ICLE++: Modeling Fine-Grained Traits for Holistic Essay Scoring

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

Abstract

The majority of the recently-developed models for automated essay scoring (AES) are evaluated solely on the ASAP corpus. However, ASAP is not without its limitations. For instance, it is not clear whether models trained on ASAP can generalize well when evaluated on other corpora. In light of these limitations, we introduce ICLE++, a corpus of persuasive student essays annotated with both holistic scores and trait-specific scores. Not only can ICLE++ be used to test the generalizability of AES models trained on ASAP, but it can also facilitate the evaluation of models developed for newer AES problems such as multi-trait scoring and cross-prompt scoring. We believe that ICLE++, which represents a culmination of our long-term effort in annotating the essays in the ICLE corpus, contributes to the set of much-needed annotated corpora for AES research.

1 Introduction

The past decade has seen considerable progress on automated essay scoring (AES), the task of automatically assigning a (holistic) score to an essay that summarizes its overall quality. The reason can in part be attributed to the public availability of annotated AES corpora where each essay is manually annotated with its holistic score. These corpora have facilitated the development and evaluation of AES models and enabled easy tracking of progress.

While several AES corpora have been developed over the years (including English corpora such as CLC-FCE (Yannakoudakis et al., 2011), ASAP[1], TOEFL11 (Blanchard et al., 2013), as well as corpora in other languages such as MERLIN[2] (multilingual), Ostling's (2013) Swedish corpus, and Horbach et al.'s (2017) German corpus), the vast majority of recently-developed AES models have been evaluated solely on the ASAP corpus (Taghipour and Ng, 2016; Uto et al., 2020; Ridley et al., 2021; Kumar et al., 2022; Chen and Li, 2023). ASAP is released as part of a Kaggle competition in 2012. While for the time being it may be fine to focus the evaluation of AES models on ASAP, continuing to do so may not be ideal for AES research in the long run, for the following reasons:

Generalizability. ASAP is composed of essays written by U.S. students between grade 7 and grade 10. A natural question, then, is: will models trained on ASAP perform well on other AES corpora? For instance, since the ASAP essays are written by native speakers of English, can the resulting models perform well on the TOEFL essays, which are written by learners of English as a second language? In addition, essay length (as measured by the number of words in the essay) has been found to be a confounding variable for holistic scoring in ASAP and other AES corpora where essays are written in a *time-restricted* setting (such as a test setting), so it is not clear how models trained on ASAP would perform if they are applied to corpora where essays are written in a *time-unrestricted* setting, such as essays written as part of a homework assignment with a length restriction (e.g., between 500 and 600 words), as length may no longer be a confounding variable in these AES corpora.

Limited feedback. While the performance of AES models has been increasing steadily over the years, these models provide little feedback on what aspects of an essay need improvement if it is assigned a low score by an AES model. As is commonly known, a holistic score is dependent on a number of *trait-specific* scores (i.e., scores along different dimensions of essay quality such as ORGANIZATION and PROMPT ADHERENCE). Hence, if an AES model can provide trait-specific scores in addition to holistic scores, the trait-specific scores can serve as feedback for students on which aspects of their essays need improvement.

[1] https://www.kaggle.com/c/asap-aes
[2] https://www.merlin-platform.eu/

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8465–8486
June 16-21, 2024 ©2024 Association for Computational Linguistics

While early heuristic-based AES models such as *e-rater* (Attali and Burstein, 2006) compute the holistic score of an essay as a weighted sum of heuristically computed trait-specific scores (and hence it is straightforward to understand which traits need improvement if the holistic score is low), the traits are primarily restricted to those that are *non-content-based* (i.e., traits that can be scored without an understanding of an essay's content, such as ORGANIZATION). Given the difficulty associated with trait-specific scoring (particularly the scoring of content-based traits such as ARGUMENT PERSUASIVENESS), few learning-based AES models explicitly exploit traits.

Nevertheless, research on trait-specific scoring, including the scoring of content-based traits, exists. For instance, we previously worked on learning-based trait-specific scoring (Persing et al., 2010; Persing and Ng, 2013, 2014, 2015), but did not study the impact of the trait-specific scores on holistic scoring. ASAP++ (Mathias and Bhattacharyya, 2018), an extension of ASAP where each essay is annotated with trait-specific scores, has facilitated the development of AES models in recent years that allow trait-specific scores to be predicted jointly with holistic scores. However, the traits that are being scored in ASAP++ are arguably too coarse-grained. Specifically, while ASAP++ provides scores for all non-content-based traits, all content-based traits (e.g., COHERENCE, PERSUASIVENESS, THESIS CLARITY) are lumped into a single trait that they call CONTENT. These coarse-grained traits not only severely limit the kind of content-based feedback that can be provided to an essay's author, but could prevent an AES model from mimicking the human essay scoring process, where different content-based traits are considered separately.[3]

Appropriateness for cross-prompt scoring. AES researchers have traditionally focused on *within-prompt* scoring, where models trained on essays written for a given prompt are applied to essays written for the same prompt. However, within-prompt scoring may not be a practical setting, as an AES model typically does not perform well on essays written for a new prompt unless it is (re)trained on essays from the new prompt. Consequently, researchers have begun investigating a new, challenging evaluation setting known as *cross-prompt* scoring: given training essays written for a set of prompts, the goal is to train a model to score essays written for prompts not seen during training.

ASAP has been used for evaluating cross-prompt models. Recall that ASAP is composed of eight prompts, including two for persuasive essays, two for narrative essays, and four for source-dependent essays. Cross-prompt evaluation is typically conducted via leave-one-prompt-out cross-validation experiments, where in each iteration essays from exactly one prompt are used for testing and the remaining ones are used for training. This implies, for instance, that a cross-prompt model that is applied to score the narrative essays from one of the prompts has been trained on not only narrative essays but also persuasive and source-dependent essays. This is a rather strange setup: since what constitutes a good persuasive essay may not be the same as what constitutes a good narrative essay (as they are scored using different rubrics), it is not clear whether it even makes sense to perform cross-prompt scoring when the training essays do not have the same type as the test essays.

In light of the above discussion, we introduce ICLE++, a corpus of persuasive essays written for 10 prompts where each essay is annotated with not only its holistic score but also its trait-specific scores. ICLE++ represents a culmination of our long-term effort in annotating the persuasive essays in the ICLE corpus that was initiated in the fall of 2009. These essays are written by university undergraduates from 16 countries who are learners of English as a foreign language. Hence, ICLE++ complements well with ASAP, where essays are written by pre-college students who are native speakers of English, enabling the evaluation of the generalizability of AES models. Unlike many existing AES corpora, essay length is no longer a confounding variable in the scoring process for ICLE++, as the vast majority of essays are between 500 and 600 words. In addition, since all essays are persuasive, ICLE++ provides a natural setup for cross-prompt scoring, where one can determine whether the knowledge acquired from the persuasive essays written for the training prompts would be useful for scoring the persuasive essays written for a new prompt. Above all, we identify a set of 10 traits that humans use when scoring persuasive essays. By scoring essays with these 10 traits, ICLE++ enables us to gain a deeper understanding of the human essay scoring process, specifically by determining the relative importance that a human puts

[3] For the sake of fairness, we should mention that the traits in ASAP++ are designed with the goal of developing models for scoring multiple traits rather than providing feedback.

Score	Description
4	essay provides ample support for its claims, uses effective vocabulary and sentence variety, organizes ideas logically, develops them well, conveys them concisely, and connects them with smooth transitions
3.5	essay offers generally sufficient support for its claims, uses appropriate vocabulary and sentence variety, organizes ideas logically, develops them well, conveys them concisely, and connects them with appropriate transitions
3	essay supports its claims adequately but the support may not be even, develops and organizes ideas reasonably well but the transitions between ideas may not be smooth, shows adequate command of language to convey ideas clearly
2.5	essay offers support that is of little relevance to its claims, provides limited logical development and organization of ideas, shows problems in language, grammar and/or sentence structure that result in a lack of clarity
2	essay provides little or no relevant support for its claims, has ideas that are not developed, illogical, and/or poorly organized, and has problems in language, grammar and/or sentence structure that often obscure meaning
1.5	essay provides little or no support for its claims, has disorganized ideas, is overly short, and has enough problems in language, grammar and/or sentence structure that make the essay nearly impossible to understand
1	essay is off topic

Table 1: Description of each Overall Quality score.

on each trait when scoring an essay holistically. We believe that ICLE++ would be a valuable resource for AES researchers. Corpus development for AES research has not been able to keep up with model development, so ICLE++ can contribute to the set of much-needed annotated AES corpora.

