Automated Essay Scoring: Recent Successes and Future Directions

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

Abstract

Automated essay scoring (AES), the task of automatically assigning a score to an essay that summarizes its quality, is a challenging task that remains largely unsolved despite more than 50 years of research. This survey paper discusses the milestones in AES research and reflects on future directions.

1 Introduction

Automated essay scoring (AES), the task of automatically scoring an essay written for a given *prompt* (i.e., writing topic, such as "Write a persuasive essay on why one should (or shouldn't) support Obamacare."), is an important educational application of natural language processing (NLP). While the vast majority of work on AES has focused on *holistic* scoring, where the goal is to assign a score to an essay that summarizes its overall quality, the past few years have seen increasing interest in *trait-specific* scoring, where the goal is to assign a score to an essay along a specific dimension of essay quality (a.k.a. trait), such as Organization, Coherence, and Prompt Adherence. Ever since its inception more than 50 years ago [Page, 1966], AES has remained an active area of research.

Researchers' unfaltering interest in AES can in part be attributed to its practical significance. For instance, AES systems have been deployed to holistically score the large number of essays written for standardized aptitude tests such as GRE and SAT every year with the goal of reducing human grading effort. In classroom settings where providing feedback to essay writers is crucial, trait-specific scoring can be employed to provide students with feedback on which essay traits need improvement if she receives a low holistic score.

While AES can naturally be recast as a text classification or regression task, where the goal is to either classify an essay as belonging to a particular score category or predict the (real-valued) score of an essay, it is arguably more challenging than many well-known text classification/regression tasks such as topic classification (the task of classifying a text document as belonging to one of a predefined set of categories). The reason is that many traits play a role in holistic scoring. To understand what these traits are, consider the rubric used to score the GRE essays written for the "Analyze an Issue" task (Table 1), which involves writing an essay that analyzes the issue described in the essay prompt by taking a stance and providing evidence in support of the chosen stance. From the five bullets that describe what typically defines a "6" (i.e., "outstanding") essay, we can infer that the traits that impact the holistic score include (1) the clarity of the essay's thesis (first bullet), (2) development (second bullet), (3) the persuasiveness of the argument (second bullet), (4) organization (third bullet), (5) coherence (third bullet), and (6) technical quality such as fluency, grammar, and mechanics (last two bullets). These traits can 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). Holistic scoring is further complicated by the fact that the content-based traits are much harder to score than their non-content-based counterparts. 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.

Score 6 (Outstanding): A 6 response presents a cogent, well-articulated analysis of the issue and conveys meaning skillfully. A typical response in this category:

- articulates a clear and insightful position on the issue in accordance with the assigned task
- develops the position fully with compelling reasons and/or persuasive examples
- sustains a well-focused, well-organized analysis, connecting ideas logically
- conveys ideas fluently and precisely, using effective vocabulary and sentence variety
- demonstrates superior facility with the conventions of standard written English (i.e., grammar, usage and mechanics), but may have minor errors

Table 1: Partial rubric for the GRE "Analyze an Issue" task.

Our goal in this paper is to provide the general AI audience with a high-level overview of AES research.[1] While AES research has primarily been conducted in the NLP and AI in

[1] In comparison to our previous survey on this topic [Ke and Ng,

Corpora	Writer's Language Level	Essay Types	# Essays	# Prompts	Score Range	Note
ASAP	US 7th to 10th grade students	A,R,N	17450	8	as small as [0-3]; as large as [0-60]	ASAP++: trait scoring for 6 prompts and 10696 essays
TOEFL11	Non-native TOEFL test takers	A	1100	8	Low, Medium, High	–

Table 2: Two popularly used corpora for holistic scoring.

Education (AIED) communities, we believe that AES would be of interest to the broader AI community. As mentioned above, AES can naturally be recast as a classification or regression task, potentially making AES an interesting testbed for applied machine learning researchers. In fact, AES has already attracted a lot of attention from machine learning researchers as one of the earlier Kaggle competitions has focused on holistic scoring. In contrast, trait scoring has received less attention, but the fact that content-based trait scoring is more challenging than holistic scoring may stimulate the interest of those in machine learning who seek to work on challenging problems. In the era of Generative AI, one could even think about using large language models (LLMs) for automatically scoring essays by presenting scoring rubrics to them as part of the instructions. However, our preliminary experiments revealed that state-of-the-art LLMs (e.g., ChatGPT[2]), when used in a zero-shot setting (where rubrics were presented without any annotated essays as demonstrations) or a few-shot setting (where rubrics in combination with a small number of annotated essays were presented) performed considerably worse than their supervised counterparts, suggesting that LLMs' interpretation of scoring rubrics is quite different from human interpretation. How to profitably exploit LLMs for AES remains an intriguing question that could be of interest to Generative AI researchers.

2 Corpora and Evaluation Metrics

Corpora. While a number of AES corpora have been developed, two of them stand out as being the most extensively used in AES research in the past couple of years. Table 2 compares these two corpora along five dimensions: (1) the types of essays present in the corpus (argumentative (A), response (R), narrative (N)); (2) the language level of the essay writers; (3) the number of essays; (4) the number of prompts; and (5) the score range of the essays (i.e., the possible scores a human rater can assign to an essay). Each prompt is associated with a rubric (see Table 1 for an example), which specifies the score range and the meaning associated with each score. Each essay in an AES corpus is scored holistically by human raters using the corresponding rubric. In some corpora, essays may additionally be scored along different traits.

