DREsS: Dataset for Rubric-based Essay Scoring on EFL Writing

Anonymous ACL submission

Abstract

Automated essay scoring (AES) is a useful tool in English as a Foreign Language (EFL) writing education, offering real-time essay scores for students and instructors. However, previous AES models were trained on essays and scores irrelevant to the practical scenarios of EFL writing education and usually provided a single holistic score due to the lack of appropriate datasets. In this paper, we release DREsS, a large-scale, standard dataset for rubric-based automated essay scoring. DREsS comprises three sub-datasets: DREsS$_{\text{New}}$, DREsS$_{\text{Std.}}$, and DREsS$_{\text{CASE}}$. We collect DREsS$_{\text{New}}$, a real-classroom dataset with 1.7K essays authored by EFL undergraduate students and scored by English education experts. We also standardize existing rubric-based essay scoring datasets as DREsS$_{\text{Std.}}$. We suggest CASE, a corruption-based augmentation strategy for essays, which generates 20K synthetic samples of DREsS$_{\text{CASE}}$ and improves the baseline results by 45.44%. DREsS will enable further research to provide a more accurate and practical AES system for EFL writing education.

		Content	*Organization*	*Language*
DREsS$_{\text{New}}$		1,782	1,782	1,782
DREsS$_{\text{Std.}}$	ASAP P7	1,569	1,569	1,569
	ASAP P8	723	723	723
	ASAP++ P1	1,785	1,785	1,785
	ASAP++ P2	1,800	1,800	1,800
	ICNALE EE	639	639	639
DREsS$_{\text{CASE}}$		3,924	15,696	981
Total		12,222	23,994	14,845

Table 1: Data statistics

1 Introduction

In writing education, automated essay scoring (AES) can provide real-time scores of students' essays to both students and instructors. For many students who are hesitant to expose their errors to instructors, the immediate assessment of their essays with AES can create a supportive environment for self-improvement in writing skills (Sun and Fan, 2022). For instructors, AES models can ease the time-consuming process of evaluation and serve as a means to validate their assessments, ensuring consistency in their evaluations.

AES systems can provide either a holistic or an analytic view of essays, but rubric-based, analytical scores are more preferred in the EFL writing education domain (Ghalib and Al-Hattami, 2015). However, there is only a limited amount of rubric-based datasets available for AES, and the rubrics are not consistent in building generalizable AES systems. Furthermore, AES datasets must be annotated by writing education experts because the scoring task requires pedagogical knowledge in English writing. To date, there is a lack of usable datasets for training rubric-based AES models, as existing AES datasets provide only overall scores and/or make use of scores annotated by non-experts.

In this paper, we release DREsS, a large-scale dataset for rubric-based essay scoring using three key rubrics: *content*, *organization*, and *language*. DREsS consists of three datasets: 1) DREsS$_{\text{New}}$ with 1,782 essays from English as a foreign language (EFL) learners and their scores assessed by experts, 2) DREsS$_{\text{Std.}}$ with 6,516 essays and scores from existing datasets, and 3) DREsS$_{\text{CASE}}$ with 20,601 synthetic essay samples. We standardize and rescale existing rubric-based datasets to align our rubrics. We also suggest CASE, a corruption-based augmentation strategy for Essays, employing three rubric-specific strategies to augment the dataset with corruption. DREsS$_{\text{CASE}}$ improves the baseline result by 45.44%.

2 Related Work

2.1 Holistic AES

ASAP Prompt 1-6 ASAP dataset [1] is widely used in AES tasks, involving eight different prompts.

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

Six out of eight prompt sets (Prompt 1-6) have a single overall score. This holistic AES includes 10K essay scoring data on source-dependent essay (Prompt 3-6) and argumentative essay (Prompt 1-2). However, these essays are graded by non-expert annotators, though the essays were written by Grade 7-10 students in the US.

TOEFL11 TOEFL11 (Blanchard et al., 2013) corpus from ETS introduced 12K TOEFL iBT essays, which are not publicly accessible now. TOEFL11 only provides a general score for essays in 3 levels (low/mid/high), which is insufficient for building a well-performing AES system.

Models The majority of the previous studies used the ASAP dataset for training and evaluation, aiming to predict the overall score of the essay only (Tay et al., 2018; Cozma et al., 2018; Wang et al., 2018; Yang et al., 2020). Enhanced AI Scoring Engine (EASE)[2] is a commonly used, open-sourced AES system based on feature extraction and statistical methods. In addition, Taghipour and Ng (2016) and Xie et al. (2022) released models based on recurrent neural networks and neural pairwise contrastive regression (NPCR) model, respectively. However, only a limited number publicly opened their models and code, highlighting the need for additional publicly available data and further validation of existing models.

