Beyond the Gold Standard in Analytic Automated Essay Scoring

Anonymous ACL submission

Abstract

Originally developed to reduce the manual burden of grading standardised language tests, Automated Essay Scoring (AES) research has long focused on holistic scoring methods which offer minimal formative feedback in the classroom. With the increasing demand for technological tools that support language acquisition, the field is turning to analytic AES (evaluating essays according to different linguistic traits). This approach holds promise for generating more detailed essay feedback, but relies on analytic scoring data that is both more cognitively demanding for humans to produce, and prone to bias. The dominant paradigm in AES is to aggregate disagreements between raters into a single gold-standard label, which fails to account for genuine examiner variability. In an attempt to make AES more representative and trustworthy, we propose to explore the sources of disagreements and lay out a novel AES system design that learns from individual raters instead of the gold standard labels.

1 Introduction

Writing practice is an essential part of learning a second-language (Graham et al., 2012; Monk, 2016). Unfortunately, assessing writing is long and tedious, and educators frequently display inconsistencies due to fatigue and biases (Uto and Ueno, 2018) which compromise the quality of their marking (Hussein et al., 2019). By providing consistent, accessible, and cheaper written assessment, **Automated Essay Scoring** (AES) has the potential to address this issue (Magliano and Graesser, 2012).[1]

In the past, AES research primarily focused on holistic scoring, i.e., summarising the quality of essays with a single score (Phillips, 2007). However, this approach fails to provide any kind of formative feedback in the classroom (Carlile et al., 2018).

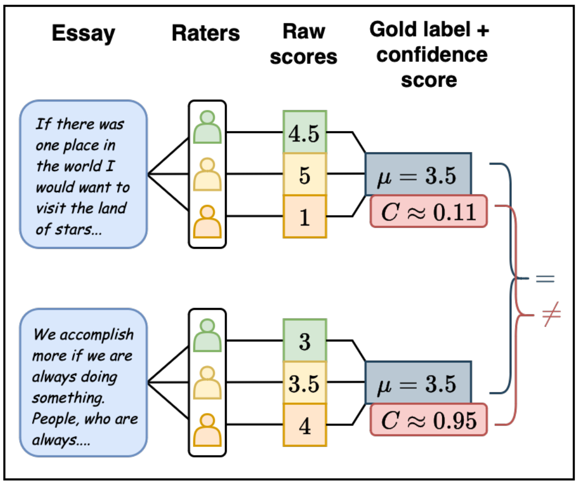

Figure 1: Two essays are multi-marked by three raters on a scale of 1–5. Their scores then are aggregated using an average, and we obtain the same mean μ. This is the gold label. We compute a confidence score C for each gold label using the variance of the raw scores (Section 4.2) and find that we can be much more confident in the second essay's gold label than the first's, despite being generally treated the same when training AES systems.

More recently, the field is turning to **analytic scoring** which involves automatically assessing essays along different dimensions to help students identify which aspects of their writing need improvement (Ke and Ng, 2019). Traits like coherence (Higgins et al., 2004), relevance to prompt (Louis and Higgins, 2010), and persuasiveness (Carlile et al., 2018) have already been studied. By breaking down essay quality into different traits, analytic AES can help a learner identify their strengths and weaknesses (e.g., Burstein et al., 2004).

However, though analytic scoring offers a pedagogically useful alternative, its implementation in real-world classrooms is not without challenges. The variety of writing tasks and ambiguity of scoring rubrics make it difficult for AES systems to consistently produce reliable scores (Xiao et al., 2025). Further, concerns over the fairness, account-

[1] We limit the discussion to the assessment of written text (or "essays") produced by **English as a Foreign Language/English as a Second Language** (EFL/ESL) students.

ability, and transparency of these systems are yet to be properly addressed (Madnani et al., 2017). These issues underscore the need for AES systems that support teacher-AI collaboration (Deane, 2013; Wilson and Roscoe, 2020) by not only producing accurate scores but also providing educators with confidence estimates, and explanations.

