]. Features marked with superscript 2 are computed using a readability package[22]. Those marked with superscript 3 are NLTK package-derived features[23]. Finally, those marked with superscript 4 are features obtained via the spaCy package[24]. The features can be categorized into nine groups:

1. *readability* features, which are derived from readability indices, such as the Coleman–Liau index;
2. *text complexity* features, which measure syntactic complexity, including metrics like the number of clauses per sentence;
3. *text variation* features, which capture the diversity of word and part-of-speech usage, such as the count of unique words;
4. *length-based* features, which include counts like the total number of words;
5. *sentiment-based* features, which assess sentiment at both the document and sentence levels, such as the proportion of positive sentences;

[21] https://github.com/textstat/textstat

[22] https://github.com/andreasvc/readability

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

[24] https://spacy.io/

6. *part-of-speech bigram* features, which encodes the count of a POS bigram that appears in the training data;

7. *pronoun-related* features, which consist of 218 word-based features designed to capture pronoun usage patterns; specifically, they include (i) the count of each pronoun (e.g., "I"), (ii) the count of pronouns belonging to each pre-defined group (e.g., first person pronouns), (iii) the number of sentences containing each pronoun, (iv) the number of sentences containing pronouns from each group, (v) the percentage of sentences containing each pronoun, and (vi) the percentage of sentences containing pronouns from each group;

8. *prompt adherence* features, which encode whether an essay is adherent to the prompt for which it is written;

9. *top-N words* features, which consist of 300 word-based features derived from the N most frequent words in the training data (with $N = 100$ in our experiments), where for each word w we compute (i) its count in an essay, (ii) the number of sentences in the essay containing w, and (iii) the percentage of sentences containing w.

The category names in Table 10 are marked with superscripts for source identification. Specifically, categories marked with superscript R are proposed by Ridley et al. (2020), categories marked with superscript U are proposed by Uto et al. (2020), and categories not marked with a superscript are proposed by Li and Ng (2024c).

The feature values are normalized as follows. Following Ridley et al. (2020), we apply min-max normalization to their features within each prompt, scaling them to the $[0, 1]$ range. Following Uto et al. (2020) and Li and Ng (2024c), we standardize their features within each prompt to achieve a mean of 0 and a standard deviation of 1.

Feature Group	Feature Name	Description
Ridley et al.'s (2020) Features (86 features)		
LB^{R}	word_count	The total number of words in the essay.
	mean_word	The average number of characters in each word.
	ess_char_len	The number of characters in the essay.
	mean_sent[3]	The average number of words in each sentence.
	characters_per_word[2]	The average number of characters in each word.
	avg_word_len	The average number of characters in each word.
	avg_words_per_sentence	The average number of words in each sentence.
	characters[2]	The number of characters in the essay.
	syllables[2]	The number of syllables in the essay.
	words[2]	The number of words in the essay.
	words_per_sentence[2]	The average number of words in each sentence.
	sentences_per_paragraph[2]	The average number of sentences in each paragraph.
	.[3]	The number of periods in the essay.
	,[3]	The number of commas in the essay.
	syll_per_word[2]	The average number of syllables in each word.
RB^{R}	automated_readability[1]	A readability metric that measures the readability of a text based on characters per word and words per sentence.
	linsear_write[1]	A readability metric developed for the U.S. Air Force to help them calculate the understandability of technical manuals, factoring in sentence length and words that are considered difficult.
	Kincaid[2]	A readability metric which estimate the readability of English texts based on sentence length and word length.
	ARI[2]	A readability metric that measures the readability of a text based on characters per word and words per sentence.
	Coleman-Liau[2]	A readability assessment that estimates the U.S. grade level required to understand a piece of text based on characters, words, and sentences.
	FleschReadingEase[2]	A readability metric that measures the readability of text based on syllables, words, and sentences. The scores are on a scale from 0 to 100, with higher scores indicating easier-to-read text.
	GunningFogIndex[2]	A readability metric that estimates the years of formal education a person needs to understand the text on the first reading.
	LIX[2]	A readability metric that considers sentence length and the percentage of long words (words with more than six characters) in a text.
	SMOGIndex[2]	A readability formula that estimates the education level needed to understand a piece of text by analyzing the number of polysyllabic words (words with three or more syllables) within the text.
	RIX[2]	A variant of the LIX readability index that only takes into account the average number of long words per sentence.
	DaleChallIndex[2]	A readability formula that uses word difficulty based on a list of familiar words, along with sentence length, to estimate the grade level required to understand a text.
	sentences[2]	The total number of sentences present in the essay.
	paragraphs[2]	The total number of paragraphs present in the essay.
	long_words[2]	The number of words that have 7 or more characters.
	complex_words[2]	The number of words that have 3 or more syllables.
	complex_words_dc[2]	The total number of words that are not in the Dale-Chall word list of 3000 words recognized by 80% of fifth graders.
TC^{R}	clause_per_s[4]	The average number of clauses per sentence.
	sent_ave_depth[4]	The average parse tree depth per sentence in each essay,
	ave_leaf_depth[4]	The average parse depth of each leaf node in the parse tree.