2.2 Rubric-based AES

ASAP Prompt 7-8 ASAP includes only two prompts (Prompt 7-8) that are rubric-based. These two rubric-based prompts consist of 1,569 and 723 essays for each respective prompt. The two prompt sets even have distinct rubrics and score ranges, which poses a challenge in leveraging both datasets for training rubric-based models. These dataset (Prompt 7-8) is also evaluated by non-expert annotators, similar to ASAP Prompt 1-6.

ASAP++ To overcome the holistic scoring of ASAP Prompt 1-6, Mathias and Bhattacharyya (2018) manually annotated rubric-based scores on those essays. However, most samples in ASAP++ were annotated by a single annotator, who is a non-expert, including non-native speakers of English. Moreover, each prompt set of ASAP++ has different attributes or rubrics to each other, which need to be more generalizable to fully leverage such dataset for AES model.

[2] https://github.com/edx/ease

Content. Paragraph is well-developed and relevant to the argument, supported with strong reasons and examples.
Organization. The argument is very effectively structured and developed, making it easy for the reader to follow the ideas and understand how the writer is building the argument. Paragraphs use coherence devices effectively while focusing on a single main idea.
Language. The writing displays sophisticated control of a wide range of vocabulary and collocations. The essay follows grammar and usage rules throughout the paper. Spelling and punctuation are correct throughout the paper.

Table 2: Explanation of rubrics

ICNALE Edited Essays ICNALE Edited Essays (EE) v3.0 (Ishikawa, 2018) presents rubric-based essay evaluation scores and fully edited versions of essays written by EFL learners from 10 countries in Asia. Even though the essays are written by EFL learners, the essay is rated and edited only by five native English speakers, non-experts in the domain of English writing education. In addition, it is not openly accessible and only consists of 639 samples.

Models The scarcity of publicly available rubric-based AES datasets poses significant obstacles to the advancement of AES research. There are industry-driven services such as IntelliMetric® (Rudner et al., 2006) and E-rater® (Attali and Burstein, 2006; Blanchard et al., 2013), but none of them are accessible to the public. In order to facilitate AES research in the academic community, it is crucial to release a publicly available rubric-based AES dataset and baseline model.

3 DREsS Dataset

We construct 1.7K samples of our newly released DREsS dataset (§3.1), 2.9K standardized samples of existing datasets (§3.2), and 20K synthetic samples augmented using CASE (§3.3). The detailed number of samples per rubric is stated in Table 1.

3.1 Dataset Collection

Dataset Details $\text{DREsS}_{\text{New}}$ includes 1,782 argumentative essays on 22 prompts, having 313.36 words and 21.19 sentences on average. Each sample in DREsS includes students' written essay, essay prompt, rubric-based scores, total score, and a test type (pre-test, post-test). The essays are scored on a range of 1 to 5, with increments of 0.5, based on the three rubrics: *content*, *organization*, and *language*. We chose such three conventional rubrics as standard criteria for scoring EFL essays, according to previous studies from the language education (Cumming, 1990; Ozfidan and Mitchell, 2022). Detailed explanations of the rubrics are shown in Table 2. The essays are written by undergraduate students whose TOEFL writing score spans from 15 to 21 and enrolled in EFL writing courses at a college in South Korea from 2020 to 2023. Most students are Korean and their ages span from 18 to 22, with an average of 19.7. During the course, students are asked to write an in-class timed essay for 40 minutes both at the start (pre-test) and the end of the semester (post-test) to measure their improvements.

Annotator Details We collect scoring data from 11 instructors, who serve as the teachers of the students who wrote the essays. All annotators are experts in English education or Linguistics and are qualified to teach EFL writing courses at a college in South Korea. To ensure consistent and reliable scoring across all instructors, they all participate in training sessions with a scoring guide and norming sessions where they develop a consensus on scores using two sample essays. Additionally, there was no significant difference among the score distribution of all instructors tested by one-way ANOVA and Tukey HSD at a p-value of 0.05.