To design transparent systems, we must first examine the data on which AES systems are typically trained: corpora of human-marked essays. Essay scoring is a difficult and subjective task, prone to rater disagreements (Brown, 2010). This is especially true for analytic scoring which is more cognitively demanding and time-consuming than holistic scoring (Hunter et al., 1996), and particularly vulnerable to rater effects (Myford and Wolfe, 2003). Despite these limitations, the dominant paradigm in Machine Learning (ML) and AES has always been to reconcile rater disagreements under one ground truth label referred to as the *gold standard* via different aggregation methods (Abercrombie et al., 2024). Not only does this neglect genuine examiner variation, but it also erases precious information about the essays (as illustrated in Figure 1) which we could use to inform better analytic AES.

With the long-term goal of improving AES systems for teacher-in-the-loop applications (Colonna, 2024), we propose to draw on **perspectivist** literature (Section 2.3) which "aims at leveraging data annotated by different individuals in order to model varied perspectives that influence their opinions and world view" (Frenda et al., 2024). In doing so, we hope to align AES systems with the diversity of rater judgements, enhancing the way in which output confidence is measured.

This PhD thesis proposal is structured as follows: Section 2 situates rater disagreements in written assessment, advocating for a perspectivist approach to data annotation in AES. Section 3 introduces relevant analytic AES datasets and techniques. Section 4 outlines our phased research plan which includes a study of disagreements in essay scoring data, the development of multi-annotator AES models, and their application to feedback generation.

2 Background

We start by contextualising and introducing perspectivist literature as an alternative approach to using annotated data for model training, and make a case that AES, and particularly analytic AES research, can benefit from this paradigm shift.

2.1 Multi-marking

Modern NLP research is highly dependent on the existence of annotated corpora for the training and evaluation of models. Thanks in part to initiatives such as SemEval or Senseval (Sabou et al., 2014), and open-competitions such as those hosted by the Kaggle[2] platform, the number of publicly available datasets is growing. And with them, best practices on how to create annotations of consistently high quality have been developed. Over the years, the "science of annotation" (Hovy, 2010) has become the subject of many dedicated conferences and workshops such as HCOMP[3] or AnnoNLP (Paun and Hovy, 2019).

Amongst the many guidelines that have been set out, it is generally considered "axiomatic" that any annotation task should be performed by at least two or more raters acting independently. This allows us to compare their rating decisions and measure the extent to which they agree (or disagree) on the same instances of data (Hovy and Lavid, 2010). Traditional agreement measures includes Krippendorff's alpha (Krippendorff, 2004) or variations of Cohen's Kappa measure (Cohen, 1960). Reporting and acting on agreement measures generally improves the overall quality of the data being collected (Snow et al., 2008; Nowak and Rüger, 2010).

2.2 Disagreements

Full agreement is rarely possible, especially for complex or subjective tasks (Hovy and Lavid, 2010), such as essay scoring, where a single "right" answer may not exist (Ovesdotter Alm, 2011). This is because having two distinct readers arrive at an identical judgement for the same piece of writing not always possible (Huot, 1990a), and there is no objective way of validating either's rating (Sadler, 2009). In fact, there is no single written evaluation standard that can be said to embody *the* ideal written product of English (Kroll, 1990). In most cases, disagreements are initially treated as a consequence of low annotation quality, and addressed through various strategies to minimise noisy data, such as annotator training (Hovy et al., 2006; Carlson et al., 2003) or reconciliation (Hovy and Lavid, 2010). Any remaining disagreements are then reduced to a single gold label by averaging (Sabou et al., 2014), majority vote (Leonardelli et al., 2021) or adjudication by an expert (Waseem and Hovy, 2016).

[2]See https://www.kaggle.com.
[3]See https://www.humancomputation.com.

Unfortunately, these approaches reduce labels to the opinion of just one individual, precisely where annotation exposes complexity (Hovy and Lavid, 2010). For instance, Plank et al. (2014b) show that disagreements in part-of-speech (POS) annotation can be systematic across domains and languages, and due to "linguistically debatable" or hard cases rather than annotation errors (e.g., possessive pronouns may be classified as determiners or pronouns). In essay scoring, raters have to reconcile their impression of the text, its particular features, and the relevant scoring rubric. Given the boundless nature of language, the latter can never be exhaustive, and markers must cope with the underspecification of rating (Lumley, 2002). Further, raters may be influenced by their cultural, political, and socio-economic background (Guerra et al., 2011; Amorim et al., 2018). And if something as prescriptive and well-documented as POS-tagging leaves room for interpretation as illustrated in Plank et al. (2014a), then the high-level descriptors typically present in essay scoring rubrics will definitely introduce ambiguity, and with it, debatable cases.

