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

Haneul Yoo Jieun Han So-Yeon Ahn Alice Oh
KAIST
`{haneul.yoo, jieun_han, ahnsoyeon}@kaist.ac.kr, alice.oh@kaist.edu`

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 with 48.9K samples in total. 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 2.3K 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 40.1K 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.[1]

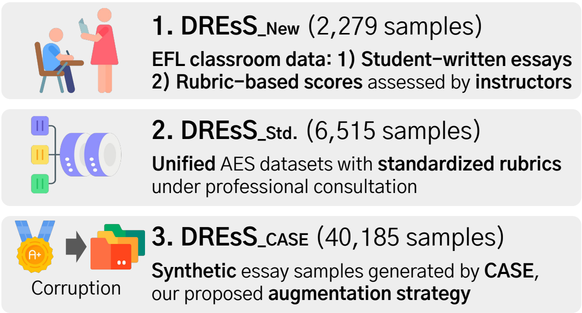


Figure 1: Data construction of DREsS

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 of 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 (**D**ataset for **R**ubric-based **E**ssay **S**coring on EFL Writing), 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 2,279 essays from English as a foreign language (EFL) learners and their scores assessed by experts, 2) DREsS$_{\text{Std.}}$ with 6,515 essays and scores from existing datasets, and 3) DREsS$_{\text{CASE}}$ with 40,185 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%.

---

[1] DREsS is publicly available at `https://haneul-yoo.github.io/dress/`.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13439–13454
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

		Content	*Organization*	*Language*
DREsS$_{\text{New}}$		2,279	2,279	2,279
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	1,799	1,799	1,799
	ICNALE EE	639	639	639
DREsS$_{\text{CASE}}$		8,307	31,086	792
Total		17,101	39,880	9,586

Table 1: Data statistics of DREsS

2 Related Work

In this section, we describe previous studies in automated essay scoring (AES) in terms of the format of predicted scores: holistic AES (§2.1) and rubric-based AES (§2.2). To date, there is only a limited amount of publicly available AES datasets, and their rubrics are inconsistent. Furthermore, their scores are usually annotated by non-experts lacking pedagogical knowledge in English writing. Here, we introduce DREsS, a publicly available, large-scale, rubric-based, real-classroom dataset, which can be used as training data for rubric-based AES systems.

2.1 Holistic AES

ASAP Prompt 1-6 ASAP dataset[2] is widely used in AES tasks, involving eight different prompts. 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, *inter alia*). Enhanced AI Scoring Engine (EASE)[3] 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. Still, only a limited number of studies publicly opened their models and codes, 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 essays (Prompt 7-8) are 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.

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

---

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

[3] `https://github.com/edx/ease`

in Asia. Even though the essays are written by EFL learners, the essay is rated and edited only by a single annotator per sample. They have five native English speakers, non-experts in the domain of English writing education in total. 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® (Blanchard et al., 2013; Attali and Burstein, 2006), but none of them are accessible to the public. Kumar et al. (2022) proposed applying a multi-task learning approach in holistic AES with ASAP and ASAP++, using traits as auxiliary tasks. Recent studies have followed up their method, introducing multi-traits AES approaches (Chen and Li, 2023; Do et al., 2023, 2024; Lee et al., 2024, *inter alia*). Still, they shed light on predicting a holistic score only due to limited data and built eight different fine-tuned models due to unconsolidated rubrics by each essay prompt. Previous studies have explored diverse non-English languages, including Chinese (Song et al., 2020; He et al., 2022), Japanese (Hirao et al., 2020), and French (Wilkens et al., 2023), while most of them have mimicked and adapted existing state-of-the-art techniques into non-English languages. 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 DREsS with 2.3K samples of our newly collected dataset (§3.1), 6.5K standardized samples of existing datasets (§3.2), and 40.1K synthetic samples augmented using CASE (§3.3). The detailed number of samples is stated in Table 1.