2 Corpus

Our corpus, ICLE++, is composed of essays selected from the International Corpus of Learner English (Granger et al., 2009), which consists of more than 6000 essays on a variety of writing topics written by university undergraduates from 16 countries and 16 native languages who are learners of English as a foreign language. 91% of the essays are persuasive. To ensure representation across the native languages of the authors, we selected mostly essays written in response to topics that are well-represented in multiple languages. This avoids many issues that may arise when certain vocabulary is used in response to a particular topic for which essays written by authors from only a few languages are available. With this criterion in mind, we selected 1008 persuasive essays for annotation. These essays are written for 10 prompts (see Appendix A) and have eight paragraphs, 32 sentences, and 582 tokens on average. Thirteen native languages are represented in these essays.

2.1 Annotation Scheme

Overall Quality (i.e., Holistic) scoring. To score the overall quality of an essay, we develop a rubric that is inspired by the one used for the GRE Argument task. We evaluate OVERALL QUALITY using a numerical score from 1 to 4 in half-point increments (for a total of seven possible scores), with a score of 4 indicating a high-quality essay and a score of 1 indicating a poorly-written essay. A description of each score can be found in the rubric shown in Table 1.

Trait scoring. In consultation with the instructors of the freshman writing course at our institution, we identify 10 traits that could impact an essay's overall quality, as described below.

PROMPT ADHERENCE concerns how related the content of an essay is to the prompt for which it is written. THESIS CLARITY refers to how clearly an author explains the thesis of her essay. ARGUMENT PERSUASIVENESS concerns the convincingness of the argument an essay makes for its thesis. DEVELOPMENT concerns whether the essay develops its main ideas with adequate elaboration and examples. COHERENCE measures whether the essay demonstrates appropriate transition between *ideas*. COHESION measures whether the essay contains appropriate words and phrases between segments. ORGANIZATION concerns how well-structured the essay is. SENTENCE STRUCTURE concerns whether the essay shows appropriate complexity and variety in sentence structure. VOCABULARY measures whether the essay shows appropriate word choice and contains advanced vocabulary. Finally, TECHNICAL QUALITY concerns whether the essay uses correct grammar, mechanics, spelling, and punctuation.

The rubrics for scoring these traits, which we designed in consultation with the instructors of the writing course at our institution, are shown in Appendix B.[4] Each trait is evaluated using a numerical

[4] As noted before, this work is a continuation of our ongoing effort in annotating the ICLE essays. In particular, we have investigated some of these traits in previous work, including ORGANIZATION (Persing et al., 2010), THESIS CLARITY (Persing and Ng, 2013), PROMPT ADHERENCE (Persing and Ng, 2014), and ARGUMENT PERSUASIVENESS (Persing and Ng, 2015). For these traits, the rubrics shown in Appendix B are therefore the same as those described in our previous work.

ICLE++ Traits	ASAP++ Traits
Prompt Adherence	–
Thesis Clarity	Content
Persuasiveness	Content
Development	Content
Organization	Organization
Coherence	Content
Cohesion	Sentence Fluency
Sentence Structure	Sentence Fluency
Vocabulary	Word Choice
Technical Quality	Conventions

Table 2: Mapping from ICLE++ traits to ASAP++ traits.

score from 1 to 4 at half-point increments. Higher scores imply higher qualities w.r.t. a given trait.

Comparison with the ASAP++ traits. One of the motivations behind our work is that the traits used in ASAP++ are too coarse-grained. Below we describe the differences between our traits and those used in ASAP++.

In ASAP++, each persuasive essay is scored along five traits, namely, Content, Word Choice, Organization, Sentence Fluency, Conventions. Table 2 shows the mapping from the ICLE++ traits to the ASAP++ traits. As can be seen, there is a one-to-one correspondence between three of the traits in ICLE++ and ASAP++, all of which are non-content-based. Two of the other non-content-based traits in ICLE++, Cohesion and Sentence Structure, can be mapped to the Sentence Fluency trait in ASAP++. Prompt Adherence is a trait used in ICLE++, but for some unknown reasons, ASAP++ uses it when scoring source-dependent but not persuasive essays. Finally, four content-based traits in ICLE++ (Thesis Clarity, Argument Persuasiveness, Development, and Coherence) are lumped into a single trait in ASAP++, Content.

Employing a *composite* trait like Content could limit the amount of feedback provided to essay writers. For example, if an essay has a low Content score, its author may not know whether the poor score can be attributed to the use of unpersuasive arguments or the poor development of ideas or both. In fact, there is no Content score (according to its rubric) that can be assigned to essays that are strong in Development and Coherence but weak in Thesis Clarity and Argument Persuasiveness.

2.2 Annotation Procedure

Our annotators were undergraduate students selected from over 30 applicants. These applicants attended a training session taught by one of the aforementioned writing instructors, during which the instructor provided an overview of the goals of this research, defined the 10 traits we identified earlier, familiarized them with the scoring rubrics, and used select essays to illustrate how they should be annotated. For example, the annotators were told that the 10 traits should be scored before Overall Quality. After the sessions, the applicants were given sample essays to score (not included in our dataset). The six who were most consistent with the expected scores were hired as our annotators. Another session was held for these annotators to discuss the mistakes they made on the sample essays.

Trait	1	1.5	2	2.5	3	3.5	4
Overall Qual.	1	15	69	279	369	254	19
Prompt Ad.	3	3	41	88	233	216	422
Thesis Clarity	0	11	78	127	299	226	265
Persuasive.	5	16	139	307	412	111	16
Development	5	4	69	241	521	146	20
Organization	8	9	99	229	494	150	17
Coherence	1	6	58	140	496	220	85
Cohesion	0	7	86	278	493	130	12
Sent. Struct.	0	7	46	206	510	210	27
Vocabulary	1	2	35	183	534	173	78
Tech. Quality	1	8	77	196	475	213	36

Table 3: Distributions of the human-annotated scores.

To ensure consistency in scoring, each essay in ICLE++ was graded by two different annotators. Discrepancies were resolved through open discussion. The distributions of scores for Overall Quality and the traits are shown in Table 3.

2.3 Inter-Annotator Agreement

We measure agreement on the annotation of Overall Quality and each of the traits for the ICLE++ essays using Krippendorff's α (Krippendorff, 2004). For the sake of completeness, we also report agreement in terms of Quadratic Weighted Kappa (QWK)[5], which is a standard metric used for evaluating holistic essay scoring systems. As we will see shortly, the agreement scores computed w.r.t. these two metrics exhibit similar trends. This is perhaps not surprising since both of them distinguish far misses from near misses when computing agreement. For this reason, we will focus our discussion below on the Krippendorff's α values.

Agreement results are shown in Table 4. W.r.t. α, all traits exhibit an agreement above 0.6, showing a correlation more significant than random

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

chance. Overall Quality has an agreement of 0.755, which suggests substantial agreement. The traits that have the highest α values are Cohesion (0.668) and Organization (0.661), whereas the one that has the lowest α value is Coherence (0.602). This is perhaps not surprising. In general, humans tend to agree on what constitutes a well-organized essay (e.g., in a 5-paragraph essay, there is an introductory paragraph, followed by three paragraphs each of which makes a unique point, and a conclusion) and whether cohesive devices (e.g., discourse connectives) are used to make a piece of text cohesive. In contrast, determining how coherent the ideas are in an essay often requires subjective interpretation that may be biased by the annotator's background.

To gain insights into how inter-annotator agreement can be improved, we show in the last three columns of Table 4 (1) the fraction of essays that received the same score by both annotators; (2) the fraction of essays for which the two annotators' scores differ by exactly 0.5; and (3) the fraction of essays for which the scores differ by more than 0.5.

Across all traits, 31–40% of the essays have scores that differ exactly by 0.5. The disagreement in these essays stems primarily from allowing the annotators to use half points when scoring the traits. For example, if an essay is not good enough to be given a 3 (w.r.t. a particular trait) but deserves a score that is close to a 3, the annotators disagreed on whether its score should be 3 or 2.5. Some annotators thought that it did not meet the "3" level, so the score should be 2.5, while others thought that its quality was closer to "3" than "2.5" and should therefore be given a score of 3. After further discussion, the annotators agreed that in such cases the essay should be given the closest possible score, which means that "3" is what should be assigned to the essay in our example. The annotators commented that having more labeled examples during the annotator training process, as well as providing a description of each half-point score (i.e., 1.5, 2.5, and 3.5) will likely improve annotator agreement.

For all but two traits (Prompt Adherence and Thesis Clarity), only 6–12% of the essays have annotator scores that differ by more than 0.5. A discussion with the annotators reveals why the percentages are higher for these two traits. For Thesis Clarity, the annotators disagreed on whether a thesis that is not explicitly stated but is clear from the essay should be given a high score. After discussion, the annotators agreed that a clear, implicit thesis should receive a high score. For Prompt Adherence, the annotators disagreed on how to handle *multi-component* prompts. For example, the prompt "The prison system is outdated. We should not punish criminals." is composed of two parts that correspond to the two sentences in the prompt. Some annotators assigned a high score to essays as long as the entire essay adheres to one of the components, while others assigned a high score only if the essay addresses all and only the components. After discussion, the annotators agreed to employ the latter definition.

