Automated Essay Scoring: A Reflection on the State of the Art

Shengjie Li and Vincent Ng

Human Language Technology Research Institute

University of Texas at Dallas

{sxl180006,vince}@hlt.utdallas.edu

Abstract—While steady progress has been made on the task of automated essay scoring (AES) in the past decade, much of the recent work in this area has focused on developing models that beat existing models on a standard evaluation dataset. While improving performance numbers remains an important goal in the short term, such a focus is not necessarily beneficial for the long-term development of the field. We reflect on the state of the art in AES research, discussing issues that we believe can encourage researchers to think bigger than improving performance numbers, with the ultimate goal of triggering discussion among AES researchers on how we should move forward.

I. INTRODUCTION

Automated Essay Scoring (AES), the task of automatically assigning a holistic score to an essay that summarizes its overall quality, is arguably one of the most important applications in natural language processing (NLP). As an example of AES, consider the essay in Table I, which is written in response to the prompt shown at the top of the table. Given the scoring rubric in Table II, an AES system should assign a score of 3 to this essay for the following reasons. First, its author takes a position but fails to provide adequate support and details. Specifically, the author talks about computers giving people entertainment and lists some general social networking websites, but there is no elaboration on how computers enhance access to entertainment. In terms of organization, while the essay has a basic structure (an introduction, three main body sentences, and a conclusion), the ideas within each body sentence are poorly connected. Finally, the transitions are rather awkward and repetitive. For example, the author uses "so as you can see" three times. While the author exhibits some awareness of the audience as the essay is addressed to a local newspaper, it is not clear whether the essay is intended to urge the editor to write an article on how computers benefit people. Given the large number of essays written by students from all over the world in both test and classroom settings, being able to automatically score essays could save a tremendous amount of manual grading effort.

Despite the fact that AES has been investigated for more than 50 years [46], the task is still far from being solved. Nevertheless, AES has been progressing steadily. In recent years, a standard recipe for publishing in this area of research seems to involve proposing a sophisticated neural model that can beat competing models on a standard evaluation dataset such as ASAP (see Section II-A for details). While improving performance numbers remains an important short-term goal

for AES research, such a focus is not necessarily beneficial for the long-term development of the field.

In this position paper, we reflect on the state of the art in AES research. Specifically, we discuss issues that we believe can encourage researchers, particularly junior researchers, to think bigger than merely improving performance numbers, with the ultimate goal of triggering discussion among AES researchers on how we should move forward.[1] These issues range from evaluation (Sections III and IV) to new tasks (Sections V and VI) to corpus annotation (Sections VII and VIII) and beyond (Section IX).

II. CURRENT STATE OF AES RESEARCH

To facilitate our discussion of the future directions of AES research, in this section, we provide an overview of the current state of AES research.[2]

A. Corpora

While a number of AES corpora have been developed, ASAP is arguably the most extensively used in AES research. Introduced as part of a 2012 Kaggle competition, the Automated Student Assessment Prize (ASAP) corpus has become a popular dataset for holistic scoring, especially given its vast collection of essays per prompt (up to 1,800 for some prompts). ASAP facilitates the development of high-performing, prompt-specific systems. ASAP++ [40] is an extension of ASAP where each essay is scored along multiple traits (i.e., dimensions of essay quality such as Coherence). Note that ASAP is composed of three types of essays (namely, narrative/descriptive essays, persuasive essays, and source-dependent essays). Not all traits are applicable to all essay types. For instance, Organization and Conventions are scored for narrative/descriptive essays and persuasive essays only.

Other English corpora that have been developed for and used in AES research include: (1) TOEFL [9], a corpus of essays from the TOEFL exam that is originally developed for the Native Language Identification task where the proficiency

---

[1] While Beigman Klebanov and Madnani's [7] theme paper focuses more broadly on the emergent uses of automated writing technologies, we focus specifically on research directions that we think are worth pursuing by AES researchers.

[2] For a comprehensive overview of AES research, we refer the reader to the books published by Shermis and Burstein [55], Shermis et al. [56] and Beigman Klebanov and Madnani [8], as well as the surveys published by our group [28], [34].

TABLE I
A SAMPLE ESSAY TAKEN FROM ESSAY SET 1 OF THE ASAP CORPUS. THE WRITING PROMPT IS SHOWN AT THE TOP.

[Prompt]
More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends.
Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.

[Essay]
Dear local newspaper, opinion about computers is that they benifet people. they help people when they need it, it gives people entertainment. So that is why computers benifet people. first reason is that a computer can help people when they need help like for instace if you have to write a paper and you need to look something up such as reseach you can use the computer or if there is a long main cwayshion then you can look up the formula. So as you can see that is why computer benifts people second resson is that it gives people entertainment, such as myspace, you twitter. And those are social networks, if you like to watch tv. or listen to music there are websites for them also. So that is why I said computers give you entertainmen too. third resson is that it helps people such as if you in different contries and you can't go see them you to them through instans or web cam. Or if you are home schooled and you dnt see other kids your age that can talk to the on the computer. So as you can see that why I say you can through the computer. So as you can see that is why I say computer are very benifical be cause they give you entertainment they let you communicate and they help people so.