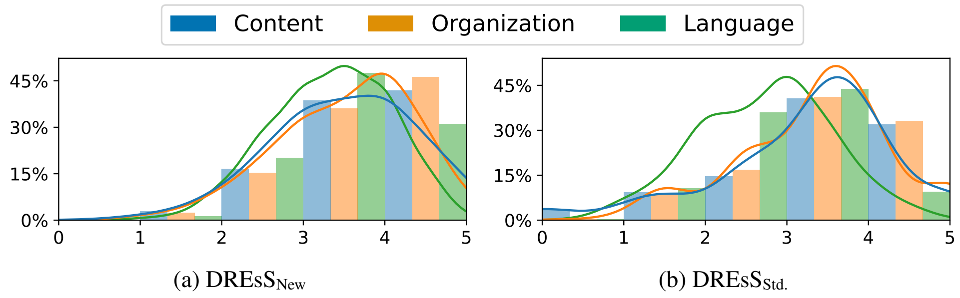
3.1 Dataset Collection

Dataset Details DREsS$_{\text{New}}$ includes 2,279 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 (the sum of three rubric-based scores), 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). Brief 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.

Rubric	Description
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: Rubric explanations

Annotator Details We collect scoring data from 11 instructors, who serve as the teachers of the students who wrote the essays. Six of them are non-native speakers, and five of them are native speakers. All annotators are experts in English education or Linguistics and are qualified to teach EFL writing courses at a college in South Korea. One instructor was allocated per essay, so the inter-annotator agreement cannot be measured. It follows that an EFL course is usually led by a single instructor, and the essays from the course are assessed by the instructor in a real-classroom setting.



(a) $\text{DREsS}_{\text{New}}$ (b) $\text{DREsS}_{\text{Std.}}$

Figure 2: Score distribution of DREsS

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. We create synthetic label based on a weighted average and then rescale the score of all rubrics into a range of 1 to 5. Detailed explanations and rationales behind standardizing weights are described in Appendix C. 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 under theoretical considerations.

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.

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

$\mathrm{n}(S_c)$ 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 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).

3.4 Score Distribution

Figure 2 shows the score distribution of $\text{DREsS}_{\text{New}}$ and $\text{DREsS}_{\text{Std.}}$ ranging from 0 to 5. The score dis-

Model	Strategy	*Content*	*Organization*	*Language*	Total
EASE (SVR)	SFT w/ DREsS	-	-	-	0.360
NPCR (Xie et al., 2022)		-	-	-	0.507
ArTS (Do et al., 2024)		0.601	0.743	0.592	0.690
BERT (Devlin et al., 2019)		**0.642**	0.750	**0.607**	0.685
Llama 3.1 8B (Grattafiori et al., 2024)		0.631	**0.771**	0.589	**0.691**
`gpt-4o`	(A) zero-shot ICL	0.310	0.322	0.231	0.304
	(B) five-shot ICL	0.361	0.475	0.367	0.428
	(C) rubric explanation	0.285	0.250	0.200	0.259
	(D) feedback generation	0.313	0.268	0.230	0.290

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

tribution of the AES dataset shows a left-skewed bell-shaped curve, following the general trends in real-classroom settings. The scarcity of samples on low scores is because instructors are reluctant to give low scores to increase students' self-efficacy and motivate them to learn (Arsyad Arrafii, 2020). To overcome the imbalance of the dataset, we propose CASE, which can generate synthetic data for all score ranges. DREsS$_{\text{CASE}}$ has the same number of samples per score.

4 Experimental Result

4.1 Baseline Result on DREsS

Table 3 shows the baseline results of rubric-based AES on DREsS. We use all three subsplits of DREsS as training data, but DREsS$_{\text{New}}$, a subsplit comprising essays and scores from real classroom settings, is used exclusively for the validation and the test sets. In other words, synthetically unified (DREsS$_{\text{Std.}}$) or augmented (DREsS$_{\text{CASE}}$) data are reserved for training to avoid incomplete or inaccurate evaluation. Detailed experimental settings are described in Appendix §A. We adopt the quadratic weighted kappa (QWK) scores, a conventional metric to evaluate the consistency between the predicted scores and the gold standard scores.

We provide the baseline results on DREsS using holistic AES models from previous studies (*i.e.,* EASE (SVR), NPCR (Xie et al., 2022), and ArTS (Do et al., 2024)), large language model (*i.e.,* `gpt-4o` from OpenAI[4] and Llama 3.1 8B (Grattafiori et al., 2024) from Meta), and BERT (Devlin et al., 2019). Note that fine-tuned BERT is the model that most state-of-the-art AES systems have leveraged. We train EASE (SVR), NPCR, ArTS, BERT, and Llama 3.1 with DREsS as supervised fine-tuning (SFT) data. We also test `gpt-4o` with four different system prompts as follows:

(A) in-context learning (ICL) with zero-shot

(B) in-context learning (ICL) with five-shots of writing prompts and essays

(C) asking the model to predict essay scores given detailed rubric explanations

(D) asking the model to predict essay scores and provide essay feedbacks that support their predicted scores.