Introduced as part of a 2012 Kaggle competition, the Automated Student Assessment Prize (ASAP[3]) corpus has become a popular dataset for training and evaluating holistic scoring models, especially given its vast collection of essays per prompt (up to 3,000 for some prompts). ASAP++ [Mathias and Bhattacharyya, 2018] is an extension of ASAP where each essay is scored along multiple traits. Note that ASAP is composed of three types of essays: narrative essays, persuasive essays, and source-dependent essays. Not all traits are applicable to all essay types. For instance, while Content is scored for all three essay types, Organization is scored for narrative essays and persuasive essays only.

The TOEFL11 corpus [Blanchard *et al.*, 2013] comprises essays from the TOEFL exam. These essays are evenly distributed across eight prompts and are authored by writers from 11 different native languages. While the corpus is initially compiled for the Native Language Identification task (the task of determining the essay writer's native language), it also provides a coarse proficiency categorization into three levels: Low, Medium, and High. Some researchers have used these proficiency labels as holistic scores for the essays when training AES systems. However, the assumption that an essay's quality can be solely determined by the language proficiency of its author remains debatable.

Other English corpora that have been developed for and used in AES research include (1) the Cambridge Learner Corpus-First Certificate in English exam [Yannakoudakis *et al.*, 2011], where each essay is scored holistically and annotated with the linguistic error types it contains; (2) the International Corpus of Learner English, where a subset of essays has been scored along multiple dimensions of essay quality, including Organization [Persing *et al.*, 2010], Thesis Clarity [Persing and Ng, 2013], Prompt Adherence [Persing and Ng, 2014], and Argument Persuasiveness [Persing and Ng, 2015]; and (3) the Argument Annotated Essays corpus [Stab and Gurevych, 2014], where each essay is scored based on the strength of its thesis [Ke *et al.*, 2019] and the persuasiveness of the argument it makes [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[4] (in Japanese), and the MERLIN dataset[5], which is composed of essays written in German, Italian, and Czech.

Evaluation metrics. The standard metric used to evaluate AES models is Quadratic Weighted Kappa (QWK). 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. Specifically, QWK is a weighted version of Kappa [Carletta, 1996] where each case of disagreement (i.e., the (rounded) system-predicted score is different from the reference/human-

2019], this survey not only covers the latest work on neural AES, but also offers a broader perspective on AES research by describing how recent work is related to early work on AES. For a comprehensive review of AES research, we refer the reader to the books published by Shermis and Burstein [2003; 2010; 2013] and Beigman Klebanov and Madnani [2021].

[2] https://chat.openai.com/

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

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

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

assigned 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. To compute QWK, we need to construct three matrices. First, a weight matrix **W** is computed as follows:

$$\mathbf{W}_{ij} = \frac{(i-j)^2}{(N-1)^2}$$

where i and j are the reference score and the predicted score respectively and N is the number of possible scores. The second matrix, **O**, is constructed where $\mathbf{O}_{ij}$ is the number of essays for which the reference score is i and the predicted score is j. Finally, an expected count matrix, **E**, is computed as the outer product of the histogram vectors corresponding to the reference ratings and the predicted ratings, and then normalized so that the sum of the elements in **E** is equal to the sum of the elements in **O**. Using these three matrices, we can compute QWK as follows:

$$\text{QWK} = 1 - \frac{\sum_{i,j} \mathbf{W}_{ij}\mathbf{O}_{ij}}{\sum_{i,j} \mathbf{W}_{ij}\mathbf{E}_{ij}}$$

Since QWK is an agreement metric, higher values are better.

Three other metrics have also been used to evaluate AES models although they are less extensively used than QWK. The first metric, Pearson's Correlation Coefficient (PCC), is a *correlation* metric that measures the correlation between the predicted scores and the reference scores. Higher PCC values are better. The second metric, mean squared error (MSE), is an *error* metric that measures the average square of the distance between a predicted score and the corresponding reference score. The intuition behind this metric is that not only should we prefer a model whose estimations are close to the reference scores, but we should also prefer one whose estimations are not too frequently very far away from the reference scores. Lower MSE scores are better. The third metric, mean absolute error (MAE), is an *error* metric that measures the average absolute distance between the predicted score and the reference score. Lower MAE scores are better.

3 Systems

Existing AES systems can be divided into three categories.[6]

3.1 Heuristic Approaches

Virtually all early AES systems are *heuristic-based* and typically possess the following characteristics (e.g., Page [1966], Elliot [2003], Attali and Burstein [2006]):

Trait-driven holistic scoring. Motivated by the human essay scoring process, the holistic score returned by a heuristic AES system is typically computed as the weighted or unweighted sum of the trait-specific 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. As mentioned before, computing content-based trait scores is challenging given that an understanding of the content (as opposed to its format) is required, and these 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.

While heuristic approaches are efficient (because no training is required) and the holistic score is interpretable (since it is computed based on the trait-specific scores), they tend to underperform competing approaches because (1) they fail to exploit content-based traits and (2) heuristic computation of even the non-content-based traits can be inaccurate.