Continued on next page

Feature Group	Feature Name	Description
	max_clause_in_s[4]	The maximum number of clauses in the sentences of the essay.
	mean_clause_l[4]	The average number of words in each clause.
SB^R	overall_positivity_score[3]	Overall, how positive the essay is.
	overall_negativity_score[3]	Overall, how negative the essay is.
	positive_sentence_prop[3]	The percentage of positive sentences in the essay.
	neutral_sentence_prop[3]	The percentage of neutral sentences in the essay.
	negative_sentence_prop[3]	The percentage of negative sentences in the essay.
TV^R	sent_var[3]	The variance of the length of sentences in the essay.
	word_var[3]	The variance of the length of words in the essay.
	stop_prop	The percentage of stopwords in the essay.
	unique_word	The total number of unique words in the essay.
	type_token_ratio[2]	The number of unique words divided by the number of words.
	wordtypes[2]	The total number of unique words present in the essay.
	tobeverb[2]	The number of "to be" verbs in the essay.
	auxverb[2]	The number of auxilllary verbs in the essay.
	conjunction[2]	The number of conjunctions in the essay.
	pronoun[2]	The number of pronouns in the essay
	preposition[2]	The number of prepositions in the essay
	nominalization[2]	The number of nominalizations in the essay
	begin_w_pronoun[2]	The number of sentences in the essay that begin with a pronoun.
	begin_w_interrogative[2]	The number of sentences in the essay that begin with an interrogative.
	begin_w_article[2]	The number of sentences in the essay that begin with an article.
	begin_w_subordination[2]	The number of sentences in the essay that begin with a subordination.
	begin_w_conjunction[2]	The number of sentences in the essay that begin with a conjunction.
	begin_w_preposition[2]	The number of sentences in the essay that begin with a preposition.
	spelling_err[3]	The number of words that are not in The Brown corpus of the NLTK package.
	prep_comma[3]	The number of preprositions and commas in the essay.
	MD[3]	The number of tokens having a POS tag of MD in the text.
	DT[3]	The number of tokens having a POS tag of DT in the text.
	TO[3]	The number of tokens having a POS tag of TO in the text.
	PRP$[3]	The number of tokens having a POS tag of PRP$ in the text.
	JJR[3]	The number of tokens having a POS tag of JJR in the text.
	WDT[3]	The number of tokens having a POS tag of WDT in the text.
	VBD[3]	The number of tokens having a POS tag of VBD in the text.
	WP[3]	The number of tokens having a POS tag of WP in the text.
	VBG[3]	The number of tokens having a POS tag of VBG in the text.
	RBR[3]	The number of tokens having a POS tag of RBR in the text.
	CC[3]	The number of tokens having a POS tag of CC in the text.
	VBP[3]	The number of tokens having a POS tag of VBP in the text.
	JJS[3]	The number of tokens having a POS tag of JJS in the text.
	VBN[3]	The number of tokens having a POS tag of VBN in the text.

Continued on next page

Feature Group	Feature Name	Description
	POS[3]	The number of tokens having a POS tag of POS in the text.
	NNS[3]	The number of tokens having a POS tag of NNS in the text.
	WRB[3]	The number of tokens having a POS tag of WRB in the text.
	JJ[3]	The number of tokens having a POS tag of JJ in the text.
	CD[3]	The number of tokens having a POS tag of CD in the text.
	NNP[3]	The number of tokens having a POS tag of NNP in the text.
	RP[3]	The number of tokens having a POS tag of RP in the text.
	RB[3]	The number of tokens having a POS tag of RB in the text.
	IN[3]	The number of tokens having a POS tag of IN in the text.
	VB[3]	The number of tokens having a POS tag of VB in the text.
	VBZ[3]	The number of tokens having a POS tag of VBZ in the text.
	NN[3]	The number of tokens having a POS tag of NN in the text.
	PRP[3]	The number of tokens having a POS tag of PRP in the text.
	Uto et al.'s (2020) Features (25 features)	
LB^{U}	syllable_count	The number of syllables in the essay.
	num_words	The number of words in the essay.
	num_sentences	The number of sentences in the essay.
	lemma_count	The number of lemmas in the essay.
	,	The number of commas in the essay.
	!	The number of exclamation marks in the essay.
	?	The number of question marks in the essay.
TV^{U}	noun_count	The number of nouns in the essay.
	verb_count	The number of verbs in the essay.
	adverb_count	The number of adverbs in the essay.
	adjective_count	The number of adjectives in the essay.
	conjunction_count	The number of conjunctions in the essay.
	spelling_error_count	The number of spelling errors in the essay.
	stopwords_count	The number of stop words in the essay.
RB^{U}	ARI	A readability metric that measures the readability of a text based on characters per word and words per sentence.
	coleman_liau	A readability assessment that estimates the U.S. grade level required to understand a piece of text based on characters, words, and sentences.
	dale_chall	A readability formula that uses word difficulty based on a list of familiar words, along with sentence length, to estimate the grade level required to understand a text.
	difficult_words	The total number of words that are not in the Dale-Chall word list of 3000 words recognized by 80% of fifth graders.
	flesch_reading_ease	A readability metric that measures the readability of text based on syllables, words, and sentences. The scores are on a scale from 0 to 100, with higher scores indicating easier-to-read text.
	flesch_kincaid_grade	A readability metric which estimate the readability of English texts based on sentence length and word length.
	gunning_fog	A readability metric that estimates the years of formal education a person needs to understand the text on the first reading.

Continued on next page

Feature Group	Feature Name	Description
	linsear_write	A readability metric developed for the U.S. Air Force to help them calculate the understandability of technical manuals, factoring in sentence length and words that are considered difficult.
	smog_index	A readability formula that estimates the education level needed to understand a piece of text by analyzing the number of polysyllabic words (words with three or more syllables) within the text.
	Part-of-speech Bigram Features (902 features)	
POSB	(DT, NN) ...	The number of appearance of the bigram (DT, NN)
	Pronoun Features (218 features)	
PRO-Pronoun Count	pronoun_cnt_I ...	The number of pronoun "I" in the essay.
PRO-Pronoun Group Count	first_person_pronoun_cnt ...	The number of first person pronouns in the essay.
PRO-Sent Pronoun	sent_cnt_I ...	The number of sentences that contain "I"
PRO-Sent Pronoun Group	sent_first_person_pronoun ...	The number of sentences that contain first person pronouns.
PRO-Sent Pronoun Portion	percentage_sent_I ...	The percentage of sentences that contain pronoun "I".
PRO-Sent Pronoun Group Portion	percentage_sent_first_person ...	The percentage of sentences that contain first person pronouns.
	Prompt Adherence Features (4 features)	
PA	max_sentence_dot_score	Dot score between the embeddings of an essay and its prompt.
	mean_sentence_dot_score	The maximum dot score between the embeddings of sentences of an essay and its prompt.
	min_sentence_dot_score	The average dot score between the embeddings of sentences of an essay and its prompt.
	dot_score	The minimum dot score between the embeddings of sentences of an essay and its prompt.
	Top-N Words Features (300 features)	
TNW-Word Count	top_n_word_count_the ...	The count of "the" in the essay.
TNW-Sent Count	top_n_num_sent_have_the ...	The number of sentences in an essay that contains "the".
TNW-Sent Portion	top_n_percentage_sent_have_the ...	The percentage of sentences in an essay that contains "the".