This essay receives a score of 3 as it takes a position but fails to provide adequate support and details. Specifically, the author talks about computers giving people entertainment and lists some general social networking websites, but there is no elaboration on how computers enhance access to entertainment. In terms of organization, the essay has a basic structure: an introduction, three main body sentences, and a conclusion, but the ideas within each body sentence are poorly connected. Moreover, the transitions are rather awkward and repetitive. For example, the author uses "so as you can see" three times. While the author exhibits some awareness of the audience as the essay is addressed to a local newspaper, it is not clear whether the essay is intended to urge the editor to write an article on how computers benefit people.

TABLE II
RUBRIC FOR SCORING ESSAYS IN ESSAY SET 1 OF THE ASAP CORPUS.

Score	Description
1	An undeveloped response that may take a position but offers no more than very minimal support. Typical elements: (1) Contains few or vague details, (2) Is awkward and fragmented, (3) May be difficult to read and understand, (4) May show no awareness of audience.
2	An under-developed response that may or may not take a position. Typical elements: (1) Contains only vague/unelaborated and/or list-like details, (2) Shows little or no evidence of organization, (3) May be awkward and confused or simplistic, (4) May show little awareness of audience.
3	A minimally-developed response that may take a position, but with inadequate support and details. Typical elements: (1) Has reasons with minimal elaboration and more general than specific details, (2) Shows some organization, (3) May be awkward in parts with few transitions, (4) Shows some awareness of audience.
4	A somewhat-developed response that takes a position and provides adequate support. Typical elements: (1) Has adequately elaborated reasons with a mix of general and specific details, (2) Shows satisfactory organization, (3) May be somewhat fluent with some transitional language, (4) Shows adequate awareness of audience.
5	A developed response that takes a clear position and provides reasonably persuasive support. Typical elements: (1) Has moderately well elaborated reasons with mostly specific details, (2) Exhibits generally strong organization, (3) May be moderately fluent with transitional language throughout, (4) May show a consistent awareness of audience.
6	A well-developed response that takes a clear and thoughtful position and provides persuasive support. Typical elements: (1) Has fully elaborated reasons with specific details, (2) Exhibits strong organization, (3) Is fluent and uses sophisticated transitional language, (4) May show a heightened awareness of audience.

label that comes with each essay (Low, Medium, or High) is used as its holistic "score" for training AES systems; (2) the Cambridge Learner Corpus-First Certificate in English exam (CLC-FCE) [71], where each essay is scored holistically and annotated with the linguistic error types it contains; (3) the International Corpus of Learner English (ICLE) [22], where a subset of essays has been scored not only holistically [35] but also along multiple dimensions of essay quality, such as Organization [47] and Argument Persuasiveness [50]; and (4) the Argument Annotated Essays (AAE) corpus [58], where each persuasive essay is scored based on the strength of its thesis [28] and the persuasiveness of its argument [27].

AES corpora in other languages exist, such as Ostling's [45] Swedish corpus, Horbach et al.'s [24] German corpus, Marinho et al.'s [39] Portuguese corpus, the GoodWriting dataset[3] (in Japanese), and the MERLIN dataset[4], which is composed of German, Italian, and Czech essays.

B. Evaluation Metric

The standard metric used to evaluate AES models is Quadratic weighted Kappa (QWK)[5]. QWK is an agreement metric that ranges from 0 to 1, but can be negative if there is agreement lower than what is expected by chance. More specifically, QWK is a weighted version of Kappa where each case of disagreement (i.e., the (rounded) predicted score is different from the reference score) is weighted by the squared difference between the reference score and the predicted score. This allows the metric to distinguish between near misses and far misses.[6]

C. Systems

AES systems can be divided into three categories:

1) Heuristic Approaches: Virtually all early AES systems are heuristic-based and typically possess the following characteristics (e.g., Elliot [18], Attali and Burstein [3]):

Trait-driven holistic scoring. Many traits play a role when human raters score an essay holistically, such as Organization, Coherence, Technical Quality (i.e., fluency, grammar, and mechanics), and Argument Persuasiveness. Motivated by the

[3]https://goodwriting.jp/wp/?lang=en
[4]https://www.merlin-platform.eu/
[5]See https://www.kaggle.com/competitions/asap-aes/overview/evaluation for details.
[6]Several other metrics have also been used although they are less popular than QWK, including Pearson's Correlation Coefficient, mean squared error and mean absolute error.

human essay scoring process, the holistic score returned by a heuristic AES system is typically computed as the weighted sum of the trait scores.

