Autoregressive Score Generation for Multi-trait Essay Scoring

Heejin Do[1] **Yunsu Kim**[2] **Gary Geunbae Lee**[1,3]
[1]Graduate School of AI, POSTECH, Republic of Korea
[2]aiXplain Inc. Los Gatos, CA, USA
[3]Department of CSE, POSTECH, Republic of Korea
{heejindo, gblee}@postech.ac.kr yunsu.kim@aixplain.com

Abstract

Recently, encoder-only pre-trained models such as BERT have been successfully applied in automated essay scoring (AES) to predict a single overall score. However, studies have yet to explore these models in multi-trait AES, possibly due to the inefficiency of replicating BERT-based models for each trait. Breaking away from the existing sole use of *encoder*, we propose an autoregressive prediction of multi-trait scores (ArTS), incorporating a *decoding* process by leveraging the pre-trained T5. Unlike prior regression or classification methods, we redefine AES as a score-generation task, allowing a single model to predict multiple scores. During decoding, the subsequent trait prediction can benefit by conditioning on the preceding trait scores. Experimental results proved the efficacy of ArTS, showing over 5% average improvements in both prompts and traits.

1 Introduction

Automated essay scoring (AES) is a prominent task to efficiently assess large volumes of essays. Currently, there is a growing trend in holistic AES to use pre-trained BERT-based models, showing promising results (Rodriguez et al., 2019; Mayfield and Black, 2020; Beseiso and Alzahrani, 2020; Yang et al., 2020; Wang et al., 2022). However, these models have yet to be explored in multi-trait AES, which evaluates essays on diverse rubrics, possibly due to the inefficiency of duplicating encoders for different traits.

Existing multi-trait scoring approaches (Mathias and Bhattacharyya, 2020; Ridley et al., 2021; Kumar et al., 2022; Do et al., 2023) typically adopted holistic scoring models (Taghipour and Ng, 2016; Dong et al., 2017), adding multiple linear layers or separate trait-specific layers for different traits. However, achieving multi-trait AES as a holistic method overlooks the trait dependencies, and constructing separate trait-specific modules is resource-inefficient, leading to inferior qualities in data-scarce traits. These limitations highlight the need for optimized multi-trait strategies.

In this paper, we propose autoregressive multi-trait scoring of essays (ArTS), which incorporates the decoding process by leveraging a pre-trained language model, T5 (Raffel et al., 2020). Moving beyond the conventional sole reliance on the encoder, we introduce a novel text-to-text AES framework. Unlike existing regression or classification approaches to output a separate numeric value, we aim at precise sequence generation by considering multi-trait scores as an entire sequence; thus, a single model can yield multi-score predictions. ArTS employs causal self-attention to capture the intrinsic relations of the traits by sequentially predicting text-transformed trait scores. The autoregressive generation allows the subsequent trait prediction to benefit from referencing preceding trait scores.

ArTS remarkably outperformed the baseline model on the ASAP and ASAP++ (Mathias and Bhattacharyya, 2018) datasets. Ablation studies and additional discussions of trait order further verify our method. Furthermore, ArTS achieved training efficiency by using a single model to generate multiple predictions across all prompts, avoiding the duplication of the same modules. Codes and datasets are available on Github[1].

2 Related Work

Early studies of AES mainly focused on holistic essay scoring that only predicts the overall score and already achieved high assessment performance (Dong and Zhang, 2016; Taghipour and Ng, 2016; Dong et al., 2017; Uto et al., 2020; Wang et al., 2022). In contrast, multi-trait scoring has been studied for detailed assessments lately, yet showing far-lagged quality. Holistic scoring structures are typically employed either for a trait-shared model followed by multiple linear layers (Hussein et al.,

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

Findings of the Association for Computational Linguistics: EACL 2024, pages 1659–1666