Trait	α	QWK	Same score	0.5 Diff	> 0.5 Diff
Overall	.755	.805	.588	.360	.052
Prompt Ad.	.615	.674	.494	.314	.192
Thesis Clarity	.631	.690	.448	.342	.210
Persuasive.	.655	.711	.523	.379	.098
Development	.618	.674	.566	.344	.090
Organization	.661	.714	.573	.338	.089
Coherence	.602	.660	.481	.403	.117
Cohesion	.668	.723	.575	.359	.066
Sent. Struct.	.633	.689	.583	.349	.068
Vocabulary	.655	.712	.561	.349	.090
Tech. Quality	.641	.698	.521	.381	.098

Table 4: Inter-annotator agreement results for each trait.

2.4 Analysis of Annotations

In this subsection, we conduct several experiments in order to gain insights into our annotations.

Correlation between Overall Quality and the traits. To understand whether the 10 traits are useful for predicting Overall Quality, we compute the Pearson Correlation Coefficient (PC) between Overall Quality and each trait. Results are shown in the first row of Table 5. As hypothesized earlier, all traits are positively correlated with Overall Quality. Though not shown in the table, all correlations are statistically significant at the $p < 0.001$ level. Given that these are persuasive essays, it should not be surprising that the trait that has the highest correlation with Overall Quality is Argument Persuasiveness (PC = 0.718). This is followed by Development (0.681) and Coherence (0.625), both of which are concerned with ideas: how well the ideas are developed and how smooth the transitions are between them. The trait that has the lowest correlation with Overall Quality is Thesis Clarity (0.497). This is somewhat unexpected, as intuitively an unclear thesis would have an adverse impact on the persuasiveness of the argument the essay makes, which would in turn lower Over-

	Prompt Adhere.	Thesis Clarity	Persuasiveness	Development	Organization	Coherence	Cohesion	Sent. Struct.	Vocabulary	Tech. Quality
Overall Qual.	0.520	0.497	0.718	0.681	0.589	0.625	0.526	0.547	0.558	0.571
Prompt Ad.		0.526	0.531	0.454	0.391	0.406	0.335	0.303	0.295	0.276
Thesis Clarity			0.516	0.412	0.410	0.417	0.286	0.313	0.325	0.284
Persuasive.				0.687	0.552	0.578	0.444	0.430	0.451	0.444
Development					0.547	0.575	0.441	0.451	0.461	0.464
Organization						0.535	0.469	0.378	0.385	0.342
Coherence							0.481	0.485	0.520	0.572
Cohesion								0.438	0.425	0.441
Sent. Struct.									0.599	0.591
Vocabulary										0.660

Table 5: Pearson Correlation values.

Trait	**Weight**
Prompt Adherence	0.076675
Thesis Clarity	0.047988
Persuasiveness	0.260711
Development	0.190239
Organization	0.120040
Coherence	0.063343
Cohesion	0.091381
Sentence Structure	0.097288
Vocabulary	0.075398
Technical Quality	0.130533
(Bias)	−0.468411

Table 6: Feature weights obtained by training a linear regressor on these traits to predict Overall Quality.

ASAP		**ICLE++**	
Feature	**PC**	**Feature**	**PC**
wordtypes	0.694	Coleman-Liau	0.310
complex_words_dc	0.653	mean_word	0.307
sentences	0.647	sylls_per_word	0.305
sents_per_para	0.636	chars_per_word	0.301
long_words	0.623	FleschReadingEase	−0.296

Table 7: Features having the highest PC values with Overall Quality in ASAP and ICLE++.

all Quality. Additional analysis is needed to determine the reason.

Correlation among the essay traits. Next, to gain insights into whether (and how) the 10 traits are correlated with each other, we compute the PC score for each pair of traits. Results are shown in Table 5. Though not shown in the table, all correlations are statistically significant at $p < 0.001$. As we can see, many correlations are relatively weak. The weak correlations are consistent with our intuition that these traits capture different aspects of essay quality. Nevertheless, there are two exceptions. First, Persuasiveness, Development, and Coherence seem to have a fairly strong correlation with each other. This is perhaps not a coincidence: as we can see in row 1 of Table 5, these are also the traits that have the strongest correlation with Overall Quality. Another group of traits that exhibits a somewhat strong correlation among themselves is composed of Sentence Structure, Vocabulary, and Technical Quality, all of which are low-level traits that are non-content-based. These results suggest that doing a poor job in scoring one of these traits will likely affect the scoring of the other two.

Trait importance. Next, we address the question of the relative importance of the traits in predicting Overall Quality. To answer this question, we train a linear regressor using the scikit-learn package[6] on all 1008 essays and examine the feature weight learned by the regressor for each trait, as a trait with a higher absolute weight implies a higher impact on Overall Quality scoring.

The feature weights and the bias term are shown in Table 6. Comparing Tables 5 and 6, we can see that the traits that have high PC values with Overall Quality also tend to be assigned large feature weights. Specifically, Argument Persuasiveness and Development, which are the traits with the highest correlations with Overall Quality, are also the traits that have the largest weights. Thesis Clarity, which has the lowest correlation with Overall Quality, is also the trait that has the smallest weight.

Comparison with ASAP. To shed some light on the differences between ASAP and ICLE++ as far as holistic scoring is concerned, we take the 86 (real-valued) prompt-independent features used in Chen and Li's (2023) AES model, PMAES, and compute the PC value between each feature and Overall Quality on ASAP and ICLE++.

Table 7 shows the five features that have the largest (positive or negative) PC values with Overall Quality for ASAP (left) and ICLE++ (right). A few points deserve mention. First, the highest-ranked features for ASAP are different from those

[6] https://scikit-learn.org/

[Prompt]

The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them

[Essay]

Is the prison system really outdated? For me it seems to be a doubtful statement . On the one hand it sounds very seductive: the freedom and equality for all human beings in the world. But this confirmation appears to be a rather turgid theory when one applies it to practice. I mean to say that when we come across a real crime, a subconscious voice immediately appeals to our common sense: “Criminals must be punished”.

But both of these notions prove to be superficial in practice. And it wouldn’t be mistaken to say that any particular case requires special considering. Different circumstances may bring a person to committing crime. Some people are put in such living (or other) conditions that crime is the only way-out when survival is threatened. And then the question which is bound to arise is, who is to be punished: the criminal or the people (or maybe the government), that put him into such position. But certainly I am not going to decide global problems here, to criticise the government or to give them my recommendations, as I hope that there exist a number of people, who (being experts in this field) are much better at this kind of activity, than I am.

Sometimes even one meeting with experienced psychologist may prevent a crime. And I have a strong belief that juvenile offenders should never be sent to prison or some other institutions of that kind. Staying among other criminals makes irreparable harm to the young people.

But in spite of these human ideas, considering human psychology, we should keep in mind that total freedom from punishment can bring the mankind unforeseen consequences. There should exist some restraining “device” in the world. Here prisons can be compared with the nuclear weapon: it is a property of civilised countries, which is not being used, but at the present time this is the only tool by means of which World Wars are prevented. Unfortunately people haven’t yet invented anything less perilous.

Turning to the fiction we will find no prison only in utopia, in that self-sufficient society where all people have everything they need, so there is absolutely no reason for crime.

To summarise what has been brooded over I need to say, that however sad it may sound, the prisons must not be abolished. But in every case of committing a crime there should be a much more careful personal approach to the criminal.

But the question keeps being open and it is up to the mankind to solve it.

Table 8: A sample essay.

for ICLE++,[7] suggesting that there are indeed differences between the essays in the two corpora such that models trained on one may not necessarily perform well on the other. Second, the PC values of the highest-ranked features for ASAP are considerably higher than those for ICLE++. In other words, these prompt-independent features appear to be more useful for scoring the ASAP essays than the ICLE++ essays, and a model that employs these features will likely achieve better results on ASAP than ICLE++.[8]

2.5 Sample Annotated Essay

To enable the reader to gain a better understanding of the 10 traits we use in ICLE++, we explain in this subsection how we annotate the sample essay shown in Table 8 using our annotation scheme.