2.3 Perspectivism

At a time when AI systems are increasingly scrutinised over bias and fairness concerns, it is not enough to assume a single "ground truth" as this can erase legitimate disagreements. Perspectivism challenges this assumption by pursuing approaches that understand and account for genuine human variability (Abercrombie et al., 2024).

A few studies have explored ways in which to use disagreements during model training. For instance, Prabhakaran et al. (2012) and Plank et al. (2014a) have tried to incorporate rater disagreements into the training loss functions: by penalising errors made on highly agreed data points more than those incurred from mislabelling complex instances (that is, with higher disagreement). Others have looked at actually modelling disagreement. Akhtar et al. (2021) divided annotators into two groups based on their polarisation (on a hate-speech classification task), and for each, compiled a different gold standard dataset to train individual classifiers. Combining these using an ensemble modelling approach outperformed previous state-of-the-art supervised classifiers for that task. More recently, Mostafazadeh Davani et al. (2022) compared three training strategies including ensembling, multi-label classification (Tsoumakas and Katakis, 2009) and multi-task learning (MTL; Caruana, 1993) on two tasks: hate-speech and emotion classification. Their results demonstrated that a MTL approach performs better than a baseline trained on aggregated gold standard labels. Additionally, these architectures provide a way to estimate uncertainty in predictions by preserving different annotators' perspectives until the prediction step. See Frenda et al. (2024) for a full survey of perspectivist approaches. We note that, to the best of our knowledge, perspectivism has not yet been investigated in the context of AES research.

In the next section, we show how (analytic) AES research exemplifies the challenges and opportunities of handling subjectivity in annotation.

2.4 Analytic Scoring

At first, AES research primarily focused on summarising the quality of essays with a single score (e.g., the Intelligent Essay Assessor™; Landauer et al., 2003) in response to the needs of large-scale standardised tests such as TOEFL, IELTS and GMAT (Chodorow and Burstein, 2004; Chen et al., 2016). But where holistic approaches fall short in terms of providing formative feedback to students in the classroom (Carlile et al., 2018), analytic scoring shows promise (Higgins et al., 2004; Louis and Higgins, 2010; Somasundaran et al., 2014; Persing and Ng, 2014; Kaneko et al., 2020).

Contrary to coarse holistic evaluations, analytic criteria consider a wide range of linguistic dimensions (or *traits*) involved in the composition of an essay (e.g., coherence, syntax, relevance to prompt, etc.) to better highlight the strengths and weaknesses of a student's writing (Carlile et al., 2018). Analytic scoring ensures that raters award appropriate scores while also revealing the grounds for their decisions to students by pointing out specific writing strengths and weaknesses (Reid, 1993, p.235). In doing so, they have the potential to reduce the apparent arbitrariness of grading (Lumley, 2002) and can easily be used as the basis for fine-grained feedback (Carlile et al., 2018; Bannò et al., 2024).

Unfortunately, due to the fuzzy nature of language (Douglas, 1997), analytic scales are more cognitively demanding to use (Cai, 2015). They also run the risk of being psychometrically redundant (Lee et al., 2010) due to rater effects (Engelhard, 1994). Moreover, the very idea that text features are independent constructs whose sum is a valid representation of the overall quality of a text is subject of debate (Huot, 1990b).

Given the complex and subjective nature of analytic essay scoring data, greater even than that of holistic scoring, we should not be blindly training models on the gold standard, and posit that analytic AES could benefit from a perspectivist approach.

3 Related Work

In this section, we review prior work in AES, with a special focus on analytic AES, introducing the datasets and main techniques relevant to our study.