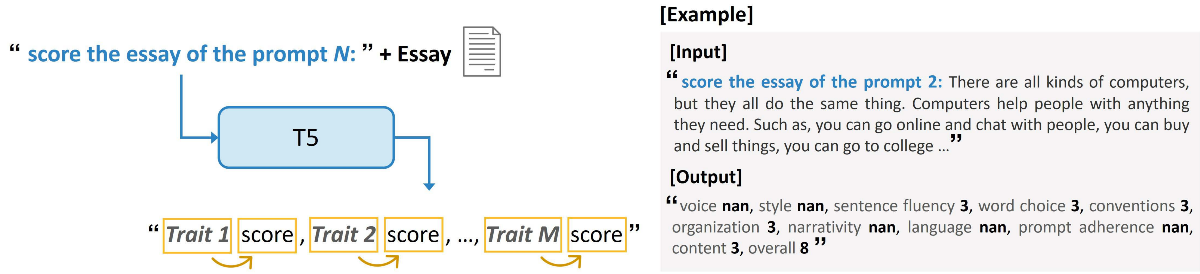

Figure 1: Proposed autoregressive multi-trait essay scoring by the fine-tuning of the T5. The example is an essay written for prompt 1, which has labeled scores for six traits. Unlabeled trait scores in the prompt are set as *nan*.

2020) or for multiple trait-specific layers (Mathias and Bhattacharyya, 2020; Ridley et al., 2021; Kumar et al., 2022; He et al., 2022; Do et al., 2023). In particular, Kumar et al. (2022) designed auxiliary trait-specific layers to assist primary trait scoring, achieving competitive results. However, to predict m trait scores, m different models containing m duplicated trait-specific layers are required, which is resource-inefficient. Moreover, the notable quality gap between trait scoring and holistic scoring highlights the need for advanced multi-trait AES.

Transformer-based pre-trained models such as BERT (Devlin et al., 2018) and GPT (Brown et al., 2020) excel across various tasks by capturing rich semantic and syntactic information via training on large-scale corpora. Recently, some studies have applied them to *holistic* AES (Rodriguez et al., 2019; Mayfield and Black, 2020; Beseiso and Alzahrani, 2020; Yang et al., 2020; Wang et al., 2022), contributing to a notable leap in the *holistic* scoring. However, they only employ encoder-only models to predict a numeric value without considering the decoder. Moreover, those BERT-based models have not been extended to multi-trait scoring, possibly due to the efficiency concerns (e.g., predicting an *Overall* score with a BERT-based model of 110M parameters took 113 hours (Kumar et al., 2022); accordingly, predicting *m* traits would require *m* times the parameters and the time). In contrast, we leverage the potential capacities of autoregressive decoding to efficiently score multiple traits with a single model, suggesting a new perspective to address AES as a text generation task instead of a classification or regression.

3 Autoregressive Essay Multi-trait Scoring (ArTS)

To predict multiple trait scores in an auto-regressive manner, we fine-tune the pre-trained encoder-decoder language model, T5. Specifically, we treat AES as a generation task to predict a single sequential text rather than multiple numeric values for traits. Subsequently, we extract each trait score from the generated text comprising the predicted trait scores along with trait names (Figure 1).

Prompt	# Essays	Traits
1	1785	Over, Content, WC, Org, SF, Conv
2	1800	Over, Content, WC, Org, SF, Conv
3	1726	Over, Content, PA, Nar, Lang
4	1772	Over, Content, PA, Nar, Lang
5	1805	Over, Content, PA, Nar, Lang
6	1800	Over, Content, PA, Nar, Lang
7	1569	Over, Content, Org, Conv, Style
8	723	Over, Content, WC, Org, SF, Conv, Voice

Table 1: Composition of the ASAP/ASAP++ combined dataset. The prompt is an instruction that defines the writing theme. Over: *Overall*, WC: *Word Choice*, Org: *Organization*, SF: *Sentence Fluency*, Conv: *Conventions*, PA: *Prompt Adherence*, Nar: *Narrativity*, Lang: *Language*.

3.1 Fine-tuning T5

T5 has achieved competitive performance in numerous natural-language processing tasks by handling various tasks using a text-to-text approach. One of the trained tasks of T5 is semantic textual similarity (STS), which is a regression task predicting a float-type similarity value between two texts. Given that T5 has been pre-trained to output a text-formed numeric value for the STS, we assume that fine-tuning the model to output an essay score will yield precise prediction. Instead of individually predicting trait scores with multiple models, our goal is to generate all trait scores with a single autoregressive prediction, thus achieving both time and resource efficiency. Using one integrated model can avoid unnecessary duplication of the same distinct models.