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 both *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 an n-gram) [Chen and He, 2013; Phandi *et al.*, 2015; Zesch *et al.*, 2015], word embeddings [Cozma *et al.*, 2018], *word category* features (e.g., whether a word is a modal) [Breland *et al.*, 1994; Farra *et al.*, 2015; McNamara *et al.*, 2015; Amorim *et al.*, 2018], and *syntactic* features (e.g., part-of-speech tag sequences) [Yannakoudakis *et al.*, 2011; Chen and He, 2013; Zesch *et al.*, 2015]. 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 is written for), *argumentation* features (e.g., the number of claims and premises in each paragraph of a persuasive essay) [Ghosh *et al.*, 2016; Wachsmuth *et al.*, 2016; Nguyen and Litman, 2018], *semantic* features (e.g., features computed based on lexico-semantic resources such as FrameNet [Baker *et al.*, 1998]) [Beigman Klebanov and Flor, 2013; Persing and Ng, 2013], and *discourse* features (e.g., local coherence features derived from Centering Theory [Grosz *et al.*, 1995]) [Morris and Hirst, 1991; Yannakoudakis and Briscoe, 2012; Somasundaran *et al.*, 2014].

[6] Virtually all approaches to AES for non-English languages are motivated by those developed for English. Hence, while our discussion in this section is focused on English AES, it can broadly be viewed as covering work in both English and non-English AES.

Focus on within-prompt scoring. In within-prompt scoring, an AES model is trained on essays written for a given prompt and then applied to test essays written for the same prompt. In other words, the same prompt is used for training and testing. While researchers have made significant progress on within-prompt scoring, some have argued that within-prompt scoring is not a practical setting for AES. The reason is that when these within-prompt scorers are applied to essays written for a new prompt, their performance often deteriorates considerably. Hence, in practice, before they are applied to score essays written for a new prompt, they need to be retrained on human-scored essays written for the new prompt. However, manually scoring essays is a time-consuming process 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 different traits are scored for each essay, such as Organization [Persing *et al.*, 2010], Thesis Clarity [Persing and Ng, 2013], Prompt Adherence [Persing and Ng, 2014], and Argument Persuasiveness [Persing and Ng, 2015]. While the scoring of content-based traits is largely ignored in heuristic approaches, researchers have begun training models for scoring content-based traits. Nevertheless, empirical results suggest that even with annotated data, the scoring of content-based traits remains a challenging task.

Researchers have mixed feelings about feature engineering for AES. Some believe that it offers the flexibility to incorporate both simple and sophisticated features into AES systems, while others think that (1) identifying useful features is time-consuming and (2) the use of low-level features deviates too much from the human essay scoring process. Nevertheless, machine learning approaches to AES have been shown to offer superior performance to their heuristic counterparts.

3.3 Neural Approaches

With the advent of the neural NLP era, the vast majority of recently-developed AES models are based on deep neural networks.[7] AES research during this period can roughly be summarized as (1) a focus on learning the distributed (i.e., real-valued vector) *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.

Within-Prompt Scoring

As noted before, in within-prompt scoring, an AES model is trained on essays written for a given prompt and then applied to test essays written for the same prompt. In other words, the same prompt is used for training and testing. Virtually all early neural approaches to AES have focused on within-prompt scoring, as described below.

Combining CNNs and RNNs. Recall that Convolutional Neural Networks (CNNs) can capture spatial dependencies whereas Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) networks can capture temporal dependencies. Motivated by this observation, Taghipour and Ng [2016] first employ a CNN to extract local features from an input essay, focusing on n-gram-level textual dependencies between words. These extracted local features are then fed into a LSTM network, generating a long-distance representation of the essay at each time step. Its final representation, which is used to predict its holistic score, is obtained by averaging the representations across all time steps.

Exploiting document structure. A weakness of Taghipour and Ng's [2016] approach is that it treats the input essay as a mere sequence of words without exploiting the essay's structure. As a result, Dong and Zhang [2016] view 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 design a two-layer CNN model for holistic scoring where the first layer creates a representation for each sentence via the n-gram features extracted from it and the second layer creates an essay representation by combining the sentence representations.

Exploiting attention pooling. Recall that an attention mechanism is a mechanism that enables us to automatically identify the important portions of an input sequence (e.g., words, phrases). Building on the efforts of Taghipour and Ng [2016] and Dong and Zhang [2016], Dong *et al.* [2017] employ an attention pooling mechanism in these models.

Despite early successes, RNNs and CNNs have their own weaknesses: RNNs are weak at capturing long-term dependencies and do not permit parallel training, while CNNs are constrained by their filter sizes. As a result, they are gradually phased out by Transformer-based neural models [Yang *et al.*, 2020; Uto *et al.*, 2020; Wang *et al.*, 2022], which not only address the aforementioned weaknesses associated with CNNs and RNNs, but also possess a vast amount of linguistic knowledge and commonsense knowledge acquired solely from large unlabeled text corpora via a process known as pre-training [Dai and Le, 2015; Peters *et al.*, 2018; Radford *et al.*, 2018; Howard and Ruder, 2018]. Many pre-trained Transformer models have been developed in the past few years, such as BERT [Devlin *et al.*, 2019], RoBERTa [Liu *et al.*, 2019], and ELECTRA [Clark *et al.*, 2020]. *Fine-tuning* these pre-trained models on task-specific training data in the usual supervised manner has enabled state-of-the-art results to be achieved on a wide variety of NLP tasks. Below we discuss work on Transformer-based AES models.