The detailed prompts are described in Appendix B.1. Considering the substantial length of writing prompt and essay, we were able to provide a maximum of 5 shots for the prompt to `gpt-4o`. We divided the samples into five distinct score ranges 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. Asking `gpt-4o` to score an essay shows high variances among the essays with the same score, implying their limitations to be applied as AES systems.

4.2 Validation of DREsS$_{\text{Std.}}$ and DREsS$_{\text{CASE}}$

Table 4 shows experimental results of rubric-based AES with different language models. We train Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020), a language model that accepts long input sequences (*i.e.,* 4,096 tokens), considering the substantial length of writing prompts and essays. In addition, we train GPT-NeoX-20B (Black et al., 2022) and Llama 3.1 8B, state-of-the-art LLMs. Nonetheless, exploiting different models

---

[4] All following experiments using `gpt-4o` in this paper was conducted from May 21, 2024 to June 5, 2024 under OpenAI API services.

Model	Strategy	*Content*	*Organization*	*Language*	Total
BERT (Devlin et al., 2019)	SFT w/ DREsS$_{\text{New}}$	**0.414**	0.311	**0.487**	0.471
Longformer (Beltagy et al., 2020)		0.409	0.312	0.475	0.463
BigBird (Zaheer et al., 2020)		0.412	0.317	0.473	0.469
GPT-NeoX-20B (Black et al., 2022)		0.410	0.313	0.446	**0.475**
Llama 3.1 8B (Grattafiori et al., 2024)		0.413	**0.375**	0.426	0.466

Table 4: Experimental results of rubric-based AES with different LMs using DREsS$_{\text{New}}$

Model	Strategy	*Content*	*Organization*	*Language*	Total
BERT (Devlin et al., 2019)	SFT w/ DREsS$_{\text{New}}$	0.414	0.311	0.487	0.471
	+ DREsS$_{\text{Std.}}$	0.599	0.593	0.587	0.551
	+ DREsS$_{\text{CASE}}$	**0.642**	0.750	**0.607**	0.685
Llama 3.1 8B (Grattafiori et al., 2024)	SFT w/ DREsS$_{\text{New}}$	0.413	0.375	0.426	0.466
	+ DREsS$_{\text{Std.}}$	0.581	0.608	0.574	0.563
	+ DREsS$_{\text{CASE}}$	0.631	**0.771**	0.589	**0.691**

Table 5: Empirical validation of data expansion in DREsS

does not significantly affect the performance of AES systems. Xie et al. (2022) also observed that leveraging different foundation models has no significant effect on AES performance, and most state-of-the-art AES methods have still leveraged BERT (Devlin et al., 2019). Therefore, based on these observations, we choose BERT and Llama 3.1 (8B) as a representative model to further evaluate and validate the effectiveness of our dataset, particularly focusing on the benefits of data standardization and synthesis.

We validate the practical benefits of data standardization (DREsS$_{\text{Std.}}$) and synthesis (DREsS$_{\text{CASE}}$) with empirical results. Both fine-tuned BERT and Llama 3.1 exhibit scalable results with the expansion of training data (Table 5). 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. Interestingly, a state-of-the-art LLM (*i.e.,* `gpt-4o`) does not outperform fine-tuned small-scale language models (*i.e.,* BERT), achieving 0.257 points lower QWK total score. Existing holistic AES models show their inability to compute rubric-based scores.

5 Discussion & Analysis

5.1 Ablation Study

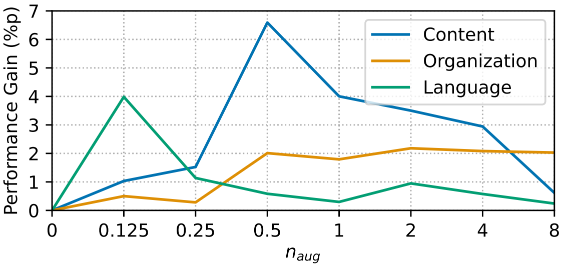


Figure 3: Ablation experimental results for CASE. n_{aug} is the number of synthetic data by each class per original data among all classes. The x-axis is a log-arithmetic scale.