Particularly, we add the prefix *"score the essay of the prompt N:"* in front of each essay as the

	Traits (←)											
Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑ (SD↓)
HISK	0.718	0.679	0.697	0.605	0.659	0.610	0.527	0.579	0.553	0.609	0.489	0.611 (-)
STL-LSTM	0.750	0.707	0.731	0.640	0.699	0.649	0.605	0.621	0.612	0.659	0.544	0.656 (-)
MTL-BiLSTM	**0.764**	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	**0.582**	0.638 (-)
ArTS (Ours)	0.754	**0.730**	**0.751**	**0.698**	**0.725**	**0.672**	**0.668**	**0.679**	**0.678**	**0.721**	0.570	**0.695** (±0.018)
ArTS-w/o *Pr*	0.690	0.723	0.751	0.691	0.725	0.655	0.656	0.644	0.648	0.673	0.530	0.671 (±0.033)

Table 2: Average QWK scores across all prompts for each **trait**. The left arrow (←) indicates the direction of the trait prediction. *SD* is the five-fold averaged standard deviation. ArTS-w/o *Pr* (shown in gray) represents the ablation results without the prompt indication. Further, **bold** text denotes the highest value, excluding ablation results.

	Prompts								
Model	1	2	3	4	5	6	7	8	AVG↑ (SD↓)
HISK	0.674	0.586	0.651	0.681	0.693	0.709	0.641	0.516	0.644 (-)
STL-LSTM	0.690	0.622	0.663	0.729	0.719	0.753	0.704	0.592	0.684 (-)
MTL-BiLSTM	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665 (-)
ArTS (Ours)	**0.708**	**0.706**	**0.704**	**0.767**	**0.723**	**0.776**	**0.749**	**0.603**	**0.717** (±0.025)
ArTS-w/o *Pr*	0.709	0.645	0.703	0.769	0.679	0.769	0.722	0.566	0.695 (±0.036)

Table 3: Average QWK scores across all traits for each **prompt**.

input and concatenate trait name and trait score sets sequentially from the least to the most data labels with a comma (,) separation (Figure 1). We hypothesize that providing the prompt number, *N*, allows more accurate guidance. Note that traits not labeled in the corresponding prompt are trained to predict *nan* values. Including *nan* values might allow the model to generate a consistent output form regardless of the prompt, leading to more reliable predictions. In particular, the model predicts traits in the following order: *Voice*, *Style*, *SF*, *WC*, *Conv*, *Org*, *Nar*, *Lang*, *PA*, *Content*, and *Overall* (Table 1). By predicting peripheral trait scores first, which are assessed in fewer prompts, and more comprehensive trait scores later, which are rated in more prompts, we reflect the actual scoring process. For example, the *Overall* score is labeled in all prompts and highly influenced by other traits, whereas the *Voice* score is only evaluated in prompt 8 (Table 1) and is relatively independent of other traits. The causal self-attention of the transformer decoder enables subsequent trait-scoring tasks to attend to prior predicted trait scores; thus, the later order of dependent and general traits is natural.

3.2 Score extraction

With the fine-tuned model, we predict and generate a text for each essay containing predicted multiple trait scores along with the trait names. Then, we extract all trait scores keyed by their name. Multiple trait scores are obtained with a single model at one inference time, eliminating the inconvenience of multiple-model training and inference. For accurate measurement, we exclude all predictions of traits whose ground truth is a *nan* value.

4 Experiment

Datasets and settings For the main experiment, we employ the widely used ASAP[2] and ASAP++[3] (Mathias and Bhattacharyya, 2018) datasets comprising English essay sets for eight prompts written by American 7–10-grade high-school students. The *Overall* score is available for all essays in the ASAP dataset; however, trait scores are only labeled for essays of prompts 7 and 8. Therefore, the ASAP++ dataset providing rated trait scores for all prompts is jointly used (Table 1). In addition, we experiment on the Feedback Prize[4] data of argumentative essays written by American 6–12-grade students. It has six labeled trait scores without prompt division: *Cohesion*, *Syntax*, *Vocabulary*, *Phraseology*, *Grammar*, and *Conventions*.