Basic fine-tuning. The $\mathtt{R}^2\mathtt{BERT}$ model proposed by Yang *et al.* [2020] represents an early effort in fine-tuning pre-trained Transformers for AES. Specifically, $\mathtt{R}^2\mathtt{BERT}$ is obtained by fine-tuning BERT using a combination of regression loss and ranking loss. The ranking loss aids the model in learning an accurate ranking order, mirroring rankings based on the reference scores.

Exploiting essay structure. In a subsequent development, Wang *et al.* [2022] propose a multi-scale BERT-based struc-

[7] Strictly speaking, deep learning approaches are a form of machine learning approaches, but they have become a league of their own given their recent successes and deserve a separate discussion.

ture that captures automatically learned features at the token, segment, and essay levels. In addition to regression and ranking losses, they implement a similarity loss function, leveraging the cosine similarity between the predicted scores and the gold scores within a batch of input essays in order to help the model learn the correlation among the essays.

Using contrastive learning. Xie *et al.* [2022] propose a holistic scoring model, `NPCR`, which, like the R^2`BERT` model mentioned earlier, integrates regression and ranking objectives, but additionally employs contrastive learning. The contrastive learning objective enables their model to acquire better essay representations in the sense that it helps to bring similar essays closer together (in terms of their representations) and push dissimilar essays further apart.

Feature engineering. A key strength of neural approaches is that features can be extracted automatically, thus obviating the need for manual feature engineering. While in all the aforementioned neural AES models the features are extracted automatically, Uto *et al.* [2020] hypothesize that neural AES models could be improved with hand-crafted features. As a result, they integrate several hand-crafted, essay-level features with the essay representations obtained from BERT and subsequently fine-tune their AES model.

Multi-trait scoring. Recall that using trait scores for holistic scoring is one of the major themes in heuristic approaches. Specifically, a pipeline architecture is adopted where the traits are heuristically scored and subsequently combined to predict the holistic score. The neural era witnesses a return of this idea of using traits for holistic scoring, but the focus is the development of *joint* models that can simultaneously predict the holistic score and the various trait-specific scores. In other words, unlike in the heuristic era, in these joint models, the traits are scored jointly rather than independently of each other using learning and not via heuristics. We refer to the underlying learning task as *multi-trait scoring*.

Early multi-trait scoring models are obtained by a simple extension of existing holistic scoring models. For instance, Hussein *et al.*'s [2020] `AES aug` model extends existing CNN- and RNN-based holistic scorers for multi-trait scoring simply by replacing the output layer with multiple output layers, one for each trait. In essence, their model allows holistic scoring and trait-specific scoring to influence each other via one or more shared layers in the network, but it does not necessarily enable trait-specific scoring to *directly* influence holistic scoring. Kumar *et al.*'s [2022] joint model is composed of multiple copies of Dong *et al.*'s [2017] holistic scorer, where each copy is responsible for scoring one trait. Similar to `AES aug`, Kumar *et al.*'s model has a representation layer that is shared by all the copies. However, to enable the predicted trait scores to influence holistic scoring, Kumar *et al.*'s model incorporates the predicted trait scores as input when predicting the holistic score.

Cross-Prompt Scoring

As noted before, within-prompt scoring may not be a practical setting for AES, as an AES scorer typically does not perform well on essays written for a new prompt unless it is (re)trained on essays from the new prompt. Motivated by this observation, researchers have begun to examine *cross-prompt* scoring: given training essays written for a set of prompts, the goal is to train a model to score essays written for prompts that are not seen during training.

Training a model that can generalize to unseen prompts is ambitious. To see why, consider the task of scoring essays written for the prompt "Write a persuasive essay on why one should (or should not) support Obamacare". Intuitively, a high-scoring essay should provide evidence(s) that can adequately support the claim of why one should (or should not) support Obamacare. However, determining whether an argument is persuasive could require domain knowledge (in this case background knowledge about Obamacare), which the model may not possess in the absence of training data for the new prompt. Nevertheless, there have been several attempts at cross-prompt scoring, as discussed below.

Recasting cross-prompt scoring as domain adaptation. AES researchers have recast cross-prompt scoring as domain adaptation [Phandi *et al.*, 2015; Cummins *et al.*, 2016]. In this setting, each essay prompt is viewed as a "domain", so that transfer learning techniques can be employed to adapt a model trained on the existing prompts (i.e., the "source domains") to a new prompt (i.e., the "target domain"). Note that Phandi *et al.*'s approach and Cummins *et al.*'s approach are both developed for a soft version of cross-prompt scoring where a small number of essays from the target prompt are available for model training in addition to a large number of essays from the source prompts. For instance, Cummins *et al.* (1) encode an essay using two vector spaces (one is shared by the source prompts and the target prompt and the other is target prompt-specific), (2) concatenate the two representations, (3) use the resulting representation to train a pairwise ranking model to rank essays, and (4) use the representation and the ranking model's weights to train a linear regressor for predicting the holistic score.