According to our rubrics, the Prompt Adherence score of this essay is 4 because it consistently stays on topic (i.e., the prison system, and whether criminals should be punished). Its Thesis Clarity score is 4 because its thesis is clear: the prisoner system should not be abandoned. Its Vocabulary score is 4 because it shows appropriate word choice and contains advanced vocabulary (e.g., “subconscious”, “superficial”). Its Cohesion score is 3.5: it uses appropriate connectives between sentences, but begins a sentence with “and” and “but” a little too excessively. Its Technical Quality score is 3.5 because the essay’s readability is not affected by the occasional technical errors (e.g., missing articles). Its Sentence Structure score is 3.5 because the essay exhibits a variety of sentence structures, but can be improved if some of the sentences can be rewritten to have simpler structures. Its Cohesion score is 3.5 because for the most part, the essay contains sensible transitions between ideas and is easy to understand. Its Organization deserves a score of 3: while the essay is fairly well-structured, it can certainly benefit from some reorganization. For example, the thesis should be stated much earlier in the essay. Its Argument Persuasiveness only deserves a score of 2, however: the only argument it presents to support its claim that the prison system should be retained is that there needs to be a mechanism for punishing people in order to maintain stability, but it is not perceived as particularly persuasive. In fact, towards the end of the essay

[7] See Appendix C for a description of these features, though many feature names are self-explanatory. For example, “wordtypes” is the total number of unique words, whereas “Coleman-Liau” and “FleschReadingEase” are readability indices.

[8] The full list of features and their PC values with Overall Quality can be found in Appendix C.

the author tried to take a somewhat neural stance by saying that whether a criminal should be punished should be considered on a case by case basis, which somewhat weakens their own argument. Its DEVELOPMENT score is 2.5: not all ideas in the essay are developed with examples or illustrations. Its OVERALL QUALITY score is 2.5 because the support it offers for its claims is not particularly persuasive and has ideas that are not well-developed. As mentioned in Section 2.1, at this score level, the essay should fare reasonably poorly for most, if not all, of the traits. Hence, even though the essay fares well on some of the non-content-based traits such as VOCABULARY, SENTENCE STRUCTURE and COHESION, this information cannot be reflected in the OVERALL QUALITY score.

In addition, recall that ASAP++ lumps all the content-based traits into a single trait CONTENT. Using a CONTENT score makes it impossible to reflect that the essay fares well on some content-based traits, including THESIS CLARITY, but poorly on other content-based traits, such as DEVELOPMENT and ARGUMENT PERSUASIVENESS.

2.6 Experiments

Next, we gauge the performance of a set of AES models on ICLE++ for both holistic scoring and trait scoring. These AES models have achieved state-of-the-art results on ASAP. For comparison purposes, we also show the results of these models on ASAP by re-running them on ASAP.

2.6.1 Experimental Setup

Since our (regression-based) models output real values, we round the outputs of each model (for both OVERALL QUALITY and trait scoring) to the nearest of the seven possible reference scores (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0) before applying our evaluation metrics, which we describe below.

Evaluation metric. We employ Quadratic Weighted Kappa (QWK) as our metric for scoring both OVERALL QUALITY and the traits.[9] Since QWK is an agreement metric, higher values are better.

Evaluation settings. We conduct experiments under two settings. In *within-prompt* scoring, we follow previous work (e.g., Taghipour and Ng (2016)) and partition the available essays into k folds such that (1) each fold contains approximately the same number of essays and (2) the distribution of essays over prompts remains more or less the same across different folds. We set k to 5 for ASAP and 10 for ICLE++, and conduct k-fold cross-validation experiments. In each fold experiment, we use one fold for testing, one fold for development, and the remaining folds for model training.

(a) Within-prompt scoring

Setting	Model	ASAP	ICLE
Without traits	Uto et al.	0.7601	0.3988
	Kumar et al.	0.6847	0.3073
With traits	Uto et al. (Simple)	0.7633	0.2839
	Uto et al. (Kumar)	0.7584	0.2776
	Kumar et al.	0.6899	0.3391
	Gold Traits	0.8799	0.8211

(b) Cross-prompt scoring

Setting	Model	ASAP	ICLE
Without traits	PMAES	0.5992	0.2509
With traits	PMAES	0.6095	0.2810
	Gold Traits	0.8345	0.8657

Table 9: Holistic scoring results.

In *cross-prompt* scoring, we partition the essays into folds by prompt, so each fold contains all and only those essays written for the same prompt. The essays in ASAP and ICLE++ are partitioned into eight folds and ten folds respectively (since there are eight prompts in ASAP and 10 prompts in ICLE++). Results are obtained by conducting leave-one-fold-out cross-validation experiments. For development, we reserve one fold for ASAP and three for ICLE++.

Regardless of which evaluation setting is used, the results we report in Tables 9 and 10 are results macro-averaged over all folds.[10]

Scoring with and without traits. For each of the aforementioned evaluation settings, we train two types of models that differ in terms of whether traits are involved in the training process. We refer to these two types of models as "scoring without traits" and "scoring with traits".

Models. For each of the two evaluation settings, we employ AES models that have achieved state-of-the-art results on ASAP in the respective setting.

For within-prompt scoring, we employ two models, namely Uto et al's model (2020) and Kumar et al.'s (2022) model. While Kumar et al.'s model already has two variants that can be used for scoring with and without traits (which correspond to their multi-task learning model and their single-task learning model respectively), Uto et al.'s model has

[9] For completeness, we also report results in terms of several other evaluation metrics. See Appendix D.1 for details.

[10] See Appendix D.2 for per-prompt results.

(a) Results on ASAP

	Content	Organization	Word Choice	Sent. Fluency	Conventions	Prompt Adhere.	Language	Narrativity
Uto et al. (Simple)	0.687	0.300	0.249	0.249	0.297	0.358	0.321	0.347
Uto et al. (Kumar)	0.644	0.254	0.242	0.237	0.258	0.341	0.303	0.323
Kumar et al.	0.612	0.511	0.552	0.571	0.473	0.682	0.593	0.652
PMAES	0.488	0.448	0.538	0.517	0.418	0.504	0.437	0.492

(b) Results on ICLE++

	Prompt Adhere.	Thesis Clarity	Persuasiveness	Development	Organization	Coherence	Cohesion	Sent. Struct.	Vocabulary	Tech. Quality
Uto et al. (Simple)	0.090	0.085	0.197	0.211	0.130	0.148	0.204	0.272	0.297	0.212
Uto et al. (Kumar)	0.082	0.057	0.160	0.215	0.060	0.146	0.141	0.142	0.231	0.220
Kumar et al.	0.189	0.041	0.192	0.241	0.144	0.156	0.234	0.400	0.318	0.346
PMAES	0.134	0.079	0.084	0.214	0.122	0.125	0.279	0.360	0.265	0.228

Table 10: Trait-specific scoring results.

only been developed for scoring without traits. For this reason, we develop two variants of Uto et al.'s model so that they can be used for scoring with traits. The first variant, Uto et al. (Simple), incorporates additional neurons into the output layer for joint prediction of trait-specific and holistic scores. The second variant, Uto et al. (Kumar), is motivated by Kumar et al.'s model. Specifically, we extend Uto et al.'s original model so that it first predicts the trait-specific scores, which are then used as features for predicting the holistic score. For cross-prompt scoring, we employ PMAES (Chen and Li, 2023) to train models for holistic scoring with and without traits.[11]

Finally, for both evaluation settings, we train a model that takes only the *gold* (i.e., human-annotated) trait-specific scores of an essay as input (i.e., without using the essay itself) and predicts its holistic score. This oracle experiment can provide an upper bound on the performance of holistic scoring with trait-specific scores.

2.7 Results and Discussion

Results of holistic scoring with and without traits on ASAP and ICLE++ for the two evaluation settings, which are expressed in terms of QWK, are shown in Table 9. A few points deserve mention. First, since the QWK scores on ASAP are considerably higher than those on ICLE++, these results suggest that ICLE++ is a more challenging corpus than ASAP. Second, the inclusion of traits does not always improves holistic scoring results: for within-prompt scoring, the results are mixed; but for cross-prompt scoring, the use of traits seems to have a generally positive impact on holistic scoring. Third, while cross-prompt scoring is generally thought to be more challenging than within-prompt scoring (and this is reflected in the ASAP results), on ICLE++ the cross-prompt results are similar to the within-prompt results. Finally, the much higher QWK scores achieved when gold traits are used suggest the usefulness of traits for holistic scoring.

Trait-specific scoring results, which are expressed in terms of QWK, are shown in Table 10. The first three rows of each subtable show within-prompt scoring results whereas the last row shows cross-prompt scoring results. A few points deserve mention. First, like the holistic scoring results, the trait scoring results on ASAP are generally higher than those on ICLE++. The poor trait scoring results on ICLE++ could explain why the use of traits has a negative effect on within-prompt holistic scoring on ICLE++ (see Table 9). Nevertheless, despite the poor trait scoring results, the use of traits slightly improves cross-prompt holistic scoring. A plausible reason could be attributed to the robustness of PMAES to the noisily predicted trait scores, but additional experiments are needed to determine the reason.

Overall, these results seem to suggest that ICLE++ presents new challenges to researchers.

3 Conclusion

We presented ICLE++, a corpus of persuasive essays annotated with both holistic scores and 10 fine-grained trait-specific scores. We believe that ICLE++ contributes to the much-needed set of annotated AES corpora and will be a valuable resource to AES researchers. We make all of our annotations publicly available.[12]

[11] An overview of each of these models as well as their implementation details can be found in Appendix E.