We utilize the T5-Base (Raffel et al., 2020) model, which is pre-trained on the Colossal Clean Crawled Corpus. For fine-tuning, we employ Seq2SeqTrainer by setting evaluation steps as 5000, early stopping patience as 2, batch size as 4, and total epoch as 15. `A100-SMX4-8` GPU is used.

Evaluation and validation For evaluation, we use the quadratic weighted kappa (QWK) (Cohen, 1968), the official metric of the dataset. QWK is well-known for effectively capturing the distance between human-rated and model-predicted scores. We use five-fold cross-validation with the same

[2]https://www.kaggle.com/c/asap-aes
[3]https://lwsam.github.io/ASAP++/lrec2018.html
[4]https://www.kaggle.com/competitions/feedback-prize-english-language-learning

	Traits											
Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑ (SD↓)
ArTS (←)	**0.754**	**0.730**	0.751	**0.698**	**0.725**	**0.672**	**0.668**	**0.679**	**0.678**	0.721	**0.570**	**0.695** (±0.018)
ArTS-*rev* (→)	0.739	0.724	0.749	0.687	0.718	0.667	0.658	0.660	0.666	0.711	0.562	0.686 (±0.021)
ArTS-*ind*	0.723	0.717	**0.752**	0.695	0.713	0.649	0.659	0.662	0.675	**0.722**	0.548	0.683 (±0.053)

Table 4: Comparison results averaged by traits. ArTS-*rev* (→) predicts traits in reverse order, and 11 different ArTS-*ind* models predict each trait individually. The left (←) and right (→) arrows denote the direction of prediction.

	Traits (→)						
Model	Conv	Gram	Phr	Voc	Syn	Coh	AVG
MTL*	0.527	0.484	0.505	0.519	0.507	0.462	0.501
ArTS	**0.659**	**0.659**	**0.639**	**0.594**	**0.628**	**0.590**	**0.628**

Table 5: Experiments with the Feedback Prize dataset. Each value is the five-fold average QWK score (Conv: *Conventions*, Gram: *Grammar*, Phr: *Phraseology*, Voc: *Vocabulary*, Syn: *Syntax*, Coh: *Cohesion*).

split as that of Taghipour and Ng (2016), as in the baseline multi-task learning (MTL) (Kumar et al., 2022), reporting five-fold averaged results. We short-list two models based on the validation loss and select the final model with the best validation result. As suggested by Taghipour and Ng (2016), we calculate QWK separately for each prompt to avoid excessively high scores when using the whole set (e.g., 0.99 QWK for *Overall* with ArTS), providing both prompt- and trait-wise averaged results.

5 Results

Our model is primarily compared with the baseline MTL-BiLSTM model (Kumar et al., 2022), multi-task learning where auxiliary multi-trait scoring tasks aid holistic scoring (Table 2). In addition, we compare our model to the HISK and STL-LSTM models, which were mainly compared to MTL. HISK is a histogram intersection string kernel with a support vector regressor (Cozma et al., 2018), and STL-LSTM is LSTM-CNN-based model (Dong et al., 2017); both models are individually applied for each trait scoring. Trait-scoring results are only presented with a graph (Kumar et al., 2022); thus, we contacted the authors and obtained exact values.

Main results ArTS exhibits a significantly improved performance, showing over 5% average improvements in both prompt- and trait-wise results (Table 2, 3). A slight decrease in *Overall* trait could be attributed to our model's general focus on all traits, as opposed to baseline models designed primarily for overall scoring. For syntactic traits (*Org*, *Conv*, *WC*, *SF*), which evaluate the structure or grammatical aspects of essays, the performance increases by an absolute 5.7–10%. This highlights that leveraging ArTS facilitates capturing essays' syntactic aspects, even with few datasets. Notably, the *Conv* trait, the most inferior trait on the baseline, shows the greatest improvement with ArTS. Remarkably enhanced semantic traits (*Content*, *PA*, *Lang*, *Nar*) further imply that our autoregressive approach adeptly encapsulates the contextual facets of writing. Further, *Style* and *Voice* traits with severely lacking (1569, 723) samples show approximately 9% advancement and a slight reduction, respectively, implying the overcoming of low-resource settings.