Heuristic trait-specific scoring. Given the lack of annotated data, each trait-specific score is computed using heuristics. For example, to compute the Organization score, which reflects how well-organized the essay is, the e-rater system [3] determines whether the essay is organized as a 5-paragraph essay where the first paragraph is the introduction, the last paragraph is the conclusion, and the middle three paragraphs each presents a key point with supporting evidence. The functional role of each paragraph (e.g., Introduction) is determined heuristically.

Focus on non-content-based traits. Traits can broadly be divided into two categories: content-based traits, which are based on the essay's content (e.g., Argument Persuasiveness, Coherence), and non-content-based traits, which are based on the surface realization of the content (e.g., Grammar, Fluency). Generally, the content-based traits are much harder to score than the non-content-based traits. For example, while Fluency and Grammaticality can be determined fairly easily using a language model and a grammar checker respectively, determining Argument Persuasiveness may require a deep understanding of the content. Content-based traits are particularly difficult to compute in the absence of labeled data. Consequently, heuristic approaches have largely focused on employing non-content-based traits for holistic scoring.

2) Machine Learning Approaches: As annotated AES corpora became publicly available in the early 2010s, the focus of AES research also started to shift from heuristic approaches to machine learning approaches, where an off-the-shelf machine learning algorithm (e.g., SVM, linear regression) is used to train a classifier or a regressor for scoring. AES research in the machine learning era has the following characteristics:

Focus on feature engineering. The focus is designing low-level and high-level features. Low-level features include length-based features (e.g., the number of tokens in the essay) [71], [62], lexical features (e.g., the presence/count of each n-gram) [10], [51], word embeddings [12], word category features (e.g., whether a word is a modal) [19], [42], and syntactic features (e.g., part-of-speech tag sequences) [10]. High-level features include readability features (i.e., metrics that reflect how easy it is to read the essay) [72], prompt-relevant features (i.e., features that encode the similarity between the essay and the prompt it was written for) [37], [6], argumentation features (e.g., the number of claims in a persuasive essay) [21], [64], [44], semantic features (e.g., features derived from lexico-semantic resources such as FrameNet [4]) [5], and discourse features (e.g., local coherence features derived from Centering Theory [23]) [70].

Focus on within-prompt scoring. In within-prompt scoring, an AES model is trained on essays written for a prompt and then applied to test essays written for the same prompt. Some have argued that within-prompt scoring is not a practical setting: when within-prompt scorers are applied to essays written for a new prompt, their performance often deteriorates

considerably. So, before they are applied to score essays written for a new prompt, they need to be retrained on scored essays written for the new prompt. However, manually scoring essays is time-consuming and requires a lot of expertise.

Learning-based trait-specific scoring. As machine learning approaches to AES became popular, researchers began to examine learning-based approaches to trait-specific scoring. The development of learning-based models for trait-specific scoring is facilitated by the release of annotated datasets where essays are scored along different essay traits [48], [49], [50]. While the scoring of content-based traits is largely ignored in heuristic approaches, researchers have begun learning models for scoring content-based traits. Nevertheless, even with annotated data, the scoring of content-based traits remains a challenging task.

3) Deep Learning Approaches: With the advent of the neural NLP era, the vast majority of recently developed AES models are deep learning-based. AES research during this period can be summarized as (1) a focus on learning the distributed representation of an essay (by adjusting the weights in a neural network) so that essays that are similar in quality will have similar representations and (2) an exploration of new, challenging AES task settings such as cross-prompt scoring and multi-trait scoring.

Early approaches. Early neural models combine CNNs and RNNs to capture spatial and temporal dependencies respectively. For instance, Taghipour and Ng [60] first use a CNN to extract n-gram-level features to capture local dependencies and then use an LSTM to generate a long-distance representation of an essay for holistic scoring. These models are subsequently replaced by Transformer-based models, which possess a vast amount of linguistic and commonsense knowledge acquired from large, unlabeled corpora. For instance, Yang et al. [69] proposed R2BERT, an AES model obtained by fine-tuning BERT. Wang et al. [65] proposed a multi-scale BERT-based structure that captures (automatically learned) features at the token, segment, and essay levels. Uto et al. [61] showed that neural AES models could be improved with hand-crafted features.

Neural AES models can be improved by exploiting document structure. Dong and Zhang [16] viewed an essay as having a two-level hierarchical structure: an essay is composed of a sequence of sentences, each of which is composed of a sequence of words. Given this view, they designed a two-layer model where the first layer creates a representation for each sentence and the second layer creates an essay representation by combining sentence representations. Further improvements can be made via attention pooling [17].