3.1 Datasets

As was noted by Ke and Ng (2019), progress in analytic AES is hindered in part by the lack of large annotated corpora needed for model training. To the best of our knowledge, only ICLE++ (Granger, 2003; Granger et al., 2009, 2020; Li and Ng, 2024), ASAP++ (Mathias and Bhattacharyya, 2018), ICNALE GRA (Ishikawa, 2020, 2023), CELA (Xue et al., 2021), and ELLIPSE (Crossley et al., 2024) have been publicly released for the English language. Of those, all but CELA have released the original, raw multi-marks, alongside the aggregated gold standard scores. See Appendix A for more information about these dataset. Importantly, Table 1 compares these datasets along various dimensions include size, and analytic traits assessed.

Put together, these datasets include scores for 34 distinct analytic trait names, ranging from low-level dimensions like "grammar" or "syntax", lexical dimensions like "word choice" or "vocabulary", to complex, discourse-level dimensions like "coherence" or "thesis clarity". Further, while some of these datasets share common trait names (e.g., "organisation"), it is important to keep in mind that each comes with very different scoring rubrics, and that the definitions of these dimensions might in fact be radically different. While this diversity can be seen as valuable, it is also an additional challenge for analytic AES research. Indeed, we cannot make any link between datasets before having properly studied how the essays were annotated. The same should be said for parallels made across studies which work with different sources of essay data.

Unfortunately, while there have been some efforts to rationalise this: notably, Li and Ng (2024, Table 2) offer a mapping between some of ICLE++'s traits and those of the ASAP++ dataset; we identify a clear gap in the field's general understanding of its analytic essay scoring datasets.

3.2 Machine Learning Approaches

Up until recently, the field of (analytic) AES mainly focused on developing effective handcrafted feature-based models (Craighead et al., 2020). Common features included grammatical errors (Andersen et al., 2013), distinctive words or part-of-speech n-grams (Page and Paulus, 1968) and essay length (Lee et al., 2008).

With the recent surge of interest in neural networks, transformer-based systems have gained favour (Ke and Ng, 2019): see Zhang and Litman (2018); Ke et al. (2019); Mayfield and Black (2020); Xue et al. (2021); Shibata and Uto (2022); Ajit Tambe and Kulkarni (2022); Dadi and Sanampudi (2023); Doi et al. (2024); Cho et al. (2024); Ding et al. (2024). These models perform on par with feature-based systems, and eliminate the need for expensive feature engineering (Qiu et al., 2020). However, this gain comes at the cost of needing increasingly large quantities of annotated data for training (Zhang et al., 2021) which can be a problem for analytic AES which lacks large datasets (Section 3.1). Additionally, neural networks are very sensitive (Uto, 2021): the models can inherit biases present in data they are trained on which can result in systematic errors and a drop in performance (Amorim et al., 2018; Huang et al., 2019; Li et al., 2020). Finally, the inherent lack of interpretability of these "black box-like models" (Kumar and Boulanger, 2020) raises ethical concerns impacting safety (Danks and London, 2017), trust (Ribeiro et al., 2016), accountability (Kroll et al., 2016), and industrial liability (Kingston, 2018).

The most recent breakthrough, brought upon by LLMs such as the GPT models (Brown et al., 2020; OpenAI et al., 2024). Thanks to their impressive performance and ease-of-use, these models are being applied to an ever-growing range of tasks, including analytic AES. So far Bannò et al. (2024), Naismith et al. (2023), Yamashita (2024) and Seßler et al. (2025) have obtained promising results with GPT-4 (OpenAI et al., 2024) for analytic AES. LLMs are now widely used as evaluators to approximate human judgements, which are otherwise very expensive to obtain (Gu et al., 2024). The "LLM-as-a-Judge" paradigm (Zheng et al., 2023) has enormous potential for AES where data is so scarce. For instance, Xiao et al. (2025) found that LLM-generated feedback and confidence scores could be used to enhance the efficiency and robustness of human graders during rating. The capability of LLMs to generate natural language explanations opens up a lot of possibilities for the field of explainability (Zhao et al., 2024). At the same time, these capabilities raise new challenges, such as hallucinated explanations (incorrect or baseless), along with their inherent opaqueness (Singh et al., 2024), and output variability (Xia et al., 2024).