Table 10: Description of the features along with their group information. Features marked with the superscript R are Ridley et al.'s (2020) features. Features marked with the superscript U are Uto et al.'s (2020) features. Group LB is composed of length-based features. Group RB is composed of readability-based features. Group TC is composed of text complexity features. Group TV is composed of text variation features. Group SB is composed of sentiment-based features. Group POSB is composed of the part-of-speech bigram features. Group PRO is composed of the pronoun-related features. Group PA is composed of the prompt adherence features. Group TNW is composed of the top-N words features.

	Prompt	Avg. # Words	# Essays	Score Range(s)
1	Write a letter to the editor of a newspaper about how computers affect society today.	365.4	1783	OVERALL: [2,12] Other: [1,6]
2	Write a letter to the editor of a newspaper about censorship in libraries	380.7	1800	OVERALL: [0,6] Other: [1,6]
3	Write a review about an article called Rough Rough Road by Joe Kurmaskie. The article will be provided.	108.5	1726	[0,3]
4	Explain why the author concludes the story the way the author did. The short story will be provided.	94.3	1772	[0,3]
5	Describe the mood created by the author in the memoir. Support your answer with relevant and specific information from the memoir	122.1	1805	[0,3]
6	Describe the difficulties that builders of the Empire State Building faced because of allowing dirigibles to dock there.	153.2	1800	[0,3]
7	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.	167.6	1569	OVERALL: [0,30] Other: [0,6]
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.	604.7	723	OVERALL: [0,60] Other: [1,12]
Overall		222.5	12978	

Table 11: The eight writing prompts in ASAP.

B Statistics on ASAP

Table 11 enumerates the eight essay prompts in ASAP. For each prompt, we additionally show the average word count, the number of essays, and the corresponding score range(s).

C Baseline Models

In this section, we provide a brief description of the baseline models used in our experiments.

HISK (Cozma et al., 2018) is a string kernel-based support vector regression model. It utilizes the histogram intersection string kernel (Ionescu et al., 2014) on character n-grams and bag-of-super-word embeddings (Butnaru and Ionescu, 2017) to obtain essay representations. These representations are then fed into ν-Support Vector Regression to predict trait scores. Each trait is predicted individually.

STL-LSTM (Dong et al., 2017) uses a CNN to generate sentence embeddings, followed by a LSTM that processes the sentence embeddings to produce essay-level embeddings. These embeddings are passed through a regression head with a sigmoid activation function to obtain the predicted trait scores. Like HISK, trait scores are predicted individually.

MTL-BiLSTM (Kumar et al., 2022) builds on the approach of STL-LSTM but employs a BiLSTM instead of a LSTM. It uses multiple regression heads (one for each trait, excluding OVERALL) to produce the predicted trait scores. The predicted trait scores, along with the essay embeddings, are then fed into a final regression head to obtain the predicted OVERALL scores. Unlike HISK and STL-LSTM, MTL-BiLSTM predicts all traits jointly.

ArTS (Do et al., 2024a) is an autoregressive sequence-to-sequence model for joint multi-trait scoring, using T5 as its backbone model. The input sequence to their model consists of a short prefix followed by an essay, while the output sequence includes multiple trait scores that are ordered by the number of prompts in which the traits appear. This ordering allows the traits predicted later in the sequence to leverage information from earlier predictions.

SaMRL (Do et al., 2024b) builds on ArTS. Instead of using supervised fine-tuning, SaMRL employs policy gradient reinforcement learning to fine-tune the T5 model with a multi-reward function that optimizes both MSE and QWK, resulting in improved trait-scoring performance.

RMTS (Chu et al., 2025) is the current state-of-the-art model for multi-trait scoring. It also builds on ArTS and employs supervised fine-tuning. However, RMTS incorporates LLM-generated rationales for each trait into the input sequence. These rationales provide explanations of how essays align with specific trait rubrics, enabling the model to achieve state-of-the-art results on ASAP/ASAP++.

D Training Details and Best-found Hyperparameters

In this section, we list the search range of each of the hyperparameters. The learning rate is searched out of the set $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times$

	Model	1	2	3	4	5	6	7	8	AVG (SD)
1	HISK	.674	.586	.651	.681	.693	.709	.641	.516	.644 (-)
2	STL-LSTM	.690	.622	.663	.729	.719	.753	.704	.592	.684 (-)
3	MTL-BiLSTM	.670	.611	.647	.708	.704	.712	.684	.581	.665 (-)
4	ArTS	.708	.706	.704	.767	.723	**.776**	**.749**	.603	.717 (±.025)
5	SaMRL	.702	**.711**	.708	.766	.722	.773	.743	.649	.722 (±.012)
6	RMTS	.716	.704	**.723**	**.772**	**.737**	.769	.736	.651	**.726** (±.042)
7	GAT-AES (Ours)	**.722**	.698	.707	.771	.720	.767	.744	**.680**	**.726** (±.009)

Table 12: Trait scoring results for each *prompt*.

$10^{-5}\}$. The hidden dimension of all node representations is searched out of the set $\{256, 512\}$. The number of text embedding nodes E is searched out of the set $\{1, 2, 4\}$. Finally, the number of hand-crafted nodes F is searched out of the set $\{1, 2, 4\}$. For both node types, we searched up to four nodes only. The reason is that we want the model to focus on interdependencies among the eleven trait nodes, and having too many embedding or feature nodes might cause the model to overly emphasize embeddings and features. Below we list the best-found hyperparameters for GAT-AES:

- learning rate $lr = 3 \times 10^{-5}$
- hidden dimension of all node representations $hid\_dim = 256$
- number of text embedding nodes $E = 1$
- number of hand-crafted feature nodes $F = 2$

It takes less than 300 hours to complete training for all experiments on a total of eight GPUs (two NVIDIA RTX A6000 48GB GPUs, two NVIDIA RTX 3090 24GB GPUs, and four NVIDIA RTX 6000 24GB GPUs).

E Trait Scoring Results for Each Prompt

We report results of trait scoring for each prompt that are averaged over five folds and applicable traits in Table 12. Several observations can be made.

First, RMTS and GAT-AES achieve the same state-of-the-art average prompt performance (0.726 QWK). However, GAT-AES shows a much lower standard deviation across folds (0.009 vs. 0.042). A similar comparison can be made using the standard deviations in Table 3, which indicate that GAT-AES is considerably more consistent and robust than RMTS.