We perform an ablation study to find the optimal number of CASE operations per each rubric. In Figure 3, we investigate how the number of CASE operations affects the performance over all rubrics for $n_{aug} = \{0.125, 0.25, 0.5, 1, 2, 4, 8\}$, where n_{aug} denotes the number of synthetic data by each class per original data among all classes (*i.e.,* the ratio of augmented data size compared to the source data size). 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. Leveraging more error patterns in new grammatical error correction (GEC) data will lead to a scalable increase in the size of DREsS$_{\text{CASE}}$ for *language*.

5.2 CASE vs. Generative Methods

We verify the quality of synthetic data using CASE compared to generative methods using LLMs. Here, we use the best-performing baseline rubric-based scoring models trained with DREsS. We measure a quadratic weighted kappa (QWK) score to measure the similarity between the gold label of the synthetic sample and the predicted score by an AES model.

For LLM to generate synthetic essays, we first give the persona of an EFL student taking an English writing course in a college for students who get TOEFL scores ranging from 15 to 21 and provide five example essays written by EFL students randomly sampled from five distinct score ranges. We then ask the model to write an essay that matches the rubric-based scores. The detailed prompts to generate synthetic EFL essays are described in Appendix B.2. We randomly sample 900 essays (100 samples per score ranging from 1.0 to 5.0 with an increment of 0.5) from CASE augmentation and synthetic samples generated by `gpt-4o`. Table 6 shows QWK scores of synthetic essays, which validate whether the essays match with their scores. We use the best-performing baseline rubric-based scoring models in Table 4, which only uses `DREsS`$_{\text{New}}$ as its training and test set. QWK score of CASE augmentation achieves 0.661 (*substantial agreement*), while the score of the generative method achieves 0.225 on average (*slight to fair agreement*). Though the detailed persona and example essays are given, `gpt-4o` fails to write an appropriate level of essays. Specifically, the predicted rubric-based scores of 900 synthetic essays from `gpt-4o` across all score ranges are $4.21_{\pm 0.65}$, $4.13_{\pm 0.63}$, and $4.30_{\pm 0.30}$ for *content*, *organization*, and *language*, respectively.

We discuss the benefit of leveraging CASE to generate synthetic essays in EFL writing for three reasons: 1) its difficulty in generating EFL students' essays, 2) low performance in scoring essays, and 3) controllability and interoperability. First of all, LLMs are hardly capable of replicating EFL learners' errors since they are mostly trained with texts from native speakers. The essays of DREsS$_{\text{New}}$ written by EFL students reveal various unique characteristics and error patterns of EFL learners. Detailed analysis is described in § 5.3. Second, we found that the state-of-the-art LLM, namely `gpt-4o`, underperforms in essay scoring tasks compared to BERT-based models, as described in Table 3. Lastly, the black-box nature of LLMs poses challenges in terms of controllability and interpretability. In contrast, our proposed CASE method offers enhanced control and interpretability. This mitigates the risks associated with over-reliance on generative methods, fostering a more robust and transparent research approach.

	Content	*Organization*	*Language*
`gpt-4o`	0.298	0.219	0.158
CASE (Ours)	0.625	0.722	0.635

Table 6: QWK scores of synthetic essays generated by two augmentation methods

5.3 In-depth Analysis

Table 7 shows quantitative analysis of essays from DREsS$_{\text{New}}$ and DREsS$_{\text{CASE}}$ compared to `gpt-4o` augmentation concerning linguistic features. Student-written essays in DREsS$_{\text{New}}$ include unique patterns of EFL learners. For instance, essays in DREsS$_{\text{New}}$ tend to be longer than synthetic essays from `gpt-4o`, with more number of sentences but easier and shorter sentences, according to Flesch reading ease (Flesch, 1948) and the number of tokens, respectively. Interestingly, EFL students use fewer unique words but frequently use unnecessary stopwords. Essays from EFL students include typos and spelling errors which cannot be made from the generation outputs of LLMs. Note that one of the major strengths of the DREsS dataset is the inclusion of errorful essays written by EFL learners in the real-world classroom.