Cross-prompt scoring. As noted above, some have argued that within-prompt scoring is not a practical setting for AES. Hence, researchers have recently begun working on the task of cross-prompt AES [53], where the goal is to train a model that can offer good performance when it is applied to score essays written for unseen prompts. A few approaches to this relatively new task of cross-prompt scoring have been developed. Cummins et al. [13] recast cross-prompt scoring

as a domain adaptation problem, where a prompt is viewed as a domain. Specifically, the goal is to use a domain adaptation method to adapt an AES model trained on the source prompts to the target prompt. Do et al. [14] incorporated as input for cross-prompt scoring essay prompt information, which ironically is not exploited by many AES models. To facilitate generalization to new prompts, Chen and Li [11] proposed a cross-prompt scoring model that seeks to make the source essay representations and the target essay representations more consistent with each other via contrastive learning, and several researchers have employed prompt-independent features in their AES models [26], [36], [54].

Multi-trait scoring. Since the holistic score of an essay is influenced by its trait-specific scores, it would be natural to develop joint models for AES that simultaneously predict the holistic score and the various trait-specific scores. Multi-trait scoring models can be obtained by several simple methods. The first method involves replacing the output layer of a holistic scoring model with multiple output layers, one for each trait [25]. The second method involves making multiple copies of a holistic scoring model where each copy is responsible for scoring one trait and the copies interact with each other via a shared representation layer [41]. The third method is similar to the second, except that the predicted trait scores are used as input to predict the holistic score [32].

LLM-based approaches. Mizumoto and Eguchi [43] investigated how prompting in large language models (LLMs) can be exploited for AES. LLM-based approaches are motivated by two key strengths of LLMs. First, LLMs possess a vast amount of commonsense knowledge that can be exploited to perform various tasks. Second, LLMs are good at understanding complex natural language instructions [2]. Given these strengths, we can ask an LLM to perform a task as complex as AES by providing instructions in the form of a prompt that may include, for instance, the rubric in a zero-shot setting, where no manually scored essays are provided as training examples (e.g., Lee et al. [33]), or a few-shot setting, where a few labeled examples are provided as part of the prompt (e.g., Mansour et al. [38]; Xiao et al. [66]).

III. AES IS MORE THAN IMPROVING QWK

What have we learned about AES over the years other than the fact that the QWK scores are improving on standard evaluation datasets? In the pre-neural NLP era when many AES researchers were focusing on feature engineering, we could at least gain some insights into what kind of features would be useful for AES through feature ablation experiments. With neural AES models, the representation of an essay is learned. Many of them do not employ hand-crafted features, and hence the output cannot be interpreted easily. We should note, however, that researchers have recently begun to explore the explainability of AES models. For instance, Kumar and Boulanger [31] applied explainable AI techniques, such as SHAP values, to enhance the transparency and interpretability of their AES system. Fiacco et al. [20] developed a methodology that aligns the functional components derived from Transformer models with human-understandable feature groups. Adadi and Berrada [1] explored various interpretability methodologies in machine learning that can be adapted to AES for increased transparency, such as LIME [52].

Recommendation #1: We recommend that researchers understand the strengths and weaknesses of the systems they developed by performing a qualitative and quantitative analysis of the system outputs. For example, while it is not uncommon for AES researchers to report results that are better than existing results when averaged over all essay prompts in the corpus, without further analysis, a reader would not know where the improvements came from. Does the proposed model perform better on all types of essays (e.g., persuasive essays, narratives) or only certain types of essays, and if it performs better on all types of essays, are there certain essay types for which the improvement is more pronounced? Does the model perform better because it can better distinguish between essays that are similar w.r.t. particular traits (e.g., Organization)? Is the model better because it scores the essays belonging to the minority classes better? Note that to answer these questions, we cannot resort to interpretability techniques. Rather, an error analysis on model outputs is needed.

IV. ASAP IS NOT THE ONLY ESSAY CORPUS

As mentioned before, in recent years many AES researchers evaluated their model solely on ASAP. A natural question is: have existing models overfitted ASAP? In other words, can models that perform well on ASAP be expected to generalize well to other essay corpora? Unfortunately, little analysis has been performed on ASAP that would enable us to gain a better understanding of whether certain corpus-specific characteristics are present in ASAP that could cause models to overfit the data. For example, what fraction of the persuasive essays in ASAP made arguments that are persuasive? If the vast majority of these essays contained persuasive arguments, then a model trained on ASAP might not perform well on a corpus where it is important to distinguish between essays with persuasive arguments and those with unpersuasive arguments.

One piece of information that we know about ASAP is that the essays it contains were written by American students in Grades 7–10. While the native languages of these students are not known, the majority of them are expected to be native speakers of English. Given this context, ASAP is not an ideal corpus for the development and evaluation of AES systems. Recall that one of the primary motivations and key driving forces behind the development of AES systems is its application to scoring essays written by takers of standardized tests such as TOEFL and SAT, where many of the exam takers are learners of English as a second language (ESL). As is well-known to essay grading researchers, the common types of error made by ESL learners who speak different native languages are not the same [71]. Hence, it is possible that the ASAP essays are simpler to score because grammatical errors may be less of an issue in these essays. In contrast, for AES of essays written by non-native speakers, AES systems may need to be combined with grammatical error correction systems for

more accurate modeling of the quality of an essay w.r.t. the Grammar trait.