[12] https://github.com/samlee946/ICLE-PlusPlus

Limitations

We believe that our work has several limitations. First, since we focus only on persuasive essays, our findings are also limited to persuasive essays. Nevertheless, we believe that our framework can be applied to annotate other types of essays. Second, our corpus is composed of essays written by university undergraduates who are non-native speakers of English. It is not clear whether the conclusions we drew from our corpus can be generalized to essays written by high school students who are native speakers of English (e.g., the essays in the ASAP dataset), for instance.

Ethics Statement

Human annotator information. All annotators were undergraduate students aged around 18-22, and were hired as student workers with full consent. All annotators were native English speakers, including male and female students from different ethnic groups residing in the United States. Annotators were compensated with an hourly rate of 10 US dollars.

Intended use of the dataset. This dataset is intended for non-profit research purposes only.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No. We do not expect any risk to be posed by the user of this dataset. Neither do we expect any financial loss associated with its use.

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? We plan to release the labeled trait scores and holistic scores with unique identifiers pointing to the source essays on a GitHub repository with the MIT license.

Is this dataset consistent with the terms of use and the intellectual property and privacy rights of people? The source essays of this dataset were obtained from ICLE, which requires a license to access. So we will distribute our annotations but not the source essays. The license grants licensee usage for non-profit research purposes only, thus our usage is compatible with the original access conditions.

Acknowledgments

We thank the anonymous reviewers for their helpful and insightful comments on an earlier draft of the paper. We would also like to thank the people who have contributed to this long-term annotation effort in one way or another over the years, including Alan Davis, Isaac Persing, Zixuan Ke, Hrishikesh Inamdar, Hui Lin, Andrew Mallon, Andrew Hubbs, Karin Khoo, Jayne Koath, Christopher Maier, and Cory Thornton.

References

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

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain.

Andrea Horbach, Dirk Scholten-Akoun, Yuning Ding, and Torsten Zesch. 2017. Fine-grained essay scoring of a complex writing task for native speakers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 357–366.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. SAGE, Thousand Oaks, CA.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, CA, USA, May 6-9, 2019, Conference Track Proceedings*.

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).

Robert Östling, André Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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

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

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

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

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13745–13753.

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

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

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

A Statistics on Essay Prompts

In this section we provide statistics on the essay prompts in ICLE++ and ASAP.

Table 11 shows the ten essay prompts in ICLE++. For each prompt, we show the average number of words in the essays, the number of native languages covered, and the number of essays annotated. For comparison purposes, Table 12 shows the eight essay prompts in ASAP. For each prompt, we show the average number of words in the essays and the number of essays annotated.

Table 13 details the average scores for Overall Quality and for each of the traits over all the essays in ICLE++, complementing the information in Table 3 by offering further insights into the score distributions. Specifically, the "Avg" column shows the average scores over all prompts whereas each of the subsequent columns shows the average scores over one of the prompts. Note that prompt 1 in this table refers to the first prompt listed in Table 11, for instance.

B Trait-Specific Rubrics

In this section, we present the rubrics we use to annotate the 10 traits for each essay in ICLE++. The rubrics are shown in Tables 14 to 23. As can be seen, we evaluate each trait using a numerical score from 1 to 4 in half-point increments (for a total of seven possible scores), with a score of 4 indicating an essay that is of high-quality w.r.t. the trait under consideration and a score of 1 indicating an essay that is of low-quality w.r.t. the trait under consideration.

Prompt	**Avg. # Words**	**Languages**	**Essays**
Some people say that in our modern world, dominated by science and technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?	575.8	13	310
Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.	586.0	13	148
The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.	585.3	13	104
In the words of the old song: "Money is the root of all evil."	623.0	10	84
In his novel {\it Animal Farm}, George Orwell wrote "All men are equal but some are more equal than others." How true is this today?	579.2	10	82
Feminists have done more harm to the cause of women than good.	583.8	10	64
All armies should consist entirely of professional soldiers: there is no value in a system of military service.	564.8	10	62
Television is the opium of the masses in modern society. Discuss.	526.5	10	58
Most University degrees are theoretical and do not prepare us for the real life. Do you agree or disagree?	552.5	10	55
Crime does not pay.	579.0	10	39

Table 11: The 10 writing prompts in ICLE++.

Prompt	**Avg. # Words**	**Essays**
Write a letter to the editor of a newspaper about how computers affect society today.	365.4	1783
Write a letter to the editor of a newspaper about censorship in libraries	380.7	1800
Write a review about an article called Rough Rough Road by Joe Kurmaskie. The article will be provided.	108.5	1726
Explain why the author concludes the story the way the author did. The short story will be provided.	94.3	1772
Describe the mood created by the author in the memoir. Support your answer with relevant and specific information from the memoir	122.1	1805
Describe the difficulties that builders of the Empire State Building faced because of allowing dirigibles to dock there.	153.2	1800
Write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience.	167.6	1569
We all understand the benefits of laughter. For example, someone once said, "Laughter is the shortest distance between two people." Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.	604.7	723

Table 12: The eight writing prompts in ASAP.

Trait	**Avg.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Overall Quality	2.91	2.89	2.92	2.85	2.64	2.83	2.99	2.90	2.73	3.11	3.12
Prompt Adhere.	3.43	3.49	3.35	3.42	3.01	3.52	3.54	3.61	3.33	3.40	3.61
Thesis Clarity	3.22	3.15	3.24	3.23	2.88	3.31	3.32	3.26	3.10	3.19	3.29
Persuasiveness	2.75	2.75	2.73	2.65	2.47	2.66	2.81	2.81	2.61	2.84	3.03
Development	2.89	2.88	2.88	2.78	2.65	2.91	2.90	2.85	2.81	3.01	3.13
Organization	2.85	2.89	2.79	2.77	2.65	2.98	2.94	2.85	2.73	2.92	3.01
Coherence	3.06	3.06	3.05	2.96	2.85	3.05	3.09	3.06	2.88	3.26	3.23
Cohesion	2.84	2.91	2.82	2.84	2.65	2.91	2.79	2.82	2.81	2.95	2.89
Sent. Structure	2.97	2.94	3.03	2.84	2.67	2.74	3.01	3.01	2.99	3.15	3.05
Vocabulary	3.03	2.86	3.11	2.92	2.74	2.81	3.12	3.05	3.03	3.31	3.15
Tech. Quality	2.95	2.83	3.03	2.82	2.86	2.60	3.01	3.05	2.95	3.12	3.07

Table 13: The average scores for Overall Quality and the 10 traits in each writing prompt in ICLE++.

Score	Description
4	essay fully addresses the prompt and **consistently stays on topic**
3	essay mostly addresses the prompt or **occasionally wanders off topic**
2	essay does not fully address the prompt or **consistently wanders off topic**
1	essay does not address the prompt at all or is **completely off topic**

Table 14: Descriptions of the Prompt Adherence scores.

Score	Description
4	essay presents a **very clear thesis** and requires little or no clarification
3	essay presents a **moderately clear thesis** but could benefit from some clarification
2	essay presents an **unclear thesis** and would greatly benefit from further clarification
1	essay presents **no thesis of any kind** and it is difficult to see what the thesis could be

Table 15: Descriptions of the Thesis Clarity scores.

Score	Description
4	essay makes a **persuasive argument** for its thesis and would convince most readers
3	essay makes a **decent argument** for its thesis and could convince some readers
2	essay makes a **poor but understandable argument** for its thesis or sometimes even **argues against it**
1	essay **does not make an argument** or it is often **unclear what the argument is**

Table 16: Descriptions of the Argument Persuasiveness scores.

Score	Description
4	essay **fully develops its main ideas** with adequate elaboration and examples
3	essay **develops most of its ideas** but could benefit from further elaboration and examples
2	essay **does not fully develop its ideas** and would greatly benefit from further elaboration
1	essay **presents numerous undeveloped ideas** with almost no elaboration or examples

Table 17: Descriptions of the Development scores.

Score	Description
4	essay contains **sensible transitions between ideas** and is usually very understandable
3	essay contains **a few slightly confusing transitions** between ideas but is still understandable
2	essay contains **multiple confusion transitions** because it switches between ideas roughly
1	essay contains **few or no transitions** and is a **highly fragmented** collection of separate ideas

Table 18: Descriptions of the Coherence scores.

Score	Description
4	essay contains **appropriate transition words and phrases** between paragraphs, sentences, and phrases, linking statements and ideas to show their connections and aid understanding
3	essay contains **some transition words or phrases** but could somewhat benefit from their use
2	essay contains **few transition words or phrases** and would greatly benefit from their use
1	essay contains **almost no transitions** and requires their use to help understand connections

Table 19: Descriptions of the Cohesion scores.