Table 8 shows two sample essays with a score of 1 under the same writing prompt. The synthetic essay from `gpt-4o` fails to reflect the EFL learners' errors, generating essays that include *content*, *organization*, and *language* features needed for a well-written essay. For *organization*, the essay from `gpt-4o` is well-structured with the use of appropriate transition signals and an appropriate thesis sentence in the first paragraph (blue text). For *content*, each body paragraph includes detailed examples to support the argument (orange text). For language, the essay does not include any grammatical errors. In contrast, the essay from DREsS$_{\text{New}}$ lacks tran-

	DREsS$_{\text{New}}$	DREsS$_{\text{CASE}}$	`gpt-4o`
# of sentences *	$20.96_{\pm 6.66}$	$22.67_{\pm 10.10}$	$16.02_{\pm 2.35}$
# of tokens *	$313.97_{\pm 96.76}$	$327.91_{\pm 56.01}$	$285.84_{\pm 69.07}$
# of tokens w/o stopwords	$162.64_{\pm 49.97}$	$167.14_{\pm 35.50}$	$165.49_{\pm 47.91}$
Type-token ratio (TTR) *	$0.43_{\pm 0.07}$	$0.43_{\pm 0.06}$	$0.51_{\pm 0.04}$
# of transition signal *	$28.03_{\pm 10.3}$	$28.18_{\pm 14.0}$	$29.61_{\pm 10.02}$
# of typos *	$4.39_{\pm 4.11}$	$6.64_{\pm 8.22}$	$0.59_{\pm 1.22}$
Flesch reading ease (Flesch, 1948) *	$53.66_{\pm 11.84}$	$59.47_{\pm 11.11}$	$29.15_{\pm 19.88}$
US grade level (Kincaid et al., 1975) *	$9.58_{\pm 2.11}$	$9.12_{\pm 2.58}$	$13.64_{\pm 3.5}$

Table 7: Quantitative analysis. The asterisk denotes a statistically significant difference between `gpt-4o` and others tested under one-way ANOVA and Tukey HSD test.

DREsS$_{\text{New}}$	`gpt-4o`
In Korea, most of high school students stay with very regular schedule and eat healthy lunch. However, in college they start drinking and often eat fast food. Their health going bad! College students have much free time and they can do their hobby more. Then, spending time to exercise. Exercise can help college student. Many college people want to make girl or boyfriend. If they start exercise and make healthy body, it can be positive effect to people.	If I have the chance to visit a foreign country for two weeks, I would like to go to Japan. Japan is a country that I have always wanted to visit because of its unique culture and beautiful landscapes. One of the main reasons I want to visit Japan is because of its culture. Japan has a very different culture from my country, and I am very interested in learning about it. I want to see the traditional Japanese ceremonies and try wearing a kimono. I also want to learn about Japanese tea ceremonies because they are very famous. Another reason I want to visit Japan is because of its beautiful landscapes. Japan has many beautiful mountains and forests that I would like to explore. I have seen pictures of Mount Fuji, and I think it is very beautiful. I would like to go hiking there and see the cherry blossoms in spring. In conclusion, I want to visit Japan because it has a unique culture and beautiful landscapes. I think it would be a very interesting and enjoyable trip.

Table 8: Sample essays whose rubric scores on content, organization, and language are all 1 out of 5. Sample-level analysis on *content* (orange), *organization* (blue), and *language* (red).

sitional signals, a thesis sentence, and supporting examples. The essay also includes a few grammatical errors and awkward phrases (red text), as it is written by EFL learners in a real-world classroom.

6 Conclusion

We release the DREsS, a large-scale, standard rubric-based essay scoring dataset with three subsets: DREsS$_{\text{New}}$, DREsS$_{\text{Std.}}$, and DREsS$_{\text{CASE}}$. DREsS$_{\text{New}}$ is the first reliable AES dataset with 2.3K 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 DREsS$_{\text{Std.}}$. We finally suggest CASE, corruption-based augmentation strategies for essays, which generates 40.1K 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 in other languages as well. We leave that as future work.

DREsS$_{\text{New}}$ 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. Also, an optimized rationale (*e.g.,* a threshold in corruption, corruption scale) will advance CASE, which we leave for future work.