Another piece of information that we know about ASAP (and AES corpora where essays are written in a time-restricted setting such as a test setting) is that essay length (as measured by the number of words) is often a confounding variable for holistic scoring [20]. Consequently, it is not clear how models trained on ASAP would perform when applied to corpora where essays are written in a time-unrestricted setting, such as essays written as part of a homework with a length restriction (e.g., between 500 and 600 words), as length may no longer be a confounding variable in these corpora.

Recommendation #2: To understand whether models trained on ASAP can generalize well when applied to other essay corpora, we recommend that AES researchers evaluate their systems on not only ASAP but at least one other publicly available corpus, preferably a corpus where the native languages of the essay writers are different from the language in which the essays were written, such as CLC-FCE [71] or a corpus where essays were written in a time-unrestricted setting, such as ICLE++ [35].

V. IS IT TIME FOR CROSS-PROMPT AES?

Cross-prompt AES is a challenging task. Below we discuss the challenges from three perspectives.

Knowledge. For an AES model to perform well on a new prompt, we need knowledge specific to the new prompt. For example, if the new prompt is "write a persuasive essay on whether capital punishment should be abolished", an AES model needs to distinguish between persuasive arguments and unpersuasive arguments for (or against) capital punishment. For cross-prompt scoring, this kind of knowledge needs to be extracted from external knowledge sources. One possibility is to prompt a LLM: a human first provides a few examples of persuasive and unpersuasive arguments for each stance, which are then used to elicit prompt-specific knowledge inherent in the LLM.

Training data. What kind of training data is needed for cross-prompt AES? If the test essays are persuasive essays, then ideally the training essays should also be persuasive essays. The reason is that the traits that affect the holistic score of a persuasive essay (e.g., Argument Persuasiveness) are not the same as those that affect the holistic score of a non-persuasive essay, even though there are traits that are common to all types of essays (e.g., Organization). If the training set and the test set are composed of different types of essays, what is learned about good essays from the training set may not necessarily be applicable to good essays in the test set. Similarly for rubrics: the rubric used to score the training essays should be the same (or at least similar to) the one used to score the test essays; otherwise, knowledge of good essays that is learned from the training set may not be transferable to the test set because essays that are good according to the training rubric could be considered bad according to the test rubric.

However, this is not how Ridley et al.'s [53] cross-prompt AES model was trained and evaluated. Specifically, they employed ASAP for training and evaluation. Recall that ASAP is composed of persuasive, narrative, and source-dependent essays written for eight prompts. Their cross-prompt AES experiments were conducted using leave-one-prompt-out cross validation, where they trained a model using essays for all but one prompt and tested it on essays for the prompt that was left out. This means that the training set may be composed of essays that have a different essay type than that in the test set. Additionally, the training and test prompts often employ differing rubrics. It is not clear whether it even makes sense to train and evaluate the model using this setup.

Models. Can within-prompt AES models be used for cross-prompt AES? The vast majority of existing within-prompt AES models do not take the essay prompt and the scoring rubric as input [65], [68]. For within-prompt AES, this may be fine, as the essays in the training set and the test set are written for the same prompt and scored using the same rubric. For cross-prompt AES, this may not be fine, as the model does not know (1) the prompt-specific knowledge that should be extracted from external sources without knowledge of the prompt and (2) the criteria for good and bad essays without the rubric. Hence, it is not clear the extent to which the performance of Ridley et al.'s [53] cross-prompt AES model is limited by the fact that it relies on neither the unseen prompt nor the rubric.

Recommendation #3: Cross-prompt AES is a challenging task that deserves the attention of AES researchers. We recommend that researchers focus on developing new annotated corpora, new methods for gathering prompt-specific knowledge, and new AES models for cross-prompt AES.

VI. TRAITS ARE USEFUL BUT UNDER-EXPLORED

Two key challenges discussed above involve building AES models that are interpretable and generalizable to unseen prompts. Next, we propose to examine the use of traits to address both challenges.

Interpretability. If a holistic scoring model is built by using only traits as inputs (such as the heuristic AES models described in Section II), then interpretability is relatively straightforward. For instance, if a linear model is used where the holistic score is computed as a weighted sum of the trait-specific scores, then the holistic score can be easily explained by the trait-specific scores. Even if a non-linear model is used, the attention weights can be used to shed light on how much influence a particular trait has on the holistic score. Not only do essay traits make interpretability straightforward, but the trait scores also serve to provide feedback to an essay's writer. For example, if someone gets a low holistic score for their essay, they can identify which aspects of the essay need improvement by checking which trait-specific scores are low.