Finally, the multi-task learning (MTL) paradigm seems to be getting a lot of attention in AES. This approach "improves learning for one task by using the information contained in the training signals of other related tasks" (Caruana, 1997, Chapter 1). It first appears in the work of Ridley et al. (2021) whose Cross-prompt Trait Scorer (CTS) is frequently used as a baseline on the ASAP++ corpus which builds on top of the Prompt Agnostic Essay Scorer (PAES; Ridley et al., 2020). Since then, all sorts of MTL analytic AES systems have been developed. See Xue et al. (2021) fine-tuned BERT on the multi-dimensional ASAP++ dataset using a shared BERT layer and trait-specific heads. Kumar et al. (2022) proposed a system whose primary task is holistic scoring, but leveraged information from analytic sub-scale scores to improve its overall performance using MTL. See also the works of Ramesh and Sanampudi (2022); Lee et al. (2023); Chen and Li (2023); Doi et al. (2024); Cho et al. (2024); Ding et al. (2024).

We note that MTL is also one of the architectures we plan to explore (Section 4.2), though to the best of our knowledge, it has never been applied to raw essay scores. In fact, not one of the studies mentioned above used raw analytic scores in lieu of the aggregated gold standard scores. This reflects a missed opportunity: treating rater disagreement as "noise" rather than signal fails to capture the full richness and variability of human judgement, which is precisely the kind of information that could enhance the transparency and reliability of AES systems in real-world settings. Thus, to the best of our knowledge, this area is yet unexplored.

4 Research Plan

We frame the following two research questions:

RQ1: How can examiner disagreements in analytic essay scoring data be used to measure and enhance confidence and performance in AES systems?

RQ2: How can analytic AES serve as a foundation for more effective automated essay feedback systems?

Through these, we hope to explore how we can best harness rater disagreements in analytic essay scoring data to improve the performance and confidence in AES and feedback systems.

4.1 Preliminary Work

As mentioned in Section 3.1, there is a lack of research into raw analytic essay scoring data. Yet most, if not all, current AES systems are trained on gold standard labels which are but a product of raw scores (Mostafazadeh Davani et al., 2022). We first seek to address this gap. Doing so will not only inform the research questions presented above, but also provide broader value to the field of AES by enhancing the interpretability of widely used datasets and enabling more meaningful comparisons across existing and future studies.

Dataset mapping. We have identified four analytic scoring datasets which have made their raw multi-marks available to us: namely ICNALE GRA, ELLIPSE, ICLE++, and parts of the ASAP++ corpus. These differ in terms of the types of essays they contain (e.g., argumentative or creative), score ranges (e.g., 1–5 or 0–10), number of raters per essay (e.g., 2 or 80), prompts, and, of course, traits assessed (Appendix A). Our first step will be to map the traits of these different datasets together, where possible. For example, comparing how "organisation" is defined in the rubrics of ICLE++ and ASAP++, and how it differs from "cohesion" which is perhaps more broadly defined in ELLIPSE. Obviously, we will have to take into account the types of essays as well. So far, Li and Ng (2024, Table 2) have mapped some of ICLE++'s traits to those of the ASAP++ dataset, for argumentative essays only, which is a small subset of the ASAP++ dataset. It is not our aim to oversimplify the problem or forcibly merge these datasets, but rather to offer a clearer understanding of how the different rubrics and annotations align or diverge. By doing so, we hope to improve the reusability of these datasets, laying the groundwork for more consistent cross-dataset comparisons in the field.

Qualitative analysis. Having done so, we will be better positioned to conduct a cross-dataset analysis of rater behaviour and scoring patterns, and will next seek to answer the following questions:

P1: What are the common patterns between the essays that have high examiner disagreement, both within and across analytic traits?

P2: Conversely, for essays that have high agreement, what are the particular features that make an essay stereotypically good or bad?

To answers these questions, we will be performing an in-depth content analysis (Mayring, 2014) of the four previously mentioned datasets. The goal of this phase is to systematically code and categorise patterns of rater agreement and disagreement across traits. Coding will begin deductively using a set of pre-defined categories informed by the rubrics of the datasets themselves (e.g., organisation, grammar, relevance to prompt) and prior studies on rater effects (e.g., halo, severity/leniency; Myford and Wolfe, 2003). Inductive coding will follow, allowing new categories to emerge from the data where rating patterns deviate from rubric norms or where disagreements appear to cluster. These codes will be applied at both the trait level (e.g., is there consistent divergence in "cohesion" scores?) and the essay level (e.g., do specific essays elicit unusually wide score variance across traits?).