3.2 Standardizing the Existing Data

We standardize and unify three existing rubric-based datasets (ASAP Prompt 7-8, ASAP++ Prompt 1-2, and ICNALE EE) to align with the three rubrics in DREsS: *content*, *organization*, and *language*. We exclude ASAP++ Prompt 3-6, whose essay type, source-dependent essays, is clearly different from argumentative essays. ASAP Prompt 7 contains four rubrics – ideas, organization, style, and convention – while Prompt 8 contains six rubrics – ideas and content, organization, voice, word choice, sentence fluency, and convention. Both sets provide scores ranging from 0 to 3. For *language*, we first create synthetic labels based on a weighted average. This involves assigning a weight of 0.66 to the style and 0.33 to the convention in ASAP Prompt 7, and assigning equal weights to voice, word choice, sentence fluency, and convention in ASAP Prompt 8. For *content* and *organization*, we utilize the existing data rubric (idea for content, organization as same) in the dataset. We then rescale the score of all rubrics into a range of 1 to 5. We repeat the same process with ASAP++ Prompt 1 and 2, which have the same attributes as ASAP Prompt 8. Similarly, for ICNALE EE dataset, we unify vocabulary, language use, and mechanics as language rubric with a weight of 0.4, 0.5, and 0.1, respectively. In the process of consolidating the writing assessment criteria, we sought professional consultation from EFL education experts and strategically grouped together those components that evaluate similar aspects.

3.3 Synthetic Data Construction

We construct synthetic data for rubric-based AES to overcome the scarcity of data and provide accurate scores for students and instructors. We introduce a corruption-based augmentation strategy for essays (CASE), which starts with a *well-written* essay and incorporates a certain portion of sentence-level errors into the synthetic essay. In subsequent experiments, we define *well-written* essays as an essay that scored 4.5 or 5.0 out of 5.0 on each criterion.

$$\mathrm{n}(S_c) = \lfloor \mathrm{n}(S_E) * (5.0 - x_i)/5.0 \rceil \quad (1)$$

$\mathrm{n}(S_c)$ is the number of corrupted sentences in the synthetic essay, and $\mathrm{n}(S_E)$ is the number of sentences in the *well-written* essay, which serves as the basis for the synthetic essay. x_i denotes the score of the synthetic essay. In this paper, we generate synthetic data with CASE under ablation study for exploring the optimal number of samples.

Content We substitute randomly-sampled sentences from *well-written* essays with out-of-domain sentences from different prompts. This is based on an assumption that sentences in *well-written* essays support the given prompt's content, meaning that sentences from the essays on different prompts convey different contents. Therefore, more number of substitutions imply higher levels of corruption in the content of the essay.

Organization We swap two randomly-sampled sentences in *well-written* essays and repeat this process based on the synthetic score, supposing that sentences in *well-written* essays are systematically structured in order. The higher number of swaps implies higher levels of corruption in the organization of the essay.

Language We substitute randomly-sampled sentences into ungrammatical sentences and repeat this process based on the synthetic score. We extract 605 ungrammatical sentences from BEA-2019 data for the shared task of grammatical error correction (GEC) (Bryant et al., 2019). We define ungrammatical sentences with the number of edits of the sentence over 10, which is the 98th percentile. The

Model	Fine-tuning Data	*Content*	*Organization*	*Language*	Total
`gpt-3.5-turbo`	N/A	0.239	0.371	0.246	0.307
EASE (SVR)	DREsS	-	-	-	0.360
NPCR (Xie et al., 2022)	DREsS	-	-	-	0.507
BERT (Devlin et al., 2019)	$\text{DREsS}_{\text{New}}$	0.414	0.311	0.487	0.471
	+ $\text{DREsS}_{\text{Std.}}$	0.599	0.593	0.587	0.551
	+ $\text{DREsS}_{\text{CASE}}$	**0.642**	**0.750**	**0.607**	**0.685**

Table 3: Baseline results of rubric-based automated essay scoring on DREsS (QWK score)

more substitutions, the more corruption is introduced in the grammar of the essay. We set such a high threshold for ungrammatical sentences because of the limitation of the current GEC dataset that inherent noise may be included, such as erroneous or incomplete correction (Rothe et al., 2021).

4 Experimental Result

Table 3 shows the baseline results of rubric-based AES on DREsS. Detailed experimental settings are described in Appendix §A. We provide the baseline results on DREsS using fine-tuned BERT (Devlin et al., 2019), which is the model that most state-of-the-art AES systems have leveraged. The same experimental results with different PLMs are provided in Appendix §B, though Xie et al. (2022) observed no significant improvements among various pre-trained language models (PLMs) in AES.

Fine-tuned BERT exhibits scalable results with the expansion of training data. In particular, the model trained with a combination of our approaches outperforms other baseline models by 45.44%, demonstrating the effectiveness of data unification and augmentation using CASE.