Cross-prompt generalizability. Recent efforts on cross-prompt AES research have involved identifying prompt-independent features, such as length-based features. While traits have been used as features for holistic scoring, they

have primarily been used for within-prompt rather than cross-prompt scoring. One reason, we believe, is that traits are rarely viewed as prompt-independent features. Given the prompt-independent nature of traits, we propose to use them for cross-prompt AES. Specifically, we propose to view the set of traits commonly known to play a role in determining the holistic score of a particular type of essay (e.g., persuasive essays) as a prompt-independent representation of these essays. Note that this trait-based essay representation is dependent on the essay type, as the traits that influence the holistic score of a persuasive essay are not necessarily the same as those that influence narrative essays, for instance.

In practice, however, using traits to address interpretability and cross-prompt generalizability is challenging. The key challenge stems from the fact that content-based traits are difficult to score accurately [11], [14], even when labeled data is available for training content-based trait-scoring models. One possibility is to leverage LLMs for scoring traits in a zero-shot setting, where we provide the rubric developed for a given trait as part of the instructions for an LLM when requesting it to score the trait.

Should we believe that LLMs can help us accurately perform trait scoring? As mentioned in Section II, state-of-the-art LLMs are very good at understanding inputs as complex as essays. Since scoring content-based traits requires an understanding of essay content, exploring the use of LLMs for trait scoring is a promising direction. Nevertheless, it is worth noting that recent prompt-based approaches to holistic scoring [33], including those that employ a chain-of-thought approach [66], are not as competitive as fine-tuned models in performance. While the underlying reasons are not yet clear, researchers have demonstrated that minor changes in the prompt [38], changes in the decoding methods [57], and variations in random seeds [15] can all result in significant performance changes. Given these results, one should expect similar challenges when prompting LLMs for trait scoring.

Recommendation #4: We recommend that a thorough investigation of the impact of traits on holistic scoring be conducted as they could be a viable solution to key problems concerning neural model interpretability and cross-prompt model generalizability. Given the difficulty in computing content-based traits, we recommend that researchers examine how LLMs can be exploited, possibly via prompting-based approaches, to score content-based traits.

VII. Corpus development is slow

The fact that ASAP is still the primary corpus used for evaluating AES models more than 10 years after its initial release is somewhat unusual in the NLP community for a task that is as popular as AES. This perhaps suggests that corpus development is seriously lagging behind model development. We believe the reasons are at least two-fold.

Assembling an essay corpus is by no means easy. For many other NLP tasks, assembling a large, unannotated corpus composed of news articles or tweets is relatively easy [63], [30]. This is not the case for essays: while we may be able to assemble a corpus of raw classroom essays, it may still require collaboration by multiple instructors over many years in order to obtain a sizeable corpus. Yet, even after multiple years of effort, the resulting corpus may still be too small to enable the pre-training of models specifically on raw essays.

There is no easy solution, however. Organizations that have access to a large number of essays written by people with different demographic characteristics should consider releasing language resources to the AES community. These resources would ideally be in the form of raw essays. We believe that this is doable: TOEFL11, for example, was released by the Educational Testing Service (ETS) in 2013, and we encourage ETS and similar organizations to continue this effort. If essays cannot be released, they may consider releasing models pre-trained on these essays. Language models pre-trained for specialized domains have been successfully deployed (e.g., the legal domain [67]). Unlike many other domains where it may suffice to have one pre-trained model per language, for AES research it may make sense to develop multiple pre-trained models that differ in terms of evaluation setting (e.g., whether the essays are written in a time-constrained or unconstrained setting), grade level (university undergraduates vs. high-school students), and native language.

Having models pre-trained on essays could enable them to acquire linguistic or even prompt-specific knowledge from raw essays. The hope is that with the availability of pre-trained language models for essays, a relatively small amount of labeled data will be needed to fine-tune them to perform specific essay-related tasks, such as AES.

Manually labeling essays is a labor-intensive and expensive procedure. The reason is that manual labeling of essays is typically performed by trained experts and cannot be reliably done via crowdsourcing [38], especially if it involves scoring along multiple traits. Even if we do have the time and resources to perform manual labeling of essays, the essay grading community has not developed a vision of what annotations would benefit the development of AES models as well as models for other essay-related tasks in the long run. Ideally, there would be one or two corpora that contain multiple layers of annotation. For example, one layer would be composed of scores (i.e., the holistic and trait-specific scores), another layer would be composed of written feedback to essay writers, and a third layer would be composed of essays that have been revised by experts. Having multiple layers of annotation in the same corpus would be beneficial from the perspective of model development, as it would allow a joint model to be built that can, for instance, exploit the inter-dependencies among different layers. For example, scoring and the generation of written feedback can be modeled simultaneously, as positive feedback is typically associated with positive scoring. In other words, inter-layer constraints can be incorporated into the model to capture such dependencies.