Second, GAT-AES shows a 2.9%-point improvement in QWK over the second-best model on prompt 8. This improvement may be attributed to GAT-AES's more accurate scoring for Voice, which is only evaluated in prompt 8. As a result, the performance gain for Voice directly contributes to the overall performance gain for prompt 8.

Third, GAT-AES does not offer the strongest results for prompts 3–6, which consist of source-dependent essays. This underperformance appears to be due to lower scores for Language and Narrativity, as only prompts 3–6 evaluate these traits.

F Analysis of How GAT-AES Captures Inter-Trait Dependencies

To show how GAT-AES captures the interdependencies between traits, we conduct a case study of how GAT-AES scores some traits.

The gold scores for Sentence Fluency, Word Choice, and Conventions for the example essay in Table 13 are all 8, indicating perfect correlations. GAT-AES scored this essay with trait scores of all 9s. While it overestimates the gold scores by 1 point, it still reflects perfect correlations. This case study suggests that GAT-AES is able to capture the strong correlations among Sentence Fluency, Word Choice, and Conventions, and to make consistent score predictions for this example.

G Additional Experimental Results

In this section, we present additional experiments, including both ablation and augmentation studies.

G.1 Leave-one-trait-out Experiments

To shed some light into how GAT-AES captures trait interdependencies, we conduct a leave-one-out type of experiment, whereby different individual traits are removed during training. Results are shown in Table 14. A few observations can be made.

First, in most cases, removing any trait negatively affects the scoring of all remaining traits. The only exceptions are: (1) when removing Narrativity in row 5, where the scoring results for Prompt Adherence and Style increased by

I believe that laughter and joy are key elements that bring families and friends together. Being able to be in the company of those who make you laugh, is a greatly valued thing. Sometimes just sitting around and telling old stories, or playing board games can leave you with a sore gut because you have been laughing so hard. Many people these days have become so caught up in their lives and sometimes forget to take a moment and just laugh. I feel sorry for these people because they are missing out on the joy and enlightenment they could be sharing with the people around them. Family vacations are always chaotic, at least in my family, but they always turn out to be a memorable experience one way or another. Every winter our family gets together and goes to @ORGANIZATION1 to stay in our cabin there. A long weekend full of good food, trips to the mountain, snowball fights and family games, is a great environment to spark some laughter. Every year we bring big family games such as @LOCATION2, @CAPS1 and @LOCATION1. My family members and I tend to be very competitive people and the volume of the room in which the game is being played, tends to escalate almost through the roof. The whole house is filled with laughter and funny arguments over things like "how in the world is that a picture of a sock!" or "That's not fair, you know the actual definition!" I enjoy these times because they are memories that you can hold onto forever. Laughter is a part of happiness, and happiness needs to be a part of life. Spending time with those who make you laugh, are those that are worthy of your time. Wiser people than myself say that "life is short." I'm starting to realize that this statement is true. If life is short, then that time should be spent in the best way it can be; moments filled with laughter.

Table 13: A sample essay whose scores for Sentence Fluency, Word Choice and Conventions are all 8.

	Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice
1	GAT–AES	.771	.742	.749	.687	.726	.694	.686	.709	.692	.699	.649
2	w/o Overall	–	.728	.738	.683	.711	.687	.653	.695	.671	.683	.626
3	w/o Content	.759	–	.741	.680	.717	.678	.656	.692	.683	.707	.628
4	w/o PA	.759	.731	–	.682	.721	.685	.676	.692	.691	.688	.620
5	w/o Lang	.758	.739	.748	–	.719	.682	.662	.693	.689	.672	.636
6	w/o Nar	.762	.739	.752	.676	–	.689	.663	.705	.684	.700	.636
7	w/o Org	.749	.736	.749	.682	.723	–	.657	.698	.687	.692	.619
8	w/o Conv	.758	.732	.743	.681	.712	.683	–	.694	.680	.699	.617
9	w/o WC	.754	.736	.742	.683	.717	.685	.669	–	.687	.695	.625
10	w/o SF	.757	.734	.748	.686	.726	.682	.665	.685	–	.693	.593
11	w/o Style	.741	.708	.721	.653	.695	.661	.627	.674	.667	–	.597
12	w/o Voice	.738	.713	.728	.674	.704	.656	.630	.675	.673	.671	–

Table 14: Results of leave-one-trait-out experiments.

.003 and .001 points in QWK, respectively; and (2) when Content is ablated (row 2, +.008 QWK) and when Conventions is ablated (row 7, +.001 QWK), both of which lead to an increase in the Style scores. Second, while in two cases the trait scoring results appear to be identical to those of GAT-AES due to rounding (specifically, Prompt Adherence in row 6 and Narrativity in row 9), both are in fact slightly worse than GAT-AES. Third, the removal of any trait seems to have a considerable negative impact on Voice. Fourth, in row 2, when removing Overall, all trait scoring results are negatively impacted. The top three impacted traits are Conventions (-.033 points in QWK), Voice (-.024 points in QWK) and Sentence Fluency (-.021 points in QWK). In addition, in row 3, when removing Prompt Adherence, all trait scoring results are negatively impacted. The top three impacted traits are Voice (-.029 points in QWK), Word Choice (-.017 points in QWK), and Overall (-.012 points in QWK). Moreover, in row 8, when removing Word Choice, all trait scoring results are negatively impacted. The top three impacted traits are Voice (-.024 points in QWK), Conventions (-.017 points in QWK) and Overall (-.017 points in QWK).

These observations suggest that: Since Voice appears only in prompt 8, which has fewer than 800 essays, we speculate that GAT-AES may have overfitted Voice by relying on (potentially multi-hop) dependencies with other traits, rather than learning from the essay embeddings or the hand-crafted features. This may explain why Voice is the most impacted trait when any trait is removed. Ridley et al. (2021) observed that the prediction of the Overall score is highly influenced by Word Choice, Prompt Adherence, and Narrativity. Our results corroborate their finding in that removing Word Choice and Prompt Adherence from GAT-AES caused a significant drop in the QWK score of Overall.

G.2 Additional Results on Nodes and Edges

Next we conduct an additional ablation experiment as well as an augmentation experiment.