We will follow this with a thematic analysis (Braun and Clarke, 2021) on a carefully curated subset of essays selected based on results from the content analysis. Specifically, we will include:

- Essays exhibiting extreme marker disagreement (e.g., with scores ranging the full scale);
- Essays that display high cross-trait disagreement (e.g., rated very highly in grammar but poorly in coherence by the same rater); and
- Essays that exemplify strong consensus, serving as contrast cases for identifying stereotypically *good* or *bad* writing.

Selection will aim for balance across datasets, genres, and prompts. These essays will be analysed in-depth to explore possible linguistic, structural, or stylistic features that may account for disagreement or consensus. Themes may include ambiguity in argument structure, unconventional grammar use, cultural variation in rhetorical style, or misalignment with rubric expectations.

Both content and thematic analyses will be completed on ATLAS.ti, a robust and well-established qualitative data analysis software (Paulus, 2023), which will support efficient coding, memoing, and cross-case comparison.

Research questions **P1** and **P2** are conceptually linked: by examining essays that provoke high disagreement (**P1**), we gain insight into the limitations or ambiguities of existing rubrics and linguistic features that challenge human raters. Conversely, analysing essays with high agreement (**P2**) helps surface the features raters appear to consistently associate with quality or poor writing.

4.2 Towards RQ1

Using the insights of the preliminary phase, we propose a new AES system that learns from individual raters instead of the gold standard labels.

Dataset. Despite our previous efforts to map the dataset traits together (**Dataset mapping**), we do not wish nor expect to merge the datasets as one. Doing so would require too many assumptions and restrict comparison with prior work. As we turn to training and evaluating a new analytic AES system, we must thus choose a dataset. Out of the four previously considered, ASAP++ is by far the largest with 12,980 essays, and has also been widely used in holistic AES research (Section A.2). Unfortunately, it is not very usable: not all essays have been multi-marked, and both the traits assessed and score ranges vary depending on the essay prompts. Instead, we will use the second largest dataset, the ELLIPSE corpus, with 6,482 essays. All of its essays have been marked by two or three raters using on a 1–5 scale using the same analytic rubric (Section A.4). Further, since this dataset was released as part of a Kaggle competition[4], the dataset comes with an established test-train split (3,911 essays in the training set and 2,571 essays in the test set). For lack of an existing set, we will use 10% of the training set for validation, aiming for balance across prompts, scores and demographics.

Baseline. As baseline, we propose to use the pretrained DeBERTa model (He et al., 2021), a state-of-the-art neural language model, which has been used in past AES research with success: for example (Hicke et al., 2023; Wang, 2024; Zhong, 2024; Huang et al., 2024). Appendix B presents how we selected this particular model. Specifically, we will fine-tune six individual DeBERTa models (one for each of the traits assessed in the ELLIPSE corpus) for regression on the gold standard labels only. Appendix C describes in detail the methodology we plan to use for these experiments.[5]

[4] See https://www.kaggle.com/competitions/feedback-prize-english-language-learning/data.

[5] All experiments presented in this proposal have been and will be conducted using shared high-performance computing resources which include three NVIDIA A100 GPUs.

Modelling. Drawing from the work by Mostafazadeh Davani et al. (2022), and for each of the six analytic traits in ELLIPSE, we will consider three different multi-annotator AES architectures which can mimic the multi-marking setting: namely ensemble, multi-label, and multi-task. We point out that some of these architectures have already been used in analytic AES in the past with success (Section 3). However, importantly, unlike prior work and our baseline, we will be training them on the raw, multi-marked essay scoring data as opposed to the gold standard labels. See Figure 2 for a schematic overview of this experimental design. Note that all variations will rest on the pre-trained language model DeBERTa.