The layers that should be included in the annotation scheme should by no means be limited to the aforementioned possibilities: it is something that needs to be discussed by AES researchers and perhaps the wider essay grading community.

Recommendation #5: We recommend that the community devote more effort to corpus development, as progress on model development will eventually be hindered by the arguably slow progress on corpus development. Research organizations should consider helping to advance the field by releasing large corpora of raw (and if possible, annotated) essays or language models that are pre-trained on such corpora. In addition, we encourage the AES community to develop a shared vision of what corpora, particularly annotated corpora, would benefit AES research in the long run. Ultimately, the community should discuss the possibility of creating a large, annotated corpus and the kind of manual annotations that are desirable to have in the corpus. This whole discussion could begin with a small, annotated essay corpus released as part of a shared task that contains some of the annotation layers described above. We believe that the shared task could bring together interested parties to discuss what the next steps should be as far as corpus development is concerned.

VIII. LLMs CAN BE USED TO ASSIST AES

So far, LLMs have primarily been used in prompt-based approaches to holistic scoring, with results that are less competitive than those of fine-tuned models [43], [59]. As researchers continue to improve prompt-based approaches, we encourage them to examine other ways of exploiting LLMs for AES research.

Rather than expect LLMs to replace humans (as in prompt-based approaches to holistic scoring), we can instead aim to use LLMs to assist humans in AES-related tasks. In particular, it is worth investigating whether LLMs can be profitably used to assist corpus creation and annotation. As noted previously, assembling a large unlabeled essay corpus and performing multi-layer annotation of such a corpus are both daunting tasks. We believe LLMs can be used to perform the ground-work for these tasks. For example, given the difficulty in assembling an essay corpus, we can examine the feasibility of asking LLMs to generate essays subject to a given set of constraints, such as essay length, grade level, and score level (given the corresponding score description taken from a rubric).[7] As another example, since corpus annotation is a time-consuming process, we can ask LLMs to perform trait scoring given the corresponding rubrics (possibly by providing the rationale behind why the essay deserves a particular score), provide written feedback (given their strong generative capabilities), and revise essays in a zero-shot fashion.

While the LLM outputs can be far from perfect, if we aim to use LLMs to assist rather than replace humans, then the goal would be for humans to serve as reviewers of the LLM outputs. For example, for corpus creation, a human can examine the LLM-generated essays to determine if they conform to the given constraints. For corpus annotation, a human can review the LLM-produced trait scores, written feedback, or even the revised essays, with the goal of determining whether reviewing

and correcting LLMs' outputs can help save a human annotator's time compared to performing these annotation tasks from scratch. If so, then there is value in using LLMs in assisting in these tasks. For instance, human trait scoring could be accelerated by understanding the LLM's rationale behind assigning a trait score.

Recommendation #6: We recommend that AES researchers explore novel ways of exploiting LLMs, especially concerning how LLMs can be used to improve human efficiency in completing daunting tasks such as corpus creation and annotation.

IX. ESSAY GRADING BEYOND SCORING

While AES has largely been tackled as a standalone task, researchers can think about how AES systems can be combined with other technologies developed for AI in education. For example, to improve the usability of AES systems in a classroom setting, AES models can be embedded in an intelligent tutoring system that can help an essay writer improve their essay in an iterative fashion through multiple rounds of feedback and revision.

Recommendation #7: We recommend that AES researchers think beyond essay scoring and develop a vision of how AES models can be combined with related AI technologies to develop more complex systems that can broaden the impact of AES research on education, ideally through discussions with other stakeholders such as teachers.

X. CONCLUSION

For the majority of the AES models proposed in the past few years, researchers have largely focused on improving performance numbers. While improving performance numbers should remain an important goal for AES, merely focusing on performance numbers is not necessarily healthy for AES research in the long run. In this position paper, we discussed a range of issues that we believe AES researchers should seriously consider as we move forward and provided seven recommendations based on our observations. We hope these recommendations will help initiate discussion among AES researchers and push the field forward.

LIMITATIONS

The views expressed in this position paper are necessarily subjective, so it is possible that the reader may disagree with some or all of our recommendations for future research directions. Nevertheless, we have tried to provide explanations behind our recommendations whenever possible. In addition, given the space limitations, we cannot provide a comprehensive overview of the related research in AES, so a reader without the relevant background (e.g., a reader who does not work on AES) may not be able to appreciate our recommendations. Nevertheless, we have tried our best to make this paper accessible to a general NLP audience.

---

[7]Care should be taken, however, to ensure that LLMs are not misused to generate essays with malicious content.