Distinguishing prompt-independent features from prompt-dependent features. The hand-crafted features researchers designed for use in machine learning approaches to AES can be broadly divided based on whether they are prompt-independent or prompt-dependent. Prompt-independent features are features that can generalize across prompts (e.g., the number of spelling errors), whereas prompt-dependent features are features whose usefulness may be dependent on the prompt (e.g., word unigrams).

Jin *et al.* [2018] propose a two-stage learning framework for cross-prompt scoring that exploits both prompt-independent and prompt-dependent features. In the first stage, they train a ranking model using only prompt-independent features on the source essays and apply this model to identify extremely good or poor target essays (via their ranks). Their assumptions are that (1) training the model on the source essays using only prompt-independent features facilitate generalization to the target prompt; and (2) essays of extreme quality can be identified using prompt-independent features. Given the ranks, the model assigns pseudo-labels to the target essays: 0 for the poor ones and 1 for the good ones. In the second stage, they train a regressor on these pseudo-labeled target essays using prompt-dependent features to predict the scores for the remaining target essays (i.e., essays that have not been assigned pseudo-labels). Their assumptions are that (1) the remaining essays are difficult to score, so prompt-dependent features need to be used that can capture prompt-specific information; and (2) since the remaining essays are not of extreme quality, using a regressor will naturally assign to them a value between 0 and 1.

Combining automatically learned features with hand-crafted features. While Jin *et al.* [2018] rely on only hand-crafted features, PAES [Ridley *et al.*, 2020], an AES model designed for cross-prompt scoring, employs both a diverse set of hand-crafted prompt-independent features and features automatically extracted from the given essay. To facilitate generalization across prompts, each input essay is represented as a sequence of part-of-speech (POS) tags.

Combining cross-prompt scoring with multi-trait scoring. Ridley *et al.* [2021] propose CTS, a model that combines cross-prompt scoring and multi-trait scoring by extending PAES. Like PAES, CTS employs a shared CNN layer combined with attention pooling to distill common features from the input essays, each of which is represented as a POS sequence. To jointly predict the holistic score and the trait scores, CTS (1) integrates a trait-specific LSTM layer with subsequent attention pooling atop these common features for extracting trait-specific features; (2) concatenates these trait-specific features with the hand-crafted prompt-independent features; and (3) uses the combined set of features to predict the score of the corresponding trait.

Incorporating prompt information. A shortcoming of the CTS model is its exclusion of essay prompts as input: without prompt information, it is not possible to determine whether the essay is off-topic or not. In light of this problem, Do *et al.* [2023] introduce ProTACT, a prompt-aware multi-trait cross-prompt scoring model that refines CTS by (1) incorporating a system to obtain a representation of the prompt; (2) combining this prompt representation with the input essay's representation using a multi-head attention mechanism [Vaswani *et al.*, 2017] and attention pooling; and (3) combining the resulting representation with prompt-independent features to predict the holistic score of the given essay. Jiang *et al.* [2023] propose another prompt-aware model, PANN, which obtains prompt adherence features by (1) calculating the cosine similarity between each word in the prompt and the essay; (2) applying RBF kernel pooling to the resulting similarity matrix; and finally (3) employing attention pooling to acquire the prompt adherence features.

Using contrastive learning. Since an essay is written for a specific prompt, its representation naturally encodes prompt-specific information. This is precisely what makes cross-prompt scoring difficult: since essay representations contain prompt-specific information, the representations of the source essays tend to be different from those of the target essays, and consequently, a model trained on the source essays will likely perform poorly on the target essays.

The solution proposed by the aforementioned approaches is to employ *features* that can generalize across prompts. These include hand-crafted features (i.e., prompt-independent features) as well as features automatically derived from generalized essay representations (e.g., essays represented as POS tag sequences).

```
I would like you to mark an essay written
by English as a foreign language (EFL)
Learners.  Each essay is assigned a
rating of 0 to 9, with 9 being the
highest and 0 the lowest.  You don't have
to explain why you assign that specific
score.  Just report a score only.  The
essay is scored based on the following
rubric.

[IELTS rubric in plain text format.]

ESSAY:
[Input essay]
```

Figure 1: Example prompt used in prompt-based AES.

Since the root cause of the generalization problem is that the representations of the source essays and target essays are different, Chen and Li [2023] propose PMAES, a model for cross-prompt scoring that seeks to make the source essay representations and the target essay representations more *consistent* with each other via contrastive learning and use the refined representation in combination with hand-crafted features for holistic scoring. Given the complexity of their approach, we refer the reader to their paper for details.

Prompt-Based Approaches

While the majority of recently-developed neural AES models involve fine-tuning a pre-trained model on AES training data, Mizumoto and Eguchi [2023] investigate a prompt-based approach to AES. Here, the word "prompt" does not refer to "essay prompt"; rather, it refers to the use of prompting in the context of LLMs. Specifically, a user interacts with a LLM via an interface where the user can enter natural language instructions on the task that she expects the LLM to perform. The entire set of instructions is known as a prompt.