Score	Description
4	essay is **well structured** and is organized in a way that logically develops an argument
3	essay is **fairly well structured** but could somewhat benefit from reorganization
2	essay is **poorly structured** and would greatly benefit from reorganization
1	essay is **completely unstructured** and requires major reorganization

Table 20: Descriptions of the Organization scores.

Score	Description
4	essay contains **numerous varied sentence structures** of appropriate complexity
3	essay contains **somewhat varied sentence structures** of moderate complexity
2	essay contains **limited sentence structures** of rather low complexity
1	essay **excessively and inappropriately repeats** the same simple sentence structures

Table 21: Descriptions of the Sentence Structure scores.

Score	Description
4	essay shows **appropriate word choice** and contains **advanced vocabulary**
3	essay shows **appropriate word choice** and contains **intermediate vocabulary**
2	essay shows **limited word choice** and contains **only beginning vocabulary**
1	essay **excessively and inappropriately repeats** the same words and/or phrases

Table 22: Descriptions of the Vocabulary scores.

Score	Description
4	essay contains **very few technical errors** that do not affect its overall readability
3	essay contains **some technical errors** that make it only somewhat difficult to read
2	essay contains **many technical errors** that make it significantly difficult to read
1	essay contains **numerous technical errors** that make it extremely difficult to read

Table 23: Descriptions of the Technical Quality scores.

C Analysis of the PMAES Features

In Table 24, we enumerate the features utilized by PMAES and provide a detailed description of each of them. These features can be divided into six categories: length-based, count-based, readability, essay complexity, essay variation, and other features. Features denoted by the superscript number 1 are derived utilizing the textstat package[13]. Those indicated by the superscript number 2 are obtained from the readability package[14]. Features marked with the superscript number 3 are obtained from the NLTK package[15]. Lastly, features annotated with the superscript number 4 are obtained from the spaCy package[16].

Table 25 presents the rank of each feature alongside its Pearson correlation with the OVERALL QUALITY score for ICLE++ and ASAP. A higher rank indicates a stronger Pearson correlation (either positive or negative) with the OVERALL QUALITY score. An interesting observation can be made: the readability features do not exhibit a strong Pearson correlation with OVERALL QUALITY in ASAP (averaging 0.06), whereas in ICLE++, the correlation is significantly higher (averaging 0.24).

In Table 26, we report the five features that have the strongest Pearson correlation with OVERALL QUALITY for each essay prompt in ASAP and ICLE++. Each column reports the statistics on a specific prompt. As can be seen, the top features for different prompts in ASAP are usually length-related features such as wordtypes and ess_char_len. In contrast, the top features in different prompts in ICLE++ demonstrate greater diversity, providing suggestive evidence that constructing a high-performing AES system could be more challenging on ICLE++ than on ASAP.

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

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

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

[16] https://spacy.io/

Feature Name	Description
	Length-based
mean_word	The average number of characters in each word.
ess_char_len	The number of characters in the essay.
mean_sent[3]	The average number of words in each sentence.
	Count-based
word_count	The total number of words in the essay.
unique_word	The total number of unique words in the essay.
characters_per_word[2]	The average number of characters in each word.
syll_per_word[2]	The average number of syllables in each word.
words_per_sentence[2]	The average number of words in each sentence.
sentences_per_paragraph[2]	The average number of sentences in each paragraph.
type_token_ratio[2]	The number of unique words divided by the number of words.
characters[2]	The number of characters in the essay.
syllables[2]	The number of syllables in the essay.
words[2]	The number of words in the essay.
wordtypes[2]	The total number of unique words present in the essay.
sentences[2]	The total number of sentences present in the essay.
paragraphs[2]	The total number of paragraphs present in the essay.
long_words[2]	The number of words that have 7 or more characters.
complex_words[2]	The number of words that have 3 or more syllables.
complex_words_dc[2]	The total number of words that are not in the Dale-Chall word list of 3000 words recognized by 80% of fifth graders.
tobeverb[2]	The number of "to be" verbs in the essay.
auxverb[2]	The number of auxilllary verbs in the essay.
conjunction[2]	The number of conjunctions in the essay.
pronoun[2]	The number of pronouns in the essay
preposition[2]	The number of prepositions in the essay
nominalization[2]	The number of nominalizations in the essay
pronoun[2]	The number of sentences in the essay that begin with a pronoun.
interrogative[2]	The number of sentences in the essay that begin with an interrogative.
article[2]	The number of sentences in the essay that begin with an article.
subordination[2]	The number of sentences in the essay that begin with a subordination.
conjunction[2]	The number of sentences in the essay that begin with a conjunction.
preposition[2]	The number of sentences in the essay that begin with a preposition.
spelling_err[3]	The number of words that are not in the Brown corpus of the NLTK package.
prep_comma[3]	The number of prepositions and commas in the essay.
MD[3]	The number of tokens having a POS tag of MD in the text.
DT[3]	The number of tokens having a POS tag of DT in the text.
TO[3]	The number of tokens having a POS tag of TO in the text.
PRP$[3]	The number of tokens having a POS tag of PRP$ in the text.
JJR[3]	The number of tokens having a POS tag of JJR in the text.
WDT[3]	The number of tokens having a POS tag of WDT in the text.
VBD[3]	The number of tokens having a POS tag of VBD in the text.
WP[3]	The number of tokens having a POS tag of WP in the text.
VBG[3]	The number of tokens having a POS tag of VBG in the text.
RBR[3]	The number of tokens having a POS tag of RBR in the text.
CC[3]	The number of tokens having a POS tag of CC in the text.
VBP[3]	The number of tokens having a POS tag of VBP in the text.
JJS[3]	The number of tokens having a POS tag of JJS in the text.
VBN[3]	The number of tokens having a POS tag of VBN in the text.
POS[3]	The number of tokens having a POS tag of POS in the text.
NNS[3]	The number of tokens having a POS tag of NNS in the text.
WRB[3]	The number of tokens having a POS tag of WRB in the text.
JJ[3]	The number of tokens having a POS tag of JJ in the text.
CD[3]	The number of tokens having a POS tag of CD in the text.
NNP[3]	The number of tokens having a POS tag of NNP in the text.
RP[3]	The number of tokens having a POS tag of RP in the text.
RB[3]	The number of tokens having a POS tag of RB in the text.
IN[3]	The number of tokens having a POS tag of IN in the text.
VB[3]	The number of tokens having a POS tag of VB in the text.
VBZ[3]	The number of tokens having a POS tag of VBZ in the text.
NN[3]	The number of tokens having a POS tag of NN in the text.
PRP[3]	The number of tokens having a POS tag of PRP in the text.
.[3]	The number of periods in the essay.

Continued on next page

Feature Name	Description
,[3]	The number of commas in the essay.
	Readibility
automated_readability[1]	A readability metric that measures the readability of a text based on characters per word and words per sentence.
linsear_write[1]	A readability metric developed for the U.S. Air Force to help them calculate the understandability of technical manuals, factoring in sentence length and words that are considered difficult.
Kincaid[2]	A readability metric which estimate the readability of English texts based on sentence length and word length.
ARI[2]	A readability metric that measures the readability of a text based on characters per word and words per sentence.
Coleman-Liau[2]	A readability assessment that estimates the U.S. grade level required to understand a piece of text based on characters, words, and sentences.
FleschReadingEase[2]	A readability metric that measures the readability of text based on syllables, words, and sentences. The scores are on a scale from 0 to 100, with higher scores indicating easier-to-read text.
GunningFogIndex[2]	A readability metric that estimates the years of formal education a person needs to understand the text on the first reading.
LIX[2]	A readability metric that considers sentence length and the percentage of long words (words with more than six characters) in a text.
SMOGIndex[2]	A readability formula that estimates the education level needed to understand a piece of text by analyzing the number of polysyllabic words (words with three or more syllables) within the text.
RIX[2]	A variant of the LIX readability index that only takes into account the average number of long words per sentence.
DaleChallIndex[2]	A readability formula that uses word difficulty based on a list of familiar words, along with sentence length, to estimate the grade level required to understand a text.
	Essay Complexity
clause_per_s[4]	The average number of clauses per sentence.
sent_avg_depth[4]	The average parse tree depth per sentence in each essay,
avg_leaf_depth[4]	The average parse depth of each leaf node in the parse tree.
max_clause_in_s[4]	The maximum number of clauses in the sentences of the essay.
mean_clause_l[4]	The average number of words in each clause.
	Essay Variation
sent_var[3]	The variance of the length of sentences in the essay.
word_var[3]	The variance of the length of words in the essay.
stop_prop	The percentage of stopwords in the essay.
	Sentiment
overall_positivity_score[3]	Overall, how positive the essay is.
overall_negativity_score[3]	Overall, how negative the essay is.
positive_sentence_prop[3]	The percentage of positive sentences in the essay.
neutral_sentence_prop[3]	The percentage of neutral sentences in the essay.
negative_sentence_prop[3]	The percentage of negative sentences in the essay.

Table 24: The features used by the PMAES system along with their descriptions.