We acknowledge that the experimental results in Table 3-4 might not fully cover state-of-the-art models in AES. Nonetheless, it is noteworthy that those results are a *baseline* for our dataset. We emphasize that the core contribution of this paper is the construction and the public release of a large-scale AES dataset (DREsS), not a proposal for AES model architecture. We believe nine different models—namely, state-of-the-art AES-specialized models (EASE, NPCR, ArTS), LLMs (GPT-4o, Llama 3.1, GPT-NeoX), and transformer-based models with different input sizes (BERT, Longformer, BigBird)—sufficiently cover empirical testing of existing models. We leave examining state-of-the-art AES models for future work, with a proposal of and comparison to a novel architecture.

Ethics Statement

All studies in this research project were conducted with the approval of our institutional review board (IRB). Annotators were fairly compensated (approximately USD 18), which exceeds the minimum wage in the Republic of Korea in 2024 (approximately USD 7.3). 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.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government (MSIP) (No. RS-2024-00443251, Accurate and Safe Multimodal, Multilingual Personalized AI Tutors). This research project has benefited from the Microsoft Accelerate Foundation Models Research (AFMR) grant program through which leading foundation models hosted by Microsoft Azure along with access to Azure credits were provided to conduct the research.

References

Mohammad Arsyad Arrafii. 2020. Grades and grade inflation: exploring teachers' grading practices in Indonesian EFL secondary school classrooms. *Pedagogy, Culture & Society*, 28(3):477–499.

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. *arXiv preprint arXiv:2004.05150*.

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.

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.

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.

Heejin Do, Yunsu Kim, and Gary Lee. 2024. 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.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal

Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata, and Mamoru Komachi. 2020. Automated essay scoring system for nonnative Japanese learners. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1250–1257, Marseille, France. European Language Resources Association.

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.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida.

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.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

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).

Wei Song, Kai Zhang, Ruiji Fu, Lizhen Liu, Ting Liu, and Miaomiao Cheng. 2020. Multi-stage pre-training for automated Chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6723–6733, Online. Association for Computational Linguistics.

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.

Sara Cushing Weigle. 2002. *Assessing Writing*, volume 1. Cambridge University Press, Englewood Cliffs, NJ.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. 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

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

Table 9: SFT configuration

We split $\text{DREsS}_{\text{New}}$ into training, validation, and test sets in a 6:2:2 ratio with a random seed of 22. We use $\text{DREsS}_{\text{Std.}}$ and $\text{DREsS}_{\text{CASE}}$, a unified or augmented data as training data only. Additionally, we separate the training, validation, and test set first and then apply CASE in Table 3. In other words, training data does not include augmented essays from high-quality essays in the test set, which prevents data leakage. The AES experiments except for ArTS, GPT-NeoX-20B, and Llama 3.1 (8B) in Table 4 were conducted under GeForce RTX 2080 Ti (4 GPUs), 256GiB system memory, and Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz (40 CPU cores) with hyperparameters denoted in Table 9. Fine-tuning ArTS, GPT-NeoX-20B, and Llama 3.1 (8B) was conducted under Quadro RTX 8000 (4 GPUs), 377GiB system memory, and Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz (48 CPU cores) with the same hyperparameters. LLM inference uses greedy decoding (*i.e.,* temperature 0.0).

B LLM Prompting

This section provides detailed system prompts used for the experiments in this paper.

B.1 Automated Essay Scoring

Table 10 illustrates four different system prompts used in experiments for Table 4.

B.2 Synthetic Essay Generation

> You are an English as a foreign language (EFL) learner taking an English writing course in a college for students who get TOEFL scores ranging from 15 to 21.
>
> ### Examples 1–5: `<five pairs of writing prompts and EFL student's essays>`
> ### Scoring criteria: `<three rubrics explanation>`
>
> Write an essay with short paragraphs about the given prompt, of which scores are `<score>` out of 5.0 for all criteria. Note that the essay should include erroneous patterns or typos from EFL students, according to the score.
>
> ### Essay prompt: `<essay_prompt>`

C Rationale Behind Standardizing

The weights are not arbitrarily chosen but were determined through expert consultation and theoretical considerations. Specifically, 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. Stylistic features, such as tone, coherence, and voice, are emphasized as higher-order concerns in writing assessment frameworks, while conventions, such as grammar and punctuation, are considered lower-order concerns. This theoretical understanding, combined with consultation with EFL education experts, informs our decision to assign a higher weight to style, particularly for argumentative essays where persuasive and expressive abilities are crucial (Weigle, 2002). For *content* and *organization*, we utilize the existing data rubric (idea for content, organization as same) in the dataset. 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.