Prompt-based approaches to AES 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 extremely good at understanding complex natural language instructions. Given these strengths, it is conceivable that we can ask a LLM to perform a task as complex as AES by providing detailed natural language instructions in the form of a prompt. A sample prompt that Mizumoto and Eguchi [2023] provide to GPT-3.5 (text-davinci-003) for AES is shown in Figure 1. As we can see, the prompt begins with a general task definition, followed by the scoring rubric (not shown due to space limitations), as well as the input essay. Given this prompt, GPT-3.5 is expected to score the given essay according to the instructions.

GPT-3.5, when used in the way described above, is effectively performing zero-shot learning for AES, where *no* manually scored essays are provided as training examples (e.g., Lee *et al.* [2024]). In other words, we are simply relying on the commonsense knowledge inherent in GPT-3.5 for AES. It is possible to use GPT-3.5 (or other LLMs) for few-shot learn-

			Trait										
Setting	**System**	Hol.	Cont.	Org.	WC	SF	Conv.	PA	Lan.	Nar.	Style	Voice	Avg.
Within-prompt STL	NPCR [Xie *et al.*, 2022]	.82	–	–	–	–	–	–	–	–	–	–	–
Within-prompt MTL	Kumar *et al.* [2022]	.76	.68	.61	.61	.60	.56	.70	.60	.67	.63	.58	.64
Cross-prompt STL	PANN [Jiang *et al.*, 2023]	.70	–	–	–	–	–	–	–	–	–	–	–
Cross-prompt MTL	ProTACT [Do *et al.*, 2023]	.67	.60	.52	.60	.59	.45	.62	.60	.64	–	–	.59

Table 3: QWK scores of state-of-the-art AES systems under different settings. The traits are Content (Cont.), Organization (Org.), Word Choice (WC), Sentence Fluency (SF), Conventions (Conv.), Prompt Adherence (PA), Language (Lang.), Narrativity (Nar.), Style and Voice. Avg. represents the QWK scores averaged over all traits, and Hol. shows the holistic scores.

ing, where a few labeled examples are provided as part of the prompt (e.g., Mansour *et al.* [2024], Xiao *et al.* [2024]).

4 The State of the Art

In Table 3, we enumerate the systems that have achieved state-of-the-art results on the ASAP dataset under four settings. As can be seen, the four settings differ in terms of (1) whether *within-prompt scoring*, where models are trained and tested on the same prompt, or *cross-prompt scoring*, where models are trained and tested on different prompts, are used; and (2) whether *single-task learning* (STL), where the model is trained to predict only the holistic score, or *multi-task learning* (MTL), where the model is jointly trained to predict the holistic score and the trait scores, is used.

Several points deserve mention. First, the within-prompt scoring results are always at least as good as the corresponding cross-prompt scoring results. This is perhaps not surprising as cross-prompt scoring is intuitively a more challenging task setting than within-prompt scoring. Nevertheless, one should keep in mind that these two sets of results are not directly comparable as the models are trained and tested on different partitions of the ASAP dataset. Second, while the MTL experiments seem to suggest that the holistic scoring results are better than the trait scoring results, the caveat is that the results are not directly comparable: since not all essays have gold trait scores, the dataset for learning holistic scores is larger than those for learning the trait-specific scores. Finally, two traits, Style and Voice, are present in only one prompt, so no cross-prompt results can be obtained for them.

5 Concluding Remarks

We conclude with a brief reflection on the future of AES.

Tasks. Two relatively new tasks in AES deserve more attention from AES researchers. Cross-prompt AES could ease the applicability of AES models to new prompts, whereas multi-trait scoring is a crucial step towards *explainable AES*. Explainable AES is important in that it not only forces an AES model to reason like human raters but also provides feedback to essay writers on which traits need improvement. Both tasks are challenging. For cross-prompt models to work well, prompt-specific knowledge is often needed. For multi-trait scoring, an understanding of essay content is needed when scoring content-based traits. LLMs could be exploited for these tasks: given that LLMs possess commonsense knowledge, prompting techniques can be designed to elicit domain knowledge for cross-prompt AES; and since LLMs are good at understanding complex inputs, it is worth examining how well they can understand inputs as complex as essays.

Data. Corpus development for AES research has not been able to keep up with model development. The majority of the recently developed models for AES are evaluated solely on the ASAP corpus. ASAP, however, is not without its limitations. Recall that the ASAP essays are written by U.S. students between grades 7 and 10 in a time-restricted setting. Hence, it is unclear whether a model trained on ASAP is generalizable to corpora composed of essays written by learners of English as a second language or those written in a time-unrestricted setting, for instance. In light of this concern, we recommend that researchers consider annotating new AES corpora. As a first step, the community can discuss what corpora would best complement ASAP and other existing corpora in evaluating the generalizability of AES models. Despite being an arduous task, we believe that corpus annotation is beneficial for the long-term development of the field.

Analysis. In AES research, results are rarely accompanied by an analysis of model outputs. Two kinds of analysis would be desirable for AES research. *Comparative analysis* seeks to reveal the relative strengths and weaknesses of two models. For instance, if results show that a newly proposed model outperforms a baseline model, comparative analysis could help us understand what exactly has been improved. In contrast, *error analysis* reveals the major sources of error a model makes and thus helps us identify areas of improvement.