ETHICS STATEMENT

We hope that the issues raised in this paper as well as the views expressed in it would encourage AES researchers to start thinking about how the field should proceed, with the eventual goal of bringing the members of the field together to develop a shared vision. Note that developing a shared vision by no means implies that we are forcing all AES researchers to do the same thing. We view the shared vision as something that could help guide the long-term development of the field, but individual researchers will still have the freedom to (1) continue to work on their own projects; (2) work on projects that contribute to the shared vision; or (3) split their time between these two things.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.

[2] Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

[3] Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v. 2.0. *Journal of Technology, Learning, and Assessment*, 4(3).

[4] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Volume 1, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

[5] Beata Beigman Klebanov and Michael Flor. 2013. Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1148–1158, Sofia, Bulgaria. Association for Computational Linguistics.

[6] Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. 2016. Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72, San Diego, CA. Association for Computational Linguistics.

[7] Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

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

[9] 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.

[10] Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA. Association for Computational Linguistics.

[11] 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.

[12] Maalina 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.

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

[14] 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.

[15] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. Preprint, arXiv:2002.06305.

[16] Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

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

[18] Scott Elliot. 2003. Intellimetric: From here to validity. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*, pages 71–86. Lawrence Erlbaum Associates, Mahwah, NJ.

[19] Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74, Denver, Colorado. Association for Computational Linguistics.

[20] James Fiacco, David Adamson, and Carolyn Rose. 2023. Towards extracting and understanding the implicit rubrics of transformer based automatic essay scoring models. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 232–241, Toronto, Canada. Association for Computational Linguistics.

[21] Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

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

[23] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

[24] 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, Copenhagen, Denmark. Association for Computational Linguistics.

[25] Mohamed A. Hussein, Hesham A. Hassan, and Mohammad Nassef. 2020. A trait-based deep learning automated essay scoring system with adaptive feedback. *International Journal of Advanced Computer Science and Applications*, 11(5).

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

[27] Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4130–4136. International Joint Conferences on Artificial Intelligence Organization.

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

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

[30] Vivek Kulkarni, Kenny Leung, and Aria Haghighi. 2022. CTM- a model for large-scale multi-view tweet topic classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech-*

nologies: Industry Track, pages 247–258, Hybrid: Seattle, Washington+ Online. Association for Computational Linguistics.

[31] Vivekanandan S. Kumar and David Boulanger. 2021. Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, 31(3):538–584.

[32] 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.

[33] Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Prompting large language models for zero-shot essay scoring via multi-trait specialization. Preprint, arXiv:2404.04941.

[34] Shengjie Li and Vincent Ng. 2024a. Automated essay scoring: Recent successes and future directions. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 8114–8122, Jeju, Republic of Korea.

[35] Shengjie Li and Vincent Ng. 2024b. ICLE++: Modeling fine-grained traits for holistic essay scoring. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8465–8486, Mexico City, Mexico. Association for Computational Linguistics.

[36] Xia Li, Minping Chen, and Jian-Yun Nie. 2020. Sednn: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems*, 210:106491.

[37] Annie Louis and Derrick Higgins. 2010. Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–95, Los Angeles, California. Association for Computational Linguistics.

[38] Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? Preprint, arXiv:2403.06149.

[39] Jeziel C. Marinho, Rafael T. Anchiêta, and Raimundo S. Moura. 2021. Essay-BR: a Brazilian corpus of essays. In *Dataset Showcase Workshop (DSW)*, pages 53–64. Porto Alegre: Sociedade Brasileira de Computação.

[40] 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).

[41] Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA→ Online. Association for Computational Linguistics.

[42] Danielle S. McNamara, Scott A. Crossley, Rod D. Roscoe, Laura K. Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

[43] Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

[44] Huy V. Nguyen and Diane J. Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5892–5899, New Orleans, Louisiana. AAAI Press.

[45] Robert Östling, André Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for Swedish. In *Proc. of the BEA Workshop*, pages 42–47.

[46] Ellis B. Page. 1967. Grading essays by computer: Progress report. In *Proceedings of the Invitational Conference on Testing Problems*.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

[53] Robert Ridley, Liang He, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 13745–13753, Online. AAAI Press.

[54] Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. ArXiv, abs/2008.01441.

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

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

[57] Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of llms. Preprint, arXiv:2402.06925.

[58] Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

[59] Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

[60] 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.

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

[62] Sowmya Vajjala. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

[63] Daniel Varab and Natalie Schluter. 2020. DaNewsroom: A large-scale Danish summarisation dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6731–6739, Marseille, France. European Language Resources Association.

[64] Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

[65] Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

[66] Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yu-fang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. Preprint, arXiv:2401.06431.

[67] Chaojun Xiao, Xueyu Hua, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, 2:79–84.

[68] 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.

[69] 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.

[70] Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 33–43, Montréal, Canada. Association for Computational Linguistics.

[71] 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.

[72] Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado. Association for Computational Linguistics.