Feature Name	Rank in ICLE++	PC in ICLE++	Rank in ASAP	PC in ASAP
wordtypes	43	0.103	1	0.694
complex_words_dc	17	0.224	2	0.653
sentences	40	-0.107	3	0.648
sentences_per_paragraph	61	-0.062	4	0.636
long_words	10	0.258	5	0.623
characters	31	0.143	6	0.603
syllables	24	0.170	7	0.594
complex_words	13	0.243	8	0.588
preposition	47	0.094	9	0.575
words	65	0.055	10	0.574
pronoun	22	-0.175	11	0.493
tobeverb	73	-0.031	12	0.487
type_token_ratio	64	0.057	13	-0.460
conjunction	52	-0.082	14	0.449
unique_word	29	0.153	15	0.416
nominalization	20	0.184	16	0.336
auxverb	76	0.026	17	0.324
ess_char_len	26	0.163	18	0.319
word_var	9	0.263	19	0.315
prep_comma	25	0.165	20	0.313
stop_prop	23	0.174	21	0.312
article	55	0.074	22	0.310
preposition	47	0.094	23	0.309
pronoun	22	-0.175	24	0.307
word_count	53	0.080	25	0.290
mean_word	2	0.307	26	0.287
,	44	0.101	27	0.283
spelling_err	37	-0.117	28	0.258
PRP	7	-0.272	29	-0.231
SMOGIndex	6	0.273	30	0.218
VBP	32	-0.135	31	-0.187
subordination	86	-0.001	32	0.172
RIX	15	0.241	33	0.162
JJ	19	0.187	34	0.159
mean_clause_l	34	0.130	35	0.158
VB	51	-0.085	36	-0.150
neutral_sentence_prop	42	-0.103	37	-0.150
characters_per_word	4	0.301	38	0.150
max_clause_in_s	81	0.008	39	0.149
VBN	45	0.100	40	0.129
VBZ	71	0.040	41	-0.125
WRB	46	-0.095	42	-0.124
WP	41	-0.106	43	-0.123
interrogative	50	-0.086	44	0.115
NNP	80	-0.013	45	0.113
negative_sentence_prop	79	-0.014	46	0.107
CC	59	0.068	47	-0.088
NNS	48	0.094	48	0.088
LIX	11	0.254	49	0.083
JJS	85	0.002	50	0.081
GunningFogIndex	14	0.242	51	0.075
MD	83	-0.005	52	-0.073
mean_sent	68	0.046	53	-0.071
clause_per_s	84	0.005	54	-0.060
POS	58	0.068	55	0.055
Coleman-Liau	1	0.310	56	0.053
DT	62	0.058	57	-0.052
WDT	60	0.063	58	0.049
VBD	18	-0.194	59	0.048
conjunction	52	-0.082	60	0.047
overall_negativity_score	77	0.019	61	0.046
PRP$	49	-0.087	62	0.043
syll_per_word	3	0.305	63	0.040
RP	74	-0.029	64	-0.039
paragraphs	63	-0.058	65	0.031
.	28	-0.158	66	-0.030
positive_sentence_prop	38	0.109	67	0.029
VBG	69	-0.045	68	0.029

Continued on next page

Feature Name	**Rank in ICLE++**	**PC in ICLE++**	**Rank in ASAP**	**PC in ASAP**
words_per_sentence	36	0.120	69	-0.029
ave_leaf_depth	35	0.124	70	0.027
ARI	16	0.229	71	0.025
RBR	78	0.017	72	0.021
Kincaid	12	0.248	73	-0.021
TO	75	-0.027	74	-0.020
NN	56	0.073	75	-0.015
sent_ave_depth	27	0.159	76	0.015
linsear_write	33	0.132	77	-0.015
CD	70	-0.045	78	0.012
FleschReadingEase	5	-0.296	79	-0.009
RB	54	0.076	80	0.009
JJR	72	0.037	81	0.009
DaleChallIndex	8	0.270	82	-0.007
automated_readability	21	0.179	83	0.003
overall_positivity_score	82	0.008	84	0.003
sent_var	57	-0.072	85	-0.002
IN	30	0.148	86	-0.001

Table 25: The rank of each feature and its PC value with Overall Quality for ICLE++ and ASAP.

(a) Features having the highest PC values with Overall Quality for each essay prompt in ASAP.

1		2		3		4		5		6		7		8	
Feature	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**
wordtypes	.828	wordtypes	.701	characters	.711	characters	.741	ess_char_len	.821	wordtypes	.708	wordtypes	.709	unique_word	.671
ess_char_len	.818	long_words	.686	ess_char_len	.711	ess_char_len	.741	characters	.821	syllables	.704	unique_word	.709	long_words	.659
characters	.817	ess_char_len	.685	wordtypes	.707	syllables	.738	syllables	.817	characters	.693	ess_char_len	.667	wordtypes	.655
syllables	.812	characters	.685	syllables	.704	wordtypes	.736	word_count	.815	ess_char_len	.693	characters	.665	complex_words	.622
word_count	.795	syllables	.684	words	.701	word_count	.734	words	.814	word_count	.681	word_count	.660	prep_comma	.613

(b) Features having the highest PC values with Overall Quality for each essay prompt in ICLE++.

1		2		3		4		5	
Feature	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**
Coleman-Liau	0.296	Coleman-Liau	0.306	characters_per_word	0.478	long_words	0.250	unique_word	0.306
FleschReadingEase	-0.287	FleschReadingEase	-0.299	mean_word	0.474	WP	0.247	RP	0.297
DaleChallIndex	0.286	SMOGIndex	0.294	Coleman-Liau	0.462	VBD	-0.245	DaleChallIndex	0.296
nominalization	0.276	RIX	0.286	word_var	0.457	complex_words	0.244	complex_words_dc	0.294
Kincaid	0.272	LIX	0.283	syll_per_word	0.452	syll_per_word	0.225	spelling_err	0.283

6		7		8		9		10	
Feature	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**	**Feature**	**PC**
SMOGIndex	0.427	linsear_write	0.384	syll_per_word	0.625	FleschReadingEase	-0.451	FleschReadingEase	-0.644
Coleman-Liau	0.392	complex_words	0.366	FleschReadingEase	-0.620	Coleman-Liau	0.447	syll_per_word	0.627
GunningFogIndex	0.381	long_words	0.356	complex_words	0.601	pronoun	-0.443	negative_sentence_prop	0.607
LIX	0.374	SMOGIndex	0.349	Coleman-Liau	0.583	conjunction	-0.440	SMOGIndex	0.589
RIX	0.372	interrogative	-0.343	word_var	0.577	PRP	-0.430	Kincaid	0.557

Table 26: Features having the highest PC values with Overall Quality for each essay prompt.

D Additional Experimental Results

In this section, we present additional experimental results, specifically results that are expressed in commonly-used evaluation metrics other than QWK (Section D.1) and per-prompt results (Section D.2).

D.1 Results in terms of Other Metrics

In Tables 27 to 29, we report holistic scoring results on ASAP and ICLE++ in terms of mean absolute error (MAE), root mean squared error (RMSE), and Pearson Correlation Coefficient, respectively. Note that the score ranges for ASAP are much larger than those for ICLE++ in some prompts. Thus the MAE and RMSE results for ASAP might appear worse than those for ICLE++. However, if we examine the agreement-based metrics (QWK in Table 9 and Pearson Correlation Coefficient in Table 29), we can observe that AES systems generally perform better on ASAP. Within each dataset, the trends exhibited by different metrics are generally consistent: higher QWK implies higher Pearson correlation, lower MAE, and lower RMSE. Note that there are a few cases where this does not apply. For instance, in the cross-prompt setting, compared to "PMAES with traits", "Gold Traits" shows a higher QWK but also higher MAE and RMSE values. Additional experiments are needed to determine the reason.

D.2 Per-Prompt Results

Tables 30a and 30b express the per-prompt holistic scoring results on ASAP and ICLE++ in terms of QWK. The rows in these two subtables can be interpreted in the same way as the rows in Table 9. Note that for the within-prompt scoring results, the QWK scores shown in the "Avg." column in these two subtables are different from the corresponding scores shown in Table 9. The reason is that the QWK scores in these subtables are obtained by macro-averaging the per-prompt QWK scores, whereas those in Table 9 are obtained by macro-averaging the QWK scores over the folds in the cross-validation experiments.

Perhaps not surprisingly, the best results are obtained using the models trained on the gold traits. To get an idea of which of the remaining models performs the best, for each task and each corpus we boldface the best result in each column. As we can see, the model that achieves the highest average QWK score for each task-corpus combination does not always outperform its counterparts on every prompt.