The typical lack of analysis of AES model outputs implies that it is not always clear what has been improved despite performance improvements. Having said that, manually conducting such analysis, particularly in the neural NLP era in which model outputs are difficult to interpret, is challenging. In the long run, researchers should examine the possibility of developing tools for automatic analysis of AES outputs. As a starting point, one could consider conducting the analysis based on the trait-specific scores, which can shed light on the comparative strengths and weaknesses of two systems w.r.t. trait scoring as well as the traits a model is weak at scoring.

Evaluation. So far AES systems have largely been evaluated intrinsically. However, if an AES system is to be deployed in a classroom setting, extrinsic evaluations, which are based on the feedback provided on the usefulness of the system by its users (e.g., teachers and essay writers), are often more important than intrinsic evaluations. We recommend that AES researchers consider broadening the impact of their research by discussing with other stakeholders on how their systems can be evaluated extrinsically.

References

[Amorim *et al.*, 2018] E. Amorim, M. Cançado, and A. Veloso. Automated essay scoring in the presence of biased ratings. In *NAACL*, 2018.

[Attali and Burstein, 2006] Y. Attali and J. Burstein. Automated essay scoring with E-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 2006.

[Baker *et al.*, 1998] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet project. In *ACL*, 1998.

[Beigman Klebanov and Flor, 2013] B. Beigman Klebanov and M. Flor. Word association profiles and their use for automated scoring of essays. In *ACL*, 2013.

[Beigman Klebanov and Madnani, 2021] B. Beigman Klebanov and N. Madnani. *Automated Essay Scoring*. In G. Hirst, editor, *Synthesis Lectures in Human Language Technologies*. Morgan & Claypool Publishers, 2021.

[Blanchard *et al.*, 2013] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013.

[Breland *et al.*, 1994] H. M. Breland, R. J. Jones, L. Jenkins, M. Paynter, J. Pollack, and Y. F. Fong. The college board vocabulary study. *ETS Research Report Series*, 1994.

[Carletta, 1996] J. Carletta. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 1996.

[Chen and He, 2013] H. Chen and B. He. Automated essay scoring by maximizing human-machine agreement. In *EMNLP*, 2013.

[Chen and Li, 2023] Y. Chen and X. Li. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *ACL*, 2023.

[Clark *et al.*, 2020] K. Clark, M. Luong, Q. V. Le, and C. D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.

[Cozma *et al.*, 2018] M. Cozma, A. Butnaru, and R. T. Ionescu. Automated essay scoring with string kernels and word embeddings. In *ACL*, 2018.

[Cummins *et al.*, 2016] R. Cummins, M. Zhang, and T. Briscoe. Constrained multi-task learning for automated essay scoring. In *ACL*, 2016.

[Dai and Le, 2015] A. M. Dai and Q. V. Le. Semi-supervised sequence learning. In *NeurIPS*, 2015.

[Devlin *et al.*, 2019] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[Do *et al.*, 2023] H. Do, Y. Kim, and G. G. Lee. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of ACL*, 2023.

[Dong and Zhang, 2016] F. Dong and Y. Zhang. Automatic features for essay scoring – an empirical study. In *EMNLP*, 2016.

[Dong *et al.*, 2017] F. Dong, Y. Zhang, and J. Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *CoNLL*, 2017.

[Elliot, 2003] S. Elliot. Intellimetric: From here to validity. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*. 2003.

[Farra *et al.*, 2015] N. Farra, S. Somasundaran, and J. Burstein. Scoring persuasive essays using opinions and their targets. In *BEA Workshop*, 2015.

[Ghosh *et al.*, 2016] D. Ghosh, A. Khanam, Y. Han, and S. Muresan. Coarse-grained argumentation features for scoring persuasive essays. In *ACL*, 2016.

[Grosz *et al.*, 1995] B. J. Grosz, A. K. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 1995.

[Horbach *et al.*, 2017] A. Horbach, D. Scholten-Akoun, Y. Ding, and T. Zesch. Fine-grained essay scoring of a complex writing task for native speakers. In *BEA Workshop*, 2017.

[Howard and Ruder, 2018] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *ACL*, 2018.

[Hussein *et al.*, 2020] M. A. Hussein, H. A. Hassan, and M. Nassef. A trait-based deep learning automated essay scoring system with adaptive feedback. *IJACSA*, 2020.

[Jiang *et al.*, 2023] Z. Jiang, T. Gao, Y. Yin, M. Liu, H. Yu, Z. Cheng, and Q. Gu. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *ACL*, 2023.

[Jin *et al.*, 2018] C. Jin, B. He, K. Hui, and L. Sun. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *ACL*, 2018.

[Ke and Ng, 2019] Z. Ke and V. Ng. Automated essay scoring: A survey of the state of the art. In *IJCAI*, 2019.

[Ke *et al.*, 2018] Z. Ke, W. Carlile, N. Gurrapadi, and V. Ng. Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays. In *IJCAI*, 2018.

[Ke *et al.*, 2019] Z. Ke, H. Inamdar, H. Lin, and V. Ng. Give me more feedback II: Annotating thesis strength and related attributes in student essays. In *ACL*, 2019.

[Kumar *et al.*, 2022] R. Kumar, S. Mathias, S. Saha, and P. Bhattacharyya. Many hands make light work: Using essay traits to automatically score essays. In *NAACL*, 2022.