The results from existing holistic AES models underscore the need to examine these models using new datasets. EASE and NPCR reported the QWK scores as 0.699 and 0.817 on ASAP, respectively. However, the scores significantly decrease when these models are trained and evaluated with DREsS dataset. This result implies that EASE and NPCR are not robust to different datasets. Additionally, these models, which are trained with ASAP, are not able to predict general rubric-based scoring.

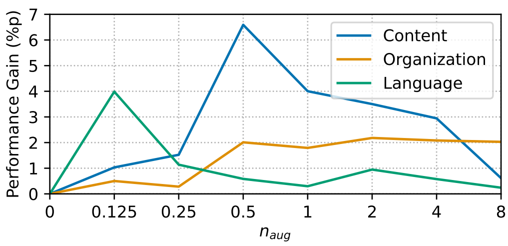

Figure 1: Ablation experimental results for CASE. n_{aug} is the number of synthetic data by each class per original data among all classes.

Asking `gpt-3.5-turbo` to score an essay achieved the worst performances among all, showing high variances among the essays with the same score. The detailed results for ChatGPT in different prompt settings are provided in Appendix §C.

We perform an ablation study to find the optimal number of CASE operations per each rubric. In Figure 1, we investigate how the number of CASE operations affects the performance over all rubrics for $n_{aug} = \{0.125, 0.25, 0.5, 1, 2, 4, 8\}$. CASE on *content*, *organization*, and *language* rubrics show their best performances on 0.5, 2, 0.125 of n_{aug}, generating a pair of synthetic essays and corresponding scores in 4.5, 18, 1.125 times, respectively. We suppose that the detailed augmentation strategies for each rubric and the small size of the original data affect the optimal number of CASE operations. *Organization*, where corruption was made within the essay and irrelevant to the size of the original data, showed the highest n_{aug}. *Content*, where the corrupted sentences were sampled from 874 *well-written* essays with 21.2 sentences on average, reported higher n_{aug} than language, where the corrupted sentences were sampled from 605 ungrammatical sentences.

5 Conclusion

We release the DREsS, a large-scale, standard rubric-based essay scoring dataset with three subsets: $\text{DREsS}_{\text{New}}$, $\text{DREsS}_{\text{Std.}}$, and $\text{DREsS}_{\text{CASE}}$. $\text{DREsS}_{\text{New}}$ is the first reliable AES dataset with 1.7K samples whose essays are authored by EFL undergraduate students and whose scores are annotated by instructors with expertise. According to previous studies from language education, we also standardize and unify existing rubric-based AES datasets as $\text{DREsS}_{\text{Std.}}$. We finally suggest CASE, corruption-based augmentation strategies for essays, which generates 20K synthetic samples and improves the baseline result by 45.44%. This work aims to encourage further AES research and practical application in EFL education.

Limitations

Our research focuses on learning *English* as a foreign language because there already exist datasets, and the current language models perform the best for English. There are many L2 learners of other languages whose writing classes can also benefit from AES. Our findings can illuminate the directions of data collection, annotation, and augmentation for L2 writing education of other languages as well. We leave that as future work.

Our DREsS dataset is collected through the EFL writing courses from a college in South Korea, and most of the essays are written by Korean EFL students. EFL students in different cultural and linguistic backgrounds might exhibit different essay-writing patterns, which might affect the distribution of scores and feedback. We suggest a further extension of collecting the DREsS dataset from diverse countries.

Our augmentation strategy primarily starts from *well-written* essays and generates erroneous essays along with corresponding scores; therefore, this approach faces challenges in synthesizing *well-written* essays. However, we believe that *well-written* essays can be reliably produced by LLMs, which have demonstrated strong capabilities in generating high-quality English text.

Ethics Statement

We expect that this paper will make a significant contribution to the application of NLP for good, particularly in the domain of NLP-driven assistance in EFL writing education. All studies in this research project were conducted with the approval of our institutional review board (IRB). To prevent any potential impact on student scores or grades, we requested students to share their essays only after the end of the EFL courses. We also acknowledged and addressed the potential risk associated with releasing a dataset containing human-written essays, especially considering privacy and personal information. To mitigate these risks, we plan to 1) employ rule-based coding and 2) conduct thorough human inspections to filter out all sensitive information. Additionally, access to our data will be granted only to researchers or practitioners who submit a consent form, ensuring responsible and ethical usage.

References

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

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. 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.

Alister Cumming. 1990. Expertise in evaluating second language compositions. *Language Testing*, 7(1):31–51.

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.