Correlation-based edges. Since we employ an all-pair configuration for trait nodes in GAT-AES, it is worth investigating whether alternative edge

	Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG (SD)
1	GAT-AES	**.771**	**.742**	.749	.687	.726	.694	**.686**	**.709**	**.692**	.699	**.649**	**.710** (±.011)
2	– w/o trait-trait edges	.754	.739	.750	.690	**.730**	.680	.653	.697	.669	.698	.593	.696 (±.010)
3	– w/ ArTS-style edges	.767	.734	.746	**.692**	.721	.688	.663	.699	.685	.704	.611	.701 (±.014)
4	– w/ correlation-based edges	.765	.741	.752	.680	.715	**.696**	.678	.702	.690	**.707**	.637	.706 (±.009)
5	– Adding prompt nodes	.764	**.742**	**.754**	.689	.723	.680	.660	.699	.683	.698	.613	.701 (±.012)

Table 15: Trait scoring results of additional ablation and augmentation experiments.

configurations for trait nodes are effective. In addition to the ArTS-style edges, another intuitive configuration is to connect only those trait nodes that exhibit a sufficiently high Pearson Correlation Coefficient. Specifically, we compute the Pearson Correlation Coefficient on the training set for each fold, and use the development set to select a threshold value to connect the trait nodes with the top-X correlation values, where X can be a value in {10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%}. The results are presented in row 4 of Table 15. As can be seen, while this approach performs significantly better than using no trait-trait edges and using ArTS-style edges, it still does not match the performance of GAT-AES. We speculate that this is because GAT-AES can leverage certain trait pairs that do not exhibit a sufficiently high Pearson Correlation Coefficient.

Adding prompt nodes. In GAT-AES, we follow prior work on multi-trait scoring (Do et al., 2024a,b) and do not include the writing prompt as input. The scoring of Prompt Adherence may have predominantly relied on hand-crafted features and/or patterns present in the essays themselves. Thus, it is worth investigating whether incorporating writing prompts into GAT-AES provides any benefit. We construct prompt embedding nodes using the same procedure applied to text embedding nodes, and connect each prompt embedding node to every trait node. The number of prompt embedding nodes is selected based on development set performance, using a range of {1, 2, 4} nodes. Results of this configuration are presented in row 5 of Table 15. As shown, while there is a minor 0.005%-point increase in the QWK score for Prompt Adherence, the average trait scoring performance decreases by 0.009% points. We speculate that adding prompt embeddings to the model may have helped the scoring of Prompt Adherence, but at the same time it introduced noise that negatively affected the scoring of other traits. More effective approaches may be needed to incorporate prompt information without compromising overall performance.

The features of setting affect the cyclist. The harsh turane proved a challenge when his water supply was low and all towns had been abandoned leaving him no refill. The heat was also a factor quickly creating him weary and tired. On the bike ride he had a glimmer of hope at an old water pump but quickly disappointed when the hot liquid tasted of battery acid. The setting set the conflict and helped achive the tone the author was looking for to affect the cyclist.

Table 16: A sample essay where GAT-AES overestimates for Language.

H Trait-Specific Error Analysis

In this section, we conduct an error analysis of GAT-AES on different traits.

H.1 Language

Language is a trait scored only in source-dependent essays (prompts 3–6). The most frequent failure case is our model assigning a score that is one point higher than the gold label, accounting for 65.33% of mispredictions. Table 16 provides a representative example. The essay contains numerous spelling and grammatical errors, such as "turane" (for "terrain"), "creating him weary" (instead of "making him weary"), "achive" (for "achieve"), and "quickly disappointed" (which should be "was quickly disappointed"). It also lacks varied grammatical structures (e.g., clauses). The gold score for Language is 1 ("Grammar and spelling show many errors. Vocabulary is limited and not very varied. Some words may be used in inappropriate places."), whereas the model predicted a 2 ("Grammar and spelling are good, with only some minor errors. Different kinds of grammatical structures may be used. The writing shows evidence of an adequate range of vocabulary."). We suspect the model overweights isolated strong vocabulary (e.g., "conflict", "tone", "glimmer of hope") while underestimating the cumulative impact of grammatical and spelling errors.

H.2 Style

Style is only scored in narrative essays (prompt 7). The model overestimates and underestimates

A time when I was patient was when I counted. I didn?t just count to ten, I counted a boatload of coins consisting of pennies, nickles, dimes, quarters, and more. It took me two hours to finish counting my money, it felt like an eternity. Counting bunch of coins won?t change the world, but it takes a lot of patience to accomplish. I still remeber the moment when I finished counting the money. I was proud of myself and l couldn?t think l could make it because l admit to be a very impatient person. Patience is important because nothing would ever be finished without it. I don?t think that I?ll be able to count my coins again, but like they say, ?Patience is a virtue.?

Table 17: A sample essay where GAT-AES overestimates for Style.

[Prompt]
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.

[Rubric for Narrativity]
Score 2: The response is somewhat interesting. Transitional and linking words are used in some places, but not everywhere.
Score 1: The response is very uninteresting and disjointed and is unable to deliver the content at all.
Score 0: The response is irrelevant / incorrect / incomplete.

[Example Essay #1]
The author concludes the story with this paragraph I think because since the girl failed her test the bud reminds her of her grandmothers long grey hair and I guess that gives her good luck so when the winter passes by and the spring time hits thats buds start to come back out and thats when she will take her test again. Thats why the author made this statement the last sentence of the story.

[Example Essay #2]
The author concludes the story w/this paragraph because she really isnt worried about when they take the test again, she?s only concerned w/the plants she picked out. Also, she knows that it will be a really long time before she takes the test again so that gives her more time to study and get more prepared and focused. The author also ends this article w/this paragraph because she wants to let the readers know that since she failed the @NUM1 time she will be back again to take it over and nothing will stop the author from failing.

Table 18: Examples for the error analysis of GAT-AES on Narrativity.

scores at similar rates (53% and 47% of failure cases, respectively). In 93% of these cases, its predictions differ from the gold labels by 1 point; in the remaining 7%, they differ by 2 points. Consider the example in Table 17: the gold label is 4 (sum of two annotators' scores) while the model predicts a 5. This implies that the model expects one annotator to assign a 2 ("Adequate command of language, including effective word choice and clear sentences, supports the writer's purpose and audience") and the other a 3 ("Command of language, including effective and compelling word choice and varied sentence structure, clearly supports the writer's purpose and audience."). We suspect the model struggles to distinguish fine-grained differences between rubric levels. While the essay effectively supports its purpose, its largely simple and direct sentence structure may have led human raters to assign a slightly lower score.