Performance. We will then compare, for each trait, the three architectures to the baseline using the evaluation metrics defined in Appendix C.3. Specifically, we will be measuring model performance based on the RMSE metric (Tyagi et al., 2022) only. Not only is it a well understood and widely used metric in ML (Karunasingha, 2022), Yannakoudakis and Cummins (2015) argues that measures of agreement (such as RMSE) are more appropriate than correlation metrics for measuring the effectiveness of AES systems. Beyond our baseline, we will also compare the performance of our systems against the leader-board of the dataset's Kaggle competition[4], and the few studies that have used ELLIPSE (e.g., Sun and Wang, 2024).

Confidence. The main novelty these models bring to AES is that we will be able to use their raw outputs to estimate how confident we should be in using an aggregate of the outputs together. Indeed, suppose we approximate each model head, or individual raw output as being a single rater's judgement. If all the outputs of our model agree, then much like when human raters agree, we should be highly confident that aggregating the raw scores together accurately conveys the quality of the essay for the considered analytic trait. If, however, the model outputs disagree, then perhaps aggregating the scores is not the best course of action.

Mostafazadeh Davani et al. (2022) propose to use the variance between the different raw model outputs as a measure of uncertainty. We describe below how to convert that into a confidence score C, with a value between 0 and 1 (as was used in Figure 1). Given that the maximal variance between three values in the 1–5 score range of ELLIPSE is $\sigma^2_{\max} \approx 3.6$ (rounded to 1 decimal place), achieved for outputs $(1, 5, 5)$ or $(1, 1, 5)$, in no particular order. Then, given any set of three raw model outputs represented as a three dimensional vector $\mathbf{x} \in [1, 5]^3$, the confidence score associated to that prediction is given by:

$$C(\mathbf{x}) = \frac{\sigma^2_{\max} - \sigma^2(\mathbf{x})}{\sigma^2_{\max}}.$$

There are of course many alternatives to this metric, and this metric will need to be properly validated. We will do so by looking at how predicted confidence correlates with the true rater disagreement, using the original raw rater scores, on the test set. We can further assess the reliability of the metric by we segmenting the test samples based on the predicted confidence scores and measure the correlation between these scores and model performance as was done by (Xiao et al., 2025). We will also explore other confidence/uncertainty metrics such as using the prediction probability from a Softmax distribution of the final output (Hendrycks and Gimpel, 2018) or Monte Carlo dropouts (Gal and Ghahramani, 2016).

4.3 Towards RQ2

Having built a series of multi-annotator AES systems for a range of essay traits, we turn our attention to the area of essay feedback: How can analytic AES serve as a foundation for more effective automated essay feedback systems?

We envision that the raw model outputs across multiple traits can form a kind of feedback *profile* for each essay, which may be mapped to specific linguistic features. Insights from our preliminary analysis (**P1** and **P2**) may help identify textual characteristics that consistently trigger high or low rater disagreement. Simply highlighting these features to learners may already provide useful formative feedback, but they could also augment existing feedback systems by offering more nuanced, trait-specific insights. Specifically, we can explore how LLMs can be used to translate raw trait scores and disagreement-informed insights into natural language explanations. These explanations could help bridge the gap between system output and learner interpretation, supporting feedback that is not only data-driven but also accessible and pedagogically meaningful. However, careful prompting and validation would be needed to ensure reliability and mitigate risks such as hallucinated feedback or overgeneralisation (Singh et al., 2024; Zhao et al., 2024).