Prompt number guidance We conducted an ablation study to investigate the effect of providing a prompt number in training. ArTS-w/o *Pr* (Table 2, 3) is the model results fine-tuned with the prefix *"score the essay:"* without the prompt number. The results indicate that clearly guiding the model with the essay's prompt number noticeably assists the scoring.

Trait prediction order To investigate the effect of the trait prediction sequence, we fine-tune T5 with the reverse order (ArTS-*rev*). Improved results when predicting general traits later in the sequence than the reverse reflect the real-world scoring, where comprehensive trait scores often rely on the other traits (Lee et al., 2010). In addition, we compare ArTS with the individual trait models (Table 4). ArTS-*ind* is the fine-tuned model to output a single trait name and score (e.g., Content 3). The results indicate that although the individual predictions highly outperform the baseline MTL model, our integrated method performs better on most traits. A single ArTS model outperforming 11 individual ArTS-*ind* models is remarkable, highlighting our model's resource efficiency along with competitive performance.

Feedback Prize dataset To provide supplementary evaluation beyond traditional benchmarks and demonstrate generalizability across diverse datasets, we employ ArTS using the Feedback

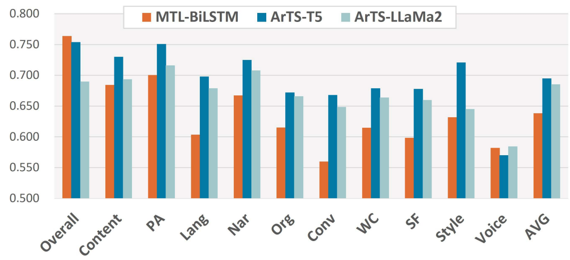

Figure 2: Results of ArTS with Llama2-13B and comparison with the baseline and ArTS with T5 models.

Prize dataset. The MTL model has not experimented with the dataset; accordingly, MTL* in Table 5 is our implementation results of the MTL with each trait scoring as the primary task and other traits as auxiliary tasks. Note that prompts are not differentiated, and all essays have identical traits in this dataset; therefore, the prompt number is excluded from the input, as in the ablation study. ArTS exhibits significantly improved QWK scores across all traits, demonstrating the broader applicability of ArTS (Table 5). A greater improvement compared to the ASAP experiments further indicates that ArTS can yield a more substantial impact in the same trait composition settings compared to the multi-prompt and different trait scenarios. Furthermore, our single-model approach outperformed MTL* in predicting all six traits simultaneously, showcasing the efficiency of our model without the need for specialized auxiliary modules for each trait scoring.

Decoder-only LLM To examine whether the decoder-only pre-trained language model alone could perform the function of autoregressive score generation, we fine-tuned the Llama2-13B model with our method (Figure 2). Noticeably, ArTS-Llama2 remarkably outperforms the baseline model for all the traits except for the *Overall* score. However, ArTS-T5 still performs better, suggesting the joint use of the encoder and decoder for AES.

Comparison with BERT-based models Recent studies in holistic AES have employed pre-trained BERT-based models and demonstrated promising scoring performances (Yang et al., 2020; Cao et al., 2020; Uto et al., 2020). However, they have not been utilized in multi-trait scoring, which confines our performance comparison solely to the *Overall* score. Their QWK results for the *Overall* scoring range from 0.790 to 0.805 (Kumar et al., 2022), surpassing our 0.754. Our result aligns with the MTL model, exhibiting lower *Overall* performance than BERT-based models but demonstrating training efficiency. Nevertheless, unlike MTL, we possess the advantage of simplicity and effectiveness by not requiring separate models for each prompt or trait and outperforming MTL in the other nine traits.

Regarding training efficiency, using BERT-based models that predict a single numeric score for multi-trait predictions would require replicating multiple models, making it resource-inefficient. For example, predicting 11 traits with a BERT model of 110M parameters would involve a substantial 110M $\times$ 11 parameters, along with increased training time. This is a probable reason for the absence of a BERT-based system for multi-trait scoring tasks. In contrast, our approach enables multi-trait predictions across all prompts with a single T5-base model of 220M parameters, taking 16.3 hours for training time. When using T5-small of 60M parameters, which also highly outperforms the baseline model (Appendix A), it took about 2.8 hours for training. Unlike existing methods, which necessitate multiple trait-specific or prompt-specific models, the ArTS with a single model demonstrates both time and resource efficiency.