H.3 Narrativity

Narrativity is a trait that is present only in the (source-dependent) essays written for prompts 3, 4, 5, and 6. The most common failure case for Narrativity is related to our model's lack of ability to verify factual or irrelevant information. In Table 18, we show a writing prompt, the rubric for Narrativity, and two example essays. The model predicts a 2 for the example essays while their gold labels are 0. These responses are incorrect because the purpose of the last paragraph is to show that the author uses Saeng's vow to take the driver's test again as a symbol of personal growth and resilience, and they should be assigned a score of 0 rather than a 2. However, since there is some basic transitional language (e.g., "That's why", "Also") in these responses, the model might be tricked into believing that they should be assigned a 2.

H.4 Word Choice, Sentence Fluency, and Conventions

The traits Word Choice, Sentence Fluency and Conventions appear in prompts 1–2 (persuasive essays) and 8 (narrative essays). The most common failure case for Word Choice lies in our model's lack of sensitivity to word appropriateness and precision. Consider the writing prompt, rubrics and example essay in Table 19. The gold Word Choice score is 3 while our model predicts a 5. In the essay, it seems like it uses many great words (e.g., "unwind", "extrodinary"). However, sometimes these usages are not precise. For example,

[Prompt]
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 views 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.
[Rubric for Word Choice]
Score 5: Words convey the intended message in an interesting, precise, and natural way appropriate to audience and purpose. The writer employs a broad range of words which have been carefully chosen and thoughtfully placed for impact. The writing is characterized by • accurate, specific words; word choices energize the writing. • fresh, vivid expression; slang, if used, seems purposeful and is effective. • vocabulary that may be striking and varied, but that is natural and not overdone. • ordinary words used in an unusual way. • words that evoke clear images; figurative language may be used.
Score 3: Language lacks precision and variety, or may be inappropriate to audience and purpose in places. The writer does not employ a variety of words, producing a sort of “generic” paper filled with familiar words and phrases. The writing is characterized by • words that work, but that rarely capture the reader’s interest. • expression that seems mundane and general; slang, if used, does not seem purposeful and is not effective. • attempts at colorful language that seem overdone or forced. • words that are accurate for the most part, although misused words may occasionally appear; technical language or jargon may be overused or inappropriately used. • reliance on clichés and overused expressions. • text that is too short to demonstrate variety.
[Rubric for Sentence Fluency]
Score 5: The writing has an easy flow and rhythm. Sentences are carefully crafted, with strong and varied structure that makes expressive oral reading easy and enjoyable. The writing is characterized by • a natural, fluent sound; it glides along with one sentence flowing into the next. • variation in sentence structure, length, and beginnings that add interest to the text. • sentence structure that enhances meaning. • control over sentence structure; fragments, if used at all, work well. • stylistic control; dialogue, if used, sounds natural.
Score 3: The writing tends to be mechanical rather than fluid. Occasional awkward constructions may force the reader to slow down or reread. The writing is characterized by • some passages that invite fluid oral reading; however, others do not. • some variety in sentence structure, length, and beginnings, although the writer falls into repetitive sentence patterns. • good control over simple sentence structures, but little control over more complex sentences; fragments, if present, may not be effective. • sentences which, although functional, lack energy. • lapses in stylistic control; dialogue, if used, may sound stilted or unnatural. • text that is too short to demonstrate variety and control.
[Rubric for Conventions]
Score 5: The writing demonstrates strong control of standard writing conventions (e.g., punctuation, spelling, capitalization, grammar and usage) and uses them effectively to enhance communication. Errors are few and minor. Conventions support readability. The writing is characterized by • strong control of conventions. • effective use of punctuation that guides the reader through the text. • correct spelling, even of more difficult words. • correct capitalization; errors, if any, are minor. • correct grammar and usage that contribute to clarity and style. • skill in using a wide range of conventions in a sufficiently long and complex piece. • little need for editing.
Score 3: The writing demonstrates limited control of standard writing conventions (e.g., punctuation, spelling, capitalization, grammar and usage). Errors begin to impede readability. The writing is characterized by • some control over basic conventions; the text may be too simple or too short to reveal mastery. • end-of-sentence punctuation that is usually correct; however, internal punctuation contains frequent errors. • spelling errors that distract the reader; misspelling of common words occurs. • capitalization errors. • errors in grammar and usage that do not block meaning but do distract the reader. • significant need for editing.
[Example Essay]
Dear @PERSON1 editor, I think that the computers have both positive and negative effects, but more positive. For example, when people get home from work or school, they immedlatly go on the computer, to check email, do work, or to play games. I think that it’s good to connect with friends and learn about the far away nature and people that we cant see, but too much of it will be bad. I know it will be bad because it can cause a lack of exercise and the computer can cut you off from your family and friends if you’re on it too much, I think it’s good to go on the computer for a limited amount of time each day, but also spend time outside with family and friends. The computer is good because of all the extrodinary things it offers and the things you can learn. I myself like to play video games on it like bubble spinner but at the same time I’ll be on facebook, facebook lets me connect with friends and family while bubble spinner helps with hand-eye coordination, when projects came along I usually do all my research there, picking up information on each site. I also do all the writing on it too so it wont look messy but instead, clean and neat. The computer also offers video of pretty much anything on t.v. and otherwise, like youtube or google videos. The computer lets me see nature, talk to friends and family, and lets me see things, like videos, that twill never ever see in books. These show that the computer has a bigger positive impact than negative and we should take advantage of the new technology that we have and not let it go to waste. The computer is a nice outlit to come home and let yourself unwind to. People say that you don’t get enough exercise but we have gym and recess at school and we become exhausted at work, which is enough exercise for the day. Computers are a nice device to have and are by far, used daily by almost everyone. So it can’t be all that bad by depriving you from nature and family because it lets you see it anyway. The effects it has on people are good because of the knowledge and communication it brings. I hope you see my point of view.

Table 19: Example essay for the error analysis of GAT-AES on WORD CHOICE, SENTENCE FLUENCY and CONVENTIONS.

