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

Shengiie Li and Vincent Ng Human Language Technology Research Institute University of Texas at Dal

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.

1 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 **??**, which is written in response to the prompt shown at the top of the table. Given the scoring rubric in Table **??**, 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 networks, but there is no elaboration on how computers enhance access to entertainment. Second, 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. Moreover, the transitions are awkward and repetitive. For example, the author uses "so as you can see" three times. Finally, 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 (Page, 1967), the task is still far from being solved.

While steady progress has been made on AES, much of the recent work in this area has focused on developing models that beat existing models on a standard evaluation dataset, often ASAP (Mathias and Bhattacharyya, 2018; Ridley et al., 2020, 2021; Li and Ng, 2024a). 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. Our goal in this position paper is to reflect on the state of the art in AES research and discuss 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.

[**Prompt**] More and more people use computers, but not everyone agrees 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.[0.5em] 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 newspape! @CAPS1 opinion about computers is that they benifetpeople. they help people when they need it, it gives people entertainment. So that is why computers benifet people. @CAPS1 first reason is that acomputer can help people when they need help like for instace if you have to write a paper andyou need to look somethingup 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 @CAPS 1 second resson is that itgives 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 themalso. So thatis why I said computers give you entertainmen too. @CAPS1 third resson is that it helpspeople such asif you in differentcontries and you can't go see them you to them through instans or web cam. Or ifyou 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 Isay computer are very benificalbe cause they give you entertainment they let you communicate and they help people so.

Table 1: A sample essay taken from Essay Set 1 of the ASAP corpus. The writing prompt is shown at the top.

2 Background

2.1 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++ (Mathias and Bhattacharyya, 2018) 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) TOEFL11 (Blanchard et al., 2013), a corpus of essays from the Native Language Identification task where the proficiency 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) (Yannakoudakis et al., 2011), where each essay is scored holistically and annotated with the linguistic error types it contains; (3) the International Corpus of Learner English (ICLE) (Granger et al., 2009), where a subset of essays has been scored not only holistically (Li and Ng, 2024b) but also along multiple dimensions of essay quality, such as Organization (Persing et al., 2010) and Argument Persuasiveness (Persing and Ng, 2015); and (4) the Argument Annotated Essays (AAE) corpus (Stab and Gurevych, 2014), where each persuasive essay is scored based on the strength of its thesis (Ke et al., 2019) and the persuasiveness of its argument (Ke et al., 2018).

AES corpora in other languages exist, such as Ostling's (2013) Swedish corpus, Horbach et al.'s (2017) German corpus, Marinho et al.'s (2021) Portuguese corpus, the GoodWriting dataset[1]](https://goodwriting.jp/wp/?lang=en) (in Japanese), and the MERLIN dataset[2]](https://www.merlin-platform.eu/), which is composed of German, Italian, and Czech essays.

2.2 Evaluation Metric

The standard metric used to evaluate AES models is Quadratic weighted Kappa (QWK).[3]) for details. QWK is an agreement metric that ranges from 0 to 1 but can be negative if there is less agreement 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.[4]

2.3 Systems

AES systems can be divided into three categories:

2.3.1 Heuristic Approaches

Virtually all early AES systems are heuristic-based and typically possess the following characteristics (e.g., Elliot (2003), Attali and Burstein (2006)):

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 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 (Attali and Burstein, 2006) 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.

[1] [https://goodwriting.jp/wp/?lang=en

[2] [https://www.merlin-platform.eu/

[3] See [https://www.kaggle.com/competitions/asap-aes/overview/evaluation](https://www.kaggle.com/competitions/asap-aes/overview/evaluation

[4] 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.

2.3.2 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) (Yannakoudakis et al., 2011; Vajjala, 2018), lexical features (e.g., the presence/count of each n-gram) (Chen and He, 2013; Phandi et al., 2015), word embeddings (Cozma et al., 2018), word category features (e.g., whether a word is a modal) (Farra et al., 2015; McNamara et al., 2015), and syntactic features (e.g., part-of-speech tag sequences) (Chen and He, 2013). High-level features include readability features (i.e., metrics that reflect how easy it is to read the essay) (Zesch et al., 2015), prompt-relevant features (i.e., features that encode the similarity between the essay and the prompt it was written for) (Louis and Higgins, 2010; Beigman Klebanov et al., 2016), argumentation features (e.g., the number of claims in a persuasive essay) (Ghosh et al., 2016; Wachsmuth et al., 2016; Nguyen and Litman, 2018), semantic features (e.g., features derived from lexico-semantic resources such as FrameNet (Baker et al., 1998)) (Beigman Klebanov and Flor, 2013), and discourse features (e.g., local coherence features derived from Centering Theory (Grosz et al., 1995)) (Yannakoudakis and Briscoe, 2012).

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 (Persing and Ng, 2013, 2014, 2015). 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.

2.3.3 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 (2016) 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. (2020) proposed R^{2}BERT, an AES model obtained by fine-tuning BERT. Wang et al. (2022) proposed a multi-scale BERT-based structure that captures (automatically learned) features at the token, segment, and essay levels. Uto et al. (2020) 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 (2016) 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 (Dong et al., 2017).

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 (Ridley et al., 2021), 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. (2016) 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. (2023) 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 (2023) 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 (Jin et al., 2018; Li et al., 2020; Ridley et al., 2020).