[Lee *et al.*, 2024] S. Lee, Y. Cai, D. Meng, Z. Wang, and Y. Wu. Prompting large language models for zero-shot essay scoring via multi-trait specialization, 2024.

[Liu *et al.*, 2019] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*, 2019.

[Mansour *et al.*, 2024] W. Mansour, S. Albatarni, S. Eltanbouly, and T. Elsayed. Can large language models automatically score proficiency of written essays?, 2024.

[Marinho *et al.*, 2021] J. C. Marinho, R. T. Anchiêta, and R. S. Moura. Essay-BR: a Brazilian corpus of essays. In *Dataset Showcase Workshop (DSW)*, 2021.

[Mathias and Bhattacharyya, 2018] S. Mathias and P. Bhattacharyya. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *LREC*, 2018.

[McNamara *et al.*, 2015] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59, 2015.

[Mizumoto and Eguchi, 2023] A. Mizumoto and M. Eguchi. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2023.

[Morris and Hirst, 1991] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 1991.

[Nguyen and Litman, 2018] H. V. Nguyen and D. J. Litman. Argument mining for improving the automated scoring of persuasive essays. In *AAAI*, 2018.

[Östling *et al.*, 2013] R. Östling, A. Smolentzov, B. Tyrefors Hinnerich, and E. Höglin. Automated essay scoring for Swedish. In *BEA Workshop*, 2013.

[Page, 1966] E. B. Page. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 1966.

[Persing and Ng, 2013] I. Persing and V. Ng. Modeling thesis clarity in student essays. In *ACL*, 2013.

[Persing and Ng, 2014] I. Persing and V. Ng. Modeling prompt adherence in student essays. In *ACL*, 2014.

[Persing and Ng, 2015] I. Persing and V. Ng. Modeling argument strength in student essays. In *ACL-IJCNLP*, 2015.

[Persing *et al.*, 2010] I. Persing, A. Davis, and V. Ng. Modeling organization in student essays. In *EMNLP*, 2010.

[Peters *et al.*, 2018] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.

[Phandi *et al.*, 2015] P. Phandi, K. M. A. Chai, and H. T. Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *EMNLP*, 2015.

[Radford *et al.*, 2018] A. Radford, K. Narasimhan, T. Salimansv, and I. Sutskever. Improving language understanding with unsupervised learning. Technical report, OpenAI, 2018.

[Ridley *et al.*, 2020] R. Ridley, L. He, X. Dai, S. Huang, and J. Chen. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *ArXiv*, 2020.

[Ridley *et al.*, 2021] R. Ridley, L. He, X.-Y. Dai, S. Huang, and J. Chen. Automated cross-prompt scoring of essay traits. In *AAAI*, 2021.

[Shermis and Burstein, 2003] M. D. Shermis and J. C. Burstein. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. 2003.

[Shermis and Burstein, 2013] M. D. Shermis and J. C. Burstein. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, 2013.

[Shermis *et al.*, 2010] M. D. Shermis, J. Burstein, D. Higgins, and K. Zechner. Automated essay scoring: Writing assessment and instruction. In *International Encyclopedia of Education*. Elsevier, 3rd edition, 2010.

[Somasundaran *et al.*, 2014] S. Somasundaran, J. Burstein, and M. Chodorow. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *COLING*, 2014.

[Stab and Gurevych, 2014] C. Stab and I. Gurevych. Annotating argument components and relations in persuasive essays. In *COLING*, 2014.

[Taghipour and Ng, 2016] K. Taghipour and H. T. Ng. A neural approach to automated essay scoring. In *EMNLP*, 2016.

[Uto *et al.*, 2020] M. Uto, Y. Xie, and M. Ueno. Neural automated essay scoring incorporating handcrafted features. In *COLING*, 2020.

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

[Vaswani *et al.*, 2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wachsmuth *et al.*, 2016] H. Wachsmuth, K. Al-Khatib, and B. Stein. Using argument mining to assess the argumentation quality of essays. In *COLING*, 2016.

[Wang *et al.*, 2022] Y. Wang, C. Wang, R. Li, and H. Lin. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *NAACL*, 2022.

[Xiao *et al.*, 2024] C. Xiao, W. Ma, S. X. Xu, K. Zhang, Y. Wang, and Q. Fu. From automation to augmentation: Large language models elevating essay scoring landscape, 2024.

[Xie *et al.*, 2022] J. Xie, K. Cai, L. Kong, J. Zhou, and W. Qu. Automated essay scoring via pairwise contrastive regression. In *COLING*, 2022.

[Yang *et al.*, 2020] R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of EMNLP*, 2020.

[Yannakoudakis and Briscoe, 2012] H. Yannakoudakis and T. Briscoe. Modeling coherence in ESOL learner texts. In *BEA Workshop*, 2012.

[Yannakoudakis *et al.*, 2011] H. Yannakoudakis, T. Briscoe, and B. Medlock. A new dataset and method for automatically grading ESOL texts. In *ACL*, 2011.

[Zesch *et al.*, 2015] T. Zesch, M. Wojatzki, and D. Scholten-Akoun. Task-independent features for automated essay grading. In *BEA Workshop*, 2015.