[Prompt]
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.
[Rubric for Content]
Score 3 : The response answers the question asked of it. Sufficient evidence from the story is used to support the points that the writer makes.
Score 2 : The response addresses some of the points. Evidence from the story supporting those points are present.
Score 1 : The response may lack information / evidence showing a lack of understanding of the text.
Score 0 : The response is irrelevant / incorrect / incomplete.
[Rubric for Prompt Adherence]
Score 3 : The response shows an excellent understanding of the meaning of the text and question, and stays on topic.
Score 2 : The response shows a good understanding of the meaning of the text and question, and occasionally wanders off topic.
Score 1 : The response shows a misreading of the text or question, or consistently wanders off topic.
Score 0 : The response is irrelevant / incorrect / incomplete.
[Example Essay]
The reason the author of the story ?Winter hibiscus? ends it the way she does is so that we know we can never give up. She fails her driving test the first time and me being less than a year from taking the test myself and no matter how many time I fail the test I will never give up. We need to preserve through every problem we go through in life. If we don?t and we just give up we will never get anything done in life. She vows in the spring and when the hibiscus is budding she will take the test again. I hope I pass my driving test the first time just like I did the learner?s permit test. Also you can?t give up on anything at all even something small like working if you stop working you will loose your job. Never give up.

Table 20: Example essay for the error analysis of GAT-AES on CONTENT and PROMPT ADHERENCE.

in the sentence “The computer is a nice outlit to come home and let yourself unwind to”, “Outlit” is likely a misspelling of “outlet”, but even so, “outlet to unwind to” is not idiomatic. Thus, our model may be overly rewarding the presence of semantically rich words without penalizing incorrect or awkward usage. Due to our model’s strong ability in capturing trait interdependencies as well as the strong correlations between WORD CHOICE, SENTENCE FLUENCY and CONVENTIONS, our model assigns a 5 to all these traits while the gold scores are all 3.

H.5 CONTENT and PROMPT ADHERENCE

CONTENT and PROMPT ADHERENCE are the only two traits in the ASAP++ dataset that are related to the content of essays. One failure case of our model is shown in Table 20, where the gold scores for CONTENT and PROMPT ADHERENCE are both 0, but our model predicts scores of 3 and 2, respectively. We speculate that this discrepancy can be attributed to the fact that the model does not have access to the source document or the writing prompt. As a result, it is misled by the seemingly on-topic essay and fails to accurately assess CONTENT and PROMPT ADHERENCE.

H.6 ORGANIZATION

The trait ORGANIZATION appears in prompts 1, 2, 7, and 8. Interestingly, our model has more overestimations for ORGANIZATION in prompts 1, 2 and 8, but has more underestimations in prompt 7. Consider the writing prompt, rubric and example given in Table 21. The gold score is 2 (two annotators gave a score of 1) while the prediction is 5. The example essay is an attempt to narrate events in sequence (trip, babysitting, fights, walk), but transitions are choppy and connections between ideas are weak. The essay contains run-on sentences and grammar issues that obscure logical flow (e.g., “I didn?t complain til I got home til my mom”). While some organization exists, it is insufficient to meet even the rubric’s Score 2 standard. GAT-AES likely overestimates because it detects a narrative thread and some chronological as evidence of logical sequencing and organization. However, the rubric emphasizes clear and strong connections between ideas, which this essay lack due to weak transitions and grammar issues.

H.7 VOICE

The trait VOICE, which appears in prompt 8 only, evaluates how clearly the writer’s personality, tone, and point of view come through in the writing. It reflects whether the writer sounds genuinely engaged with the topic and how well they connect with the intended audience. GAT-AES has a tendency to overestimate an essay’s VOICE score. Consider the writing prompt, rubric, and example essay in Table 22. The gold score is 6 (two annotators gave

[Prompt]

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.

[Rubric for Organization]

Score 3: Organization and connections between ideas and/or events are clear and logically sequenced.
Score 2: Organization and connections between ideas and/or events are logically sequenced.
Score 1: Organization and connections between ideas and/or events are weak.
Score 0: No organization evident.

[Example Essay]

One time I was pacient was when I went to @LOCATION1 with my @CAPS1 + cousin. I had to baby sit my cousin @CAPS2 while his room pack up cause they were making back to @LOCATION2. My cousin can be a handful sometimes and always gets his way. But I was pacient and waited til the trip was away. I didn?t complain til I got home til my mom. Me and my cousin got in lots of fights on that vacation. But I still love him . I had to be very pacient with my cousin and even maybe do what he wanted but I knew his mom was very proud of me. I remember when we went for a walk to a pond to find alligators. But then @CAPS2 wanted to go to the pack.I knew I had to give in (considering that hes @NUM1 and immature). But I just went with the flow and we went to the pack. That was a time that I had to be pacient.

Table 21: Example essay for the error analysis of GAT-AES on ORGANIZATION.

3 points) while the predicted score is 9. Our model likely overestimated the score because it was influenced by the liveliness and energy in the narrative, such as the playful scenes ("WACK! The bunney conected...") and moments of humor ("@CAPS66 from planet meanie are we?"). These features mimic the surface characteristics of a high-scoring voice, but upon a closer examination, the voice is inconsistent and often mechanical, as described under the Score 3 criteria. The narrative lacks a sustained and appropriate voice, frequently becoming overly casual or erratic, and failing to maintain a clear sense of audience awareness, which are key traits expected at Score 5 or 6. Thus, while the essay is energetic, the commitment and consistency needed for a higher score are not sufficiently developed. We speculate that the overestimation of our model on VOICE is due to it capturing only surface-level features for VOICE.

[Prompt]
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.