Multi-trait scoring. Since the holistic score of an essay is influenced by its trait-specific scores, it is natural to train a model to score an essay along multiple traits. One way to do so is to train multiple models, each of which is responsible for scoring one trait. An alternative is to train a single model that jointly predicts multiple trait scores. There are at least three approaches in the literature: (1) replacing the output layer of a holistic scoring model with multiple output layers, one for each trait (Hussein et al., 2020); (2) 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 (Mathias and Bhattacharyya, 2020); and (3) using the predicted trait scores as input to predict the holistic score (Kumar et al., 2022).

LLM-based approaches. Mizumoto and Eguchi (2023) 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 (Anthropic, 2024). 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. (2024)), or a few-shot setting, where a few labeled examples are provided as part of the prompt (e.g., Mansour et al. (2024); Xiao et al. (2024)).

3 Recommendations

3.1 Recommendation 1: Understand systems via analysis, not only metrics

There are various interpretability methodologies in machine learning that can be adapted to AES for increased transparency, such as LIME (Ribeiro et al., 2016).

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.

3.2 Recommendation 2: Evaluate beyond ASAP

What have we learned about AES over the years other than the fact that the QWK scores are improving on standard evaluation datasets? 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 conducted to address this question.

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 (Yannakoudakis et al., 2011) or a corpus where essays were written in a time-unrestricted setting, such as ICLE++ (Li and Ng, 2024b).

3.3 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 (2021) 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 evaluated it on essays for the held-out prompt. As a result, the training set could contain essays of a different type from those in the test set and could be scored using a different rubric. Consequently, it is not clear whether the performance obtained by their model on the held-out prompt truly reflects its ability to generalize to a new prompt.

Metrics. When evaluating cross-prompt AES models, one should ensure that the evaluation is meaningful. In particular, comparing QWK scores across prompts that have different rubrics and score ranges can be misleading. More analysis is needed to understand what is being measured and whether improvements are due to genuine cross-prompt generalization.

3.4 Recommendation 3: Reconsider evaluation and framing for cross-prompt AES

Recommendation 3: We recommend that researchers carefully design experimental settings for cross-prompt AES to ensure that (1) the training and test essays are of the same essay type, (2) the training and test essays are scored using the same (or highly similar) rubric, and (3) the reported performance reflects genuine generalization to unseen prompts rather than artifacts of the evaluation setup. [ILLEGIBLE]

3.5 Recommendation 4: Investigate trait impacts and LLMs for content-based traits

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 (Lee et al., 2024), including those that employ a chain-of-thought approach (Xiao et al., 2024), 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 (Mansour et al., 2024), changes in the decoding methods (Shi et al., 2024), and variations in random seeds (Dodge et al., 2020) 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.

4 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 (Varab and Schluter, 2020; Kulkarni et al., 2022). 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.

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 (Mansour et al., 2024), 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 such a corpus could facilitate not only the development of AES systems but also the development of systems that provide feedback to essay writers and systems that help essay writers revise their essays.

Recommendation 5: We recommend that the AES community allocate more effort to corpus development and annotation, and work towards creating richly annotated corpora that go beyond holistic scores.

5 Conclusion

In this position paper, we discussed a range of issues that need to be addressed for the long-term development of AES. We hope that our discussion can encourage AES researchers to think bigger than improving performance numbers on a standard evaluation dataset. We believe that a shared vision is something that could help guide the AES community on how we should move forward.

Ethics Statement

Automated essay scoring models can have significant societal impact, especially when used in educational settings. Researchers and practitioners should consider issues such as fairness, transparency, and the consequences of deployment. [ILLEGIBLE]

Acknowledgements

[MISSING]

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. Claude 3.5 sonnet system card. Preprint, arXiv:2410.10065.

[3] Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater$^{®}$ v.2. *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 *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

[5] Beata Beigman Klebanov and Michael Flor. 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.

[6] Beata Beigman Klebanov, Nitin Madnani, and Jill Burstein. 2016. Using discourse markers for automatic essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1061–1071, Austin, Texas. Association for Computational Linguistics.

[7] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. In *ETS Research Report Series.*

[8] Hongbo Chen and Ben He. 2013. 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.

[9] Menglin Chen and Shengjie Li. 2023. Cross-prompt essay scoring via contrastive learning. [ILLEGIBLE]

[10] Mihai Cozma, Andrei Butnaru, and Cristian Cercel. 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Fei Dong, Yue Zhang, and Jie Yang. 2011. 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.

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

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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).

[24] 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.

[25] 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.

[26] Zixtan 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.

[27] 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.

[28] 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 Technologies: Industry Track*, pages 247–258, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Shengie 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.

[34] 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.

[35] 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.

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

[37] Jeziel C. Marinho, Rafael T. Anchi6ta, 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 Computaqdo.

[38] 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).

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] Robert Ostling, Andr6 Smolentzov, Bj6rn Tyrefors Hinnerich, and Erik Hoglin. 2013. Automated essay scoring for Swedish. In *Proc. of the BEA Workshop*, pages 42–47.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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, Be4ing, China. Association for Computational Linguistics.

[49] 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.

[50] 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.

[51] 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.

[52] 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, abs12008.0144 1, 17887.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

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

[65] 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.

[66] Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qi. 2022. Automated essay scoring via pairwise contrastive regression. [MISSING]

[67] Jianhui Yang, Lixin Duan, and [ILLEGIBLE]. 2020. R^2BERT: [ILLEGIBLE] [MISSING]

[68] 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.

[69] 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, Montreal, Canada. Association for Computational Linguistics.

[70] 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.

Score	Rubric description
	1
response may a no more very support. (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 general reasons with 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 mini-	[MISSING] 5