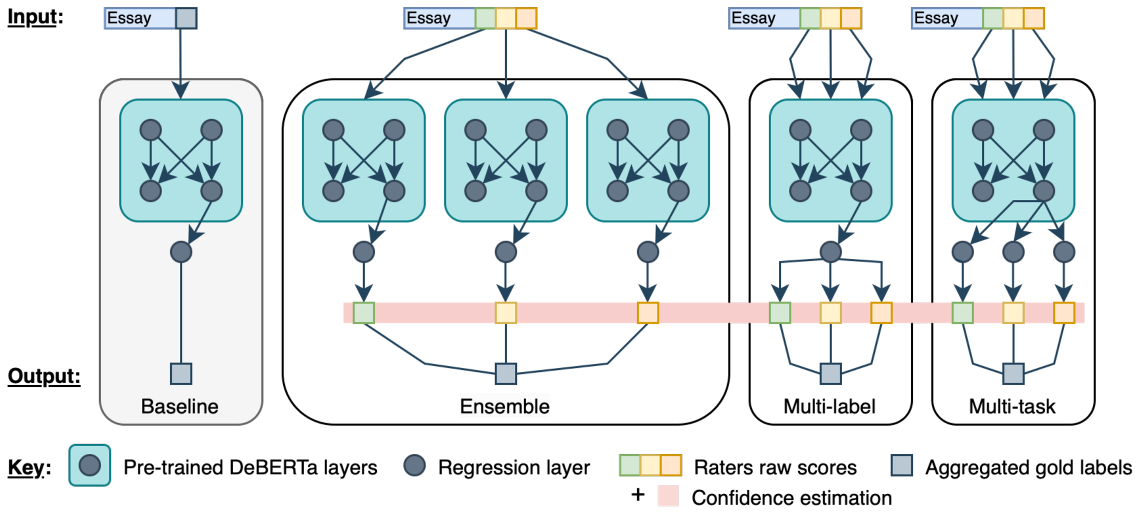

Figure 2: Schematic overview of the multi-annotator AES models (ensemble, multi-label, and MTL) and baseline we plan to build for each analytic trait in ELLIPSE. Adapted from Mostafazadeh Davani et al. (2022, Figure 1).

Evaluating the effectiveness of this kind of approach to feedback will ideally require engagement with actual users: teachers and students. To that end, we will design a small-scale, controlled user study, time and resources permitting. In particular, we may draw from Wilson and Roscoe (2020) which measured the effectiveness of their approach through a series of metrics: writing self-efficacy, holistic writing quality, and performance on a state English language arts test, and teachers' perceptions of the AES system's social validity. Particular attention would be given to how disagreement-informed feedback compares with more conventional, rule-based or gold-standard approaches.

We consider this a longer-term, exploratory extension of our project, recognising that user-facing feedback is a complex and iterative design challenge. If direct user testing is not feasible within the current project scope, we will instead rely on proxy evaluations—such as alignment with rubric criteria, interpretability assessments, or expert annotation studies—to ensure pedagogical relevance and practical utility. Ultimately, our goal to contribute to a learner-centred vision of AES that supports teaching and learning in meaningful ways.

5 Summary

In this PhD proposal, we explored the idea that we can advance analytic AES research by harnessing examiner disagreements, rather than viewing them as "noise" that should be quietened. We propose to build a series of multi-annotator models to mimic a multi-marker setting and output automated raw scores. By placing the original raters of the training data at the centre of our design, our solution will not only help measure how confident we can be in the model's aggregated output, but also prove more transparent than traditional approaches. And by focusing on analytic scoring, we will be able to use our suite of models to generate fine-grained feedback, offering more tailored and effective guidance to learners. A key part of this work will require conducting a systematic qualitative analysis of rater disagreement in analytic essay scoring data. By improving interpretability, surfacing uncertainty, and enabling richer feedback, we hope to contribute to the development of AES systems that are designed for real-world classroom use.

Overall, we believe the project is feasible within the time-frame of a PhD. The phased research plan outlines the work will look to complete over the next 18 months. The availability of public multi-marked analytic AES datasets makes this work both timely and well-grounded.

6 Limitations

The primary limitation of this study is the lack of large, publicly-available multi-marked analytic AES datasets. While our approach seeks to better model rater variability and improve representation in AES systems, most of the datasets we draw from have been annotated by no more than two or three raters per essay (see Appendix A). This relatively shallow annotation depth may limit the extent to which we can robustly capture and model inter-rater variation, particularly for traits that are inherently more subjective or rubric-dependent. Importantly, we note that this is not a limitation unique to this study, but a broader challenge across AES.

A related constraint concerns language coverage. All of the datasets used in this study are in English, which was also our particular focus. [1] However, this limits the immediate applicability of our findings to English-language educational contexts. Future work could extend this approach to other languages as suitable multi-marked datasets become available. Such extensions would be essential for ensuring that AES advancements benefit a more diverse set of learners and writing contexts.

Finally, although our use of qualitative methods (content and thematic analysis) enriches the interpretability of findings, these approaches carry inherent subjectivity. Researcher bias in coding and theme development is a known limitation of qualitative work. To mitigate this, we will use a transparent and iterative coding process