[Rubric for Voice]
Score 5: The writer has chosen a voice appropriate for the topic, purpose, and audience. The writer demonstrates commitment to the topic, and there is a sense of “writing to be read.” The writing is expressive, engaging, or sincere. The writing is characterized by an appropriate level of closeness to or distance from the audience (e.g., a narrative should have a strong personal voice, while an expository piece may require extensive use of outside resources and a more academic voice; nevertheless, both should be engaging, lively, or interesting. Technical writing may require greater distance.). • a strong sense of audience; the writer seems to be aware of the reader and of how to communicate the message most effectively. The reader may discern the writer behind the words and feel a sense of interaction. • a sense that the topic has come to life; when appropriate, the writing may show originality, liveliness, honesty, conviction, excitement, humor, or suspense.
Score 4: A voice is present. The writer seems committed to the topic, and there may be a sense of “writing to be read.” In places, the writing is expressive, engaging, or sincere. The writing is characterized by • a suitable level of closeness to or distance from the audience. • a sense of audience; the writer seems to be aware of the reader but has not consistently employed an appropriate voice. The reader may glimpse the writer behind the words and feel a sense of interaction in places. • liveliness, sincerity, or humor when appropriate; however, at times the writing may be either inappropriately casual or personal, or inappropriately formal and stiff.
Score 3: The writer’s commitment to the topic seems inconsistent. A sense of the writer may emerge at times; however, the voice is either inappropriately personal or inappropriately impersonal. The writing is characterized by • a limited sense of audience; the writer’s awareness of the reader is unclear. • an occasional sense of the writer behind the words; however, the voice may shift or disappear a line or two later and the writing become somewhat mechanical. • a limited ability to shift to a more objective voice when necessary. • text that is too short to demonstrate a consistent and appropriate voice.

[Example Essay]
I could just laugh to death. someone said. I sacned the area. “@CAPS37?” I said. “@CAPS1 here son.” I looked @CAPS1 the stairs leading to @CAPS2 avenue. There was the sorce of the voice i had herd. @CAPS35 was only an old man probualy homeless by the looks of him. He had worn shoes, dirty blue jeans, a riped shirt under a worn jacket. “@CAPS3, @CAPS31 @CAPS33 a little lost”The old man said. “@CAPS4, I’m @CAPS73 lost” said @PERSON1. His pocket started to vibrate. @PERSON1 reached into his pocket an pulled out his cell. He fliped @CAPS35 open and read: @CAPS31 going to be here soon? Tommey typed: ya bro just going onto @CAPS2 ave now. Message sent. @PERSON1 started to climb the steps. “@CAPS5 @CAPS31 later old man” said @PERSON1 as he dashed of to meet his friends. @PERSON1 saw his friends in site @CAPS62 were all standing outside of the @CAPS6 @CAPS7 bakery. “@CAPS3,guys @CAPS33 @CAPS35’s @PERSON1!” @CAPS34 @PERSON2. @PERSON2 ran at @LOCATION1. @PERSON1 scuped her @CAPS1 and spun her around one before puting her @CAPS66 @CAPS19. @CAPS62 proceeded to walk over to the group. “@CAPS37 took @CAPS31 so long dude” said @CAPS9 @CAPS10. “@CAPS11, @CAPS63 the short cut i took turned out to be, @CAPS73 to short” @PERSON1 repiled. “@CAPS12 don’t mind him he’s just in a bad mood” said @CAPS13. “@CAPS14” replied @CAPS9 @CAPS10. “@CAPS3, guys can we go into the @CAPS18 @CAPS19’s toy store!!” inqured @PERSON2. “@CAPS20 that’s fine” replied @CAPS13 @CAPS22. The group proceeded @CAPS19 @CAPS2 avenue @CAPS24 turned right onto @ORGANIZATION1 street, @CAPS24 another right onto @CAPS25 avenue. The store was located halfway @CAPS19 in the middle of other boutiques. “@CAPS26 were here” said @PERSON2 as she spun around around from excitement. “@CAPS24 let’s go in” said @CAPS9 @CAPS10. “@CAPS30 old greblin!” said @PERSON2. “shh mu ha ha ha” @CAPS31 got told said @CAPS13. “@CAPS32” he grumbled. Once inside the store @CAPS35 was like somone turned @CAPS9 the lights @CAPS9 bright. @CAPS73 only that but the store was filled from top to bottom with @CAPS38 but toys. The toys ranged from small to large to huge. There were all sorts of diffrent types. “@CAPS33” @CAPS34 @PERSON2. “@CAPS35’s a @CAPS43 bunny!”“@CAPS35’ll smother @CAPS31 in your sleep twerp” “@CAPS37’s eating @CAPS31?” said @PERSON1. “@CAPS38” grumbbled @CAPS9 @CAPS10. WACK! The bunney conected with @CAPS9 @CAPS10’s face. “@CAPS43 meenie”“@CAPS44, that does @CAPS35” he took a lion off the shelf and tossed @CAPS35 at @LOCATION2. “@CAPS3, don’t be mean to her!” said @CAPS13 as she picted @CAPS1 a monester an flung @CAPS35 at @CAPS9 @CAPS10. @CAPS35 missed and hit @PERSON1. @PERSON1 flung @CAPS35 @CAPS66 at @CAPS9 @CAPS10. “That’s @CAPS35” I @CAPS9 @CAPS10 challege @CAPS31. SMAK. “@CAPS52” @CAPS53 this. @CAPS53 that. @CAPS54 @CAPS54 @CAPS31 scum. @CAPS31 can’t get me. WAM! Ha ha i’m ausome. TWACK! “@CAPS55” @CAPS34 some one. @CAPS62 @CAPS56.“@CAPS31 kids are in @CAPS43 trouble!” the owner was yelling. “@CAPS31 better put all these toys @CAPS66 where @CAPS62 were” he warned.“@CAPS59! I think we got a little caried away” said @CAPS13. “A little!! My toy store is a mess!now get cleaning” “@CAPS61” @CAPS62 all sighed. “@CAPS63 that was cool” siad @CAPS9 @CAPS10. “@CAPS66 from planet meanie are we?” teased @PERSON1. “@CAPS67 i @CAPS20 am” replied @CAPS9 @CAPS10. “ya, that was fun” said @CAPS13. “@CAPS71 @CAPS35 was!!” replied @PERSON2. “We all looked pretty stupid hu” said @CAPS13 blushing. “ @CAPS67, but how many people can say @CAPS62 had a toy fight in a toy store” replied @PERSON1. “@CAPS73 many i guess!”“@CAPS74 talking more cleaning, clean, clean!” @CAPS34 the store owner.“yes, sir” @PERSON2 soluted the owner.“@CAPS76 @CAPS31...”“@CAPS77” The group of kids laughed.

Table 22: Example essay for the error analysis of GAT-AES on VOICE.