6 Conclusion

In this paper, we introduce an autoregressive multi-trait scoring of essays that leverages the capacity of the pre-trained language model, T5. Our model exhibits remarkably improved results, demonstrating its ability to overcome far-lagging multi-trait-scoring performances. Furthermore, our approach allows a single model to make multi-trait score predictions across all prompts, avoiding the use of redundant modules and promoting simplicity and training efficiency. This indicates that a new paradigm of generating score sequences holds profound implications for future AES, opening new avenues for advanced multi-trait scoring.

Limitations

We identified three limitations of this study. First, although our method achieved competitive results even in low-resource settings, it showed some performance degradation when confronted with extremely limited amounts of data, e.g., the *Voice* trait with less than 1000 samples. This might be attributed to the inherent susceptibility of language models influenced by training data magnitude (Mehrafarin et al., 2022). Second, additional analysis regarding the prediction order can further enhance the scoring quality. Currently, the order is set from rare to frequent traits, which are decided by the number of rated prompts. In future work, we aim to explore more effective ordering strategies through detailed analysis. Lastly, a comprehensive exploration of other pre-trained models could shed more light on future AES. Previously, pre-trained models have only been applied for single-holistic scoring in AES. This could be attributed to the burdensome size of the pre-trained model to approach by constructing duplicated multiple trait-specific layers, unlike existing LSTM and attention-pooling-based models. Therefore, we could not directly compare our model to existing BERT-based systems for each trait scoring. However, as we have demonstrated the autoregressive approach to aid multi-trait AES, we plan to comprehensively investigate other alternative encoder-decoder or GPT-based models as the next step.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00223, Development of digital therapeutics to improve communication ability of autism spectrum disorder patients, 50%) and (No.2019-0-01906, Artificial Intelligence Graduate School Program (POSTECH), 50%).

References

Majdi Beseiso and Saleh Alzahrani. 2020. An empirical analysis of bert embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications*, 11(10).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1011–1020.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt-and trait relation-aware cross-prompt essay trait scoring. *arXiv preprint arXiv:2305.16826*.

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

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

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

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

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.

Yong-Won Lee, Claudia Gentile, and Robert Kantor. 2010. Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, 31(3):391–417.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

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

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.

Houman Mehrafarin, Sara Rajaee, and Mohammad Taher Pilehvar. 2022. On the importance of data size in probing fine-tuned models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 228–238, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891.

Masaki Uto, 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.

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

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569.

A Effect of Model Size

We examine the impact of the pre-trained T5 model size (Table 6). In additional experiments, we utilize T5-Small, T5-Base, and T5-Large, which contain 60 million, 220 million, and 770 million parameters, respectively. Experimental settings are all set as described in our main paper (Section 4).

For both trait-wise and prompt-wise results, overall performance improvements are observed as the model size increases. In particular, the *Voice* trait with only 723 samples, including all training, development, and test sets, outperforms the baseline with ArTS-Large. This result highlights that utilizing larger models could boost the effect of our method, assisting even in severely low-resource environments.

B Comprehensive Results of Additional Experiments

Due to the space constraint, only trait-wise results have been reported for additional experiments in Section 5. In this section, we present both trait-wise and prompt-wise results for each experiment and numerical results for ArTS-Llama2, which are only shown in the graph figure.

C Error Analysis in Prompt Number Guidance

In Section 5, we investigated the impact of providing a prompt when fine-tuning as an ablation study (Table 2, 3). While the QWK results clearly demonstrated the effect of informing the prompt number, we conducted additional error case analysis. In particular, we find out that training with the *"score the essay:"* prefix without providing a prompt (ArTS-w/o *Pr*) often brings in *out-of-range* scoring cases, influencing negatively on the overall QWK score. Each prompt has different score ranges for multiple traits, and we named the *out-of-range* prediction for the prediction that is not inside the corresponding prompt's score range. While there are a five-fold total of 66 *out-of-range* test predictions in ArTS-w/o *Pr*, only one *out-of-range* predictions are observed in ArTS. Note that ArTS is fine-tuned with the prefix *"score the essay of the prompt N:"*. Most out-of-range cases are cases where an essay was mistaken for a different prompt and incorrectly graded based on the range of that prompt. The error case analysis proves that our strategy of prefixing with the prompt number provides clear evidence