(a) Within-prompt scoring

Setting	Model	ASAP	ICLE
Without traits	Uto et al.	1.0245	0.3441
	Kumar et al.	1.1452	0.5529
With traits	Uto et al. (Simple)	1.0085	0.3820
	Uto et al. (Kumar)	1.0229	0.3782
	Kumar et al.	1.1228	0.5288
	Gold Traits	0.9630	0.1902

(b) Cross-prompt scoring

Setting	Model	ASAP	ICLE
Without traits	PMAES	1.9150	0.3979
With traits	PMAES	1.4557	0.3664
	Gold Traits	1.5809	0.1882

Table 27: Holistic scoring results in terms of mean absolute error (MAE).

(a) Within-prompt scoring

Setting	Model	ASAP	ICLE
Without traits	Uto et al.	1.4254	0.4837
	Kumar et al.	1.5540	0.6804
With traits	Uto et al. (Simple)	1.4056	0.5141
	Uto et al. (Kumar)	1.4261	0.5103
	Kumar et al.	1.5364	0.6495
	Gold Traits	1.4401	0.3039

(b) Cross-prompt scoring

Setting	Model	ASAP	ICLE
Without traits	PMAES	2.3806	0.5383
With traits	PMAES	1.8543	0.4977
	Gold Traits	1.9907	0.3147

Table 28: Holistic scoring results in terms of root mean squared error (RMSE).

(a) Within-prompt scoring

Setting	Model	ASAP	ICLE
Without traits	Uto et al.	0.7679	0.4617
	Kumar et al.	0.7111	0.3755
With traits	Uto et al. (Simple)	0.7720	0.3980
	Uto et al. (Kumar)	0.7682	0.3742
	Kumar et al.	0.7176	0.4001
	Gold Traits	0.8891	0.8890

(b) Cross-prompt scoring

Setting	Model	ASAP	ICLE
Without traits	PMAES	0.6815	0.2986
With traits	PMAES	0.7029	0.3574
	Gold Traits	0.8878	0.8787

Table 29: Holistic scoring results in terms of Pearson Correlation Coefficient.

E Overview of the Models

In this section, we give an overview of the models we use in our experiments as well as their implementation details.

E.1 Kumar et al.'s Model

Kumar et al.'s (2022) system is the state-of-the-art model on the ASAP++ dataset that performs

(a) Results on ASAP

Task	Setting	Model	1	2	3	4	5	6	7	8	Avg.
Within-prompt	w/o traits	Uto et al.	**.793**	.646	.690	**.818**	.800	**.812**	.758	.641	.745
		Kumar et al.	.749	.606	.672	.713	.784	.748	.687	.434	.674
	w/ traits	Uto et al. (Simple)	.786	**.712**	**.697**	.809	**.848**	.794	**.827**	**.659**	**.766**
		Uto et al. (Kumar)	.786	.709	.693	.814	**.848**	.794	.809	.639	.762
		Kumar et al.	.755	.606	.693	.686	.764	.762	.738	.490	.687
		Gold Traits	.852	.903	.914	.947	.884	.901	.816	.863	.885
Cross-prompt	w/o traits	PMAES	**.775**	.568	**.540**	.573	**.694**	**.559**	.667	.418	.599
	w/ traits	PMAES	.661	**.667**	.494	**.629**	.617	.461	**.738**	**.609**	**.609**
		Gold Traits	.835	.875	.906	.947	.865	.901	.462	.884	.834

(b) Results on ICLE++

Task	Setting	Model	1	2	3	4	5	6	7	8	9	10	Avg.
Within-prompt	w/o traits	Uto et al.	**.374**	**.429**	**.554**	**.431**	**.263**	.291	.325	**.601**	**.354**	.449	**.407**
		Kumar et al.	.320	.304	.281	.310	.085	**.376**	.307	.447	.192	.346	.297
	w/ traits	Uto et al. (Simple)	.303	.302	.442	.267	.106	.150	.328	.362	.305	.233	.280
		Uto et al. (Kumar)	.222	.320	.405	.218	.090	.268	.267	.443	.235	.260	.273
		Kumar et al.	.323	.393	.377	.206	.177	.330	**.338**	.345	.169	**.461**	.312
		Gold Traits	.791	.794	.789	.690	.847	.770	.834	.804	.813	.834	.796
Cross-prompt	w/o traits	PMAES	**.198**	.197	.339	.234	.112	.233	.074	.442	.201	**.479**	.251
	w/ traits	PMAES	.162	**.269**	**.397**	**.240**	**.163**	**.314**	**.077**	**.519**	**.269**	.405	**.281**
		Gold Traits	.841	.855	.855	.775	.939	.852	.879	.864	.888	.909	.866

Table 30: Prompt-specific holistic scoring results.

within-prompt multi-task learning using trait information. The system contains a stack of layers for each of the trait scores as well as the holistic score. Within each stack, it obtains different levels of representation of the input essay using a CNN layer and a LSTM layer. First, to obtain a representation for each sentence in the essay, it passes the GloVe embedding (Pennington et al., 2014) of each token in the sentence to the CNN layer and applies the attention pooling operation over the output of the CNN layer. Then, it obtains the document-level representation of the essay by passing the resulting sentence-level representations to the LSTM layer and applying attention pooling to the hidden states of the LSTM layer. For each of the trait scoring stacks, the document-level representation is passed to a dense layer to predict the corresponding trait score. After that, these predicted trait scores and the document-level representation from the holistic scoring stack are then passed into a dense layer to predict the holistic score. The training process of this system minimizes the MSE loss.

E.2 Uto et al.'s Model

Uto et al's (2020) system is simple yet effective. The authors concatenate the embedding of the input essay obtained by the BERT model with several hand-crafted essay-level features such as length-based features and count-based features. Subsequently, they pass this representation to a linear layer to get the predicted essay score. By fine-tuning BERT with the hand-crafted features, they achieved state-of-the-art performance in holistic essay scoring on the ASAP dataset at the time.

To make Uto et al.'s system predict trait scores, we experiment with two approaches: (1) Uto et al. (simple), which merely extends the number of output neurons in the final linear layer, and (2) Uto et al. (Kumar), where an architecture similar to Kumar et al.'s (2022) system is employed, initially predicting trait scores and subsequently using both the essay embeddings and the trait scores for holistic score prediction.

E.3 The PMAES Model

The PMAES model, introduced by Chen and Li (2023), focuses on cross-prompt essay scoring. To facilitate cross-prompt essay scoring, the authors propose a prompt-mapping framework in which the training prompts are divided into *source prompts* and *target prompts*, and the goal is to employ contrastive learning to align the essay representations from the source and target prompts. Their prompt-mapping framework consists of a source-to-target prompt mapping procedure and a target-to-source prompt mapping procedure. The source-to-target prompt mapping procedure operates as follows. First, for source essay i and its essay representation r_i, a source-to-target mapping representation $\hat{r_i}$ is first obtained by (1) taking the dot product of each source essay representation vector with the transpose of the matrix that consists of all target essay representations, and (2) multiplying the resultant product by a matrix of learnable parameters. To align the essay representations in the source and target prompts, the model then considers the pairs $(r_i, \hat{r_i})$ as positive samples while treating (r_i, r_j) as negative samples for source essays i and j. The target-to-source prompt mapping procedure works similarly. The prompt mapping procedures are optimized using the contrastive learning loss as defined by Chen et al. (2020). The final step of PMAES involves predicting the holistic score, specifically by adding linear layers atop the essay representation that has been concatenated with hand-crafted features.

E.4 Implementation Details

For all models, we tune two hyperparameters on development data, the learning rate and the dropout rate. Specifically, we experiment with learning rates of 1×10^{-3}, 1×10^{-4}, 3×10^{-4}, 6×10^{-4}, 1×10^{-5}, and 3×10^{-5}, and dropout rates of $0.1, 0.2, 0.3, 0.4$, and 0.5. All models are executed with the random seed set to 11.

Kumar's model is trained for 150 epochs for ICLE++ and 100 epochs for ASAP. We use 1×10^{-3} as the learning rate, 64 as the batch size, AdamW (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as the optimizer, and $0.1 \times \{\text{total number of update steps}\}$ as the number of warm-up steps. This system is trained on a single RTX 3090. It takes around 4 hours to finish the training process.

Uto et al.'s model along with its variations are all trained for 50 epochs for ICLE++ and 20 epochs for ASAP. We use 6×10^{-4} as the learning rate, 64 as the batch size, AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ as the optimizer, and $0.1 \times \{\text{total number of update steps}\}$ as the number of warm-up steps. This system is trained on a single RTX A6000. It takes around 4 hours to finish the training process.

PMAES is trained for 50 epochs for ICLE++ and 20 epochs for ASAP. We use 3×10^{-4} as the learning rate, and Adam (Kingma and Ba, 2015) with $\lambda_1 = 0.5$ and $\tau = 0.1$ as the optimizer. This system is trained on a single RTX A6000. It takes around 5 hours to finish the training process.

The linear regressor that is trained on gold traits is implemented using the scikit-learn package. All hyper-parameters of the linear regressor are set to their default values.