Thikra K Ghalib and Abdulghani A Al-Hattami. 2015. Holistic versus analytic evaluation of efl writing: A case study. *English Language Teaching*, 8(7):225–236.

Shinichiro Ishikawa. 2018. The icnale edited essays; a dataset for analysis of l2 english learner essays based on a new integrative viewpoint. *English Corpus Studies*, 25:117–130.

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

Burhan Ozfidan and Connie Mitchell. 2022. Assessment of students' argumentative writing: A rubric development. *Journal of Ethnic and Cultural Studies*, 9(2):pp. 121–133.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. 2006. An evaluation of intellimetric™ essay scoring system. *The Journal of Technology, Learning and Assessment*, 4(4).

Bo Sun and Tingting Fan. 2022. The effects of an awe-aided assessment approach on business english writing performance and writing anxiety: A contextual consideration. *Studies in Educational Evaluation*, 72:101123.

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.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

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

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.

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, Online. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Appendix

A Experimental Settings

We split our data into train/dev/test with 6:2:2 ratio with a seed of 22. The AES experiments were conducted under GeForce RTX 2080 Ti (4 GPUs), 128GiB system memory, and Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz (20 CPU cores) with hyperparameters denoted in Table 4.

Hyperparameter	Value
Batch Size	32
Number of epochs	10
Early Stopping Patience	5
Learning Rate	2e-5
Learning Rate Scheduler	Linear
Optimizer	AdamW

Table 4: Model configuration

B Rubric-based AES with Different LMs

Model	Content	Organization	Language	Total
BERT (2019)	0.414	0.311	0.487	0.471
Longformer (2020)	0.409	0.312	0.475	0.463
BigBird (2020)	0.412	0.317	0.473	0.469
GPT-NeoX (2022)	0.410	0.313	0.446	0.475

Table 5: Experimental results on rubric-based AES with different fine-tuned LMs

Experimental results of rubric-based AES with different LMs are provided in Table 5, showing no significant difference among different LMs. Xie et al. (2022) also observed that leveraging different LMs has no significant effect on AES performance, and most state-of-the-art AES methods have leveraged BERT (Devlin et al., 2019).

C Rubric-based AES with ChatGPT

Prompt	Content	Organization	Language	Total
(A)	0.320	0.248	0.359	0.336
(B)	0.330	0.328	0.306	0.346
(C)	0.357	0.278	0.342	0.364
(D)	0.336	0.361	0.272	0.385

Table 6: Rubric-based quadratic weighted kappa (QWK) scores of ChatGPT (`gpt-3.5-turbo`, temperature: 0) with diverse prompts.
(A): standard zero-shot prompting, (B): 2-shot prompting, (C): zero-shot with rubric explanation (D): zero-shot with feedback generation.

(A)	Q. Please score the essay with three rubrics: content, organization, and language. ### Answer format: {content: score[x], organization: score[y], language: score[z]} score = [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0] Please answer only in the above dictionary format, without feedback. ### prompt: `<essay prompt>` ### essay: `<student's essay>` A:
(B)	Q. Please score the essay with three rubrics: content, organization, and language. ### Answer format: {content: score[x], organization: score[y], language: score[z]} score = [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0] Please answer only in the above dictionary format, without feedback. ### 1-shot example: ### 2-shot example: ### prompt: `<essay prompt>` ### essay: `<student's essay>` A:
(C)	Q. Please score the essay with three rubrics: content, organization, and language. `<three rubrics explanation>` ### Answer format: {content: score[x], organization: score[y], language: score[z]} score = [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0] Please answer only in the above dictionary format, without feedback. ### prompt: `<essay prompt>` ### essay: `<student's essay>` A:
(D)	Q. Please score the essay with three rubrics: content, organization, and language. ### Answer format: {content: score[x], organization: score[y], language: score[z], content_fbk: chatgpt_con_fbk, org_fbk: chatgpt_org_fbk, lang_fbk: chatgpt_lang_fbk} score = [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0] Please answer only in the above dictionary format, with feedback. ### prompt: `<essay prompt>` ### essay: `<student's essay>` A:

Table 7: Different prompts for ChatGPT to get rubric-based scores. Refer to Table 6 for descriptions of setting (A)–(D).

Table 6 shows AES results of ChatGPT with different prompts described in Table 7. Considering the substantial length of the essay and feedback, we were able to provide a maximum of 2 shots for the prompt to `gpt-3.5-turbo`. To examine 2-shot prompting performance, we divided the samples into two distinct groups and computed the average total score for each group. Subsequently, we randomly sampled a single essay in each group, ensuring that its total score corresponded to the calculated mean value.