	Traits (←)											
Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑ (SD↓)
MTL-BiLSTM (baseline)	**0.764**	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.638 (-)
ArTS-Small	0.712	0.695	0.720	0.667	0.711	0.630	0.606	0.631	0.625	0.694	0.474	0.651 (±0.026)
ArTS-Base (Ours)	0.754	**0.730**	**0.751**	0.698	0.725	0.672	0.668	0.679	0.678	**0.721**	0.570	0.695 (±0.018)
ArTS-Large	0.751	**0.730**	0.750	**0.701**	**0.728**	**0.675**	**0.682**	**0.680**	**0.680**	0.715	**0.603**	**0.700** (±0.024)

Table 6: Experimental results of fine-tuning ArTS with T5-Small, T5-Base, and T5-Large models. The left arrow (←) denotes the direction of trait prediction. Each value denotes the average QWK scores across all prompts for each **trait**.

	Prompts								
Model	1	2	3	4	5	6	7	8	AVG↑ (SD↓)
MTL-BiLSTM (baseline)	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665 (-)
ArTS-Small	0.696	0.669	0.682	0.732	0.712	0.743	0.712	0.492	0.680 (±0.029)
ArTS-Base (Ours)	**0.708**	**0.706**	0.704	**0.767**	0.723	**0.776**	**0.749**	0.603	0.717 (±0.025)
ArTS-Large	0.701	0.698	**0.705**	0.766	**0.725**	0.773	0.743	**0.635**	**0.718** (±0.030)

Table 7: Average QWK scores across all traits for each **prompt**.

	Traits (←)											
Model	Overall	Content	PA	Lang	Nar	Org	Conv	WC	SF	Style	Voice	AVG↑ (SD↓)
MTL-BiLSTM (baseline)	**0.764**	0.685	0.701	0.604	0.668	0.615	0.560	0.615	0.598	0.632	0.582	0.638 (-)
ArTS (Ours)	0.754	**0.730**	0.751	**0.698**	**0.725**	**0.672**	**0.668**	**0.679**	**0.678**	0.721	0.570	**0.695** (±0.018)
ArTS-*rev* (→)	0.739	0.724	0.749	0.687	0.718	0.667	0.658	0.660	0.666	0.711	0.562	0.686 (±0.021)
ArTS-*ind*	0.723	0.717	**0.752**	0.695	0.713	0.649	0.659	0.662	0.675	**0.722**	0.548	0.683 (±0.053)
ArTS-*Llama2*	0.690	0.694	0.716	0.679	0.708	0.666	0.649	0.664	0.660	0.645	**0.584**	0.685 (±0.034)

Table 8: Comprehensive results of models, which are described in Section 5. Each value denotes the average QWK scores across all prompts for each **trait**. ArTS-*rev* (→) predicts traits in reverse order, and 11 different ArTS-*ind* models predict each trait individually. ArTS-*Llama2* denotes the fine-tuned results of the Llama2-13B model.

	Prompts								
Model	1	2	3	4	5	6	7	8	AVG↑ (SD↓)
MTL-BiLSTM (baseline)	0.670	0.611	0.647	0.708	0.704	0.712	0.684	0.581	0.665 (-)
ArTS (Ours)	**0.708**	**0.706**	0.704	**0.767**	0.723	**0.776**	**0.749**	**0.603**	**0.717** (±0.025)
ArTS-*rev* (→)	0.700	0.683	0.702	0.763	**0.730**	0.767	0.734	0.586	0.708 (±0.027)
ArTS-*ind*	0.695	0.679	**0.705**	0.762	0.721	0.756	0.734	0.578	0.704 (±0.041)
ArTS-*Llama2*	0.702	0.641	0.700	0.721	0.691	0.736	0.700	0.592	0.685 (±0.030)

Table 9: Average QWK scores across all traits for each **prompt**.

to the model about essay scoring, especially when there are numerous prompts.