D Additional Experimental Results

We rigorously investigate the efficacy of CASE

(A)	Please score the essay with three rubrics: content, organization, and language. ### Answer format: {content: Float, organization: Float, language: Float} Note that the float values of scores are within [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]. Please answer only in the above JSON format. ### prompt: `<essay prompt>` ### essay: `<student's essay>`
(B)	Please score the essay with three rubrics: content, organization, and language. ### Answer format: {content: Float, organization: Float, language: Float} Note that the float values of scores are within [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]. Please answer only in the above JSON format. ### Examples 1–5: ### prompt: `<essay prompt>` ### essay: `<student's essay>`
(C)	Please score the essay with three rubrics: content, organization, and language. `<three rubrics explanation>` ### Answer format: {content: Float, organization: Float, language: Float} Note that the float values of scores are within [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]. Please answer only in the above JSON format. ### prompt: `<essay prompt>` ### essay: `<student's essay`
(D)	Please score the essay with three rubrics: content, organization, and language. ### Answer format: {content: Float, organization: Float, language: Float, content_feedback: String, organization_feedback: String, language_feedback: String} Note that the float values of scores are within [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0]. Please answer only in the above JSON format, with feedback. ### prompt: `<essay prompt>` ### essay: `<student's essay>`

Table 10: Four different prompts for `gpt-4o` to get rubric-based scores in the last four rows of Table 4

	Content	Organization	Language	Total
SFT w/ DREsS$_{\text{New}}$	0.411	0.375	0.425	0.464
+ DREsS$_{\text{CASE}}$	0.634	0.780	0.588	0.692
+ `gpt-4o`	0.452	0.377	0.408	0.467

Table 11: Experimental results of augmentation techniques in AES models with the identical training steps

in training AES models by conducting a more controlled experiment using Llama 3.1 (8B) as a foundation model for supervised fine-tuning (SFT). We fine-tune the models with DREsS$_{\text{New}}$, DREsS$_{\text{CASE}}$, and synthetic data generated by `gpt-4o`. Differing from the main experiments (Table 3–5), Table 11 shows the experimental results where the training steps as 5,000 to ensure that the number of training samples is identical. Notably, adding synthetic data generated by `gpt-4o` for fine-tuning shows a minimal impact, especially achieving the worst performance in Language. Aligning with the findings in §5.2, generative methods are not applicable for essay augmentations.

E Datasheet for Dataset

In this section, we document DREsS following the format of Datasheets for Datasets (Gebru et al., 2021). The details on the composition and the collection process of the CSRT dataset are described in the main text.

E.1 Motivation

1. **For what purpose was the dataset created?** We aim to construct a large-scale, standard, rubric-based dataset for automated essay scoring (AES) to build AES systems that meet the needs of both instructors and students.
2. **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** The authors constructed DREsS by 1) collecting new essays and scores from the writing courses in their institution, 2) standardizing existing works, and 3) synthesizing new samples.
3. **Who funded the creation of the dataset?** See the Acknowledgments and Disclosure of Funding section.

E.2 Preprocessing/cleaning/labeling

1. **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** No. Instead, we conduct rule-based post-processing and human inspection to filter out sensitive information.
2. **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** N/A
3. **Is the software that was used to preprocess/clean/label the data available?** N/A

E.3 Uses

1. **Has the dataset been used for any tasks already?** No.
2. **Is there a repository that links to any or all papers or systems that use the dataset?** N/A
3. **What (other) tasks could the dataset be used for?** DREsS can be used as a training and evaluation dataset for automated essay scoring tasks.

E.4 Distribution

1. **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** Yes, the dataset is open to the public who submitted a consent form.
2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** The dataset will be distributed through our website.
3. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** The dataset will be distributed under the MIT license.
4. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** No.
5. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.

E.5 Maintenance

1. **Who will be supporting/hosting/maintaining the dataset?** The authors of this paper will maintain DREsS.
2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** The owner/curator/manager(s) of the dataset are the authors of this paper. They can be contacted through the emails on the first page of the main text.
3. **Is there an erratum?** We will release an erratum at the GitHub repository if errors are found in the future.
4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** Yes, the dataset will be updated whenever it can be extended to other red-teaming benchmarks. These updates will be posted on the main web page for the dataset.
5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** N/A
6. **Will older versions of the dataset continue to be supported/hosted/maintained?** Yes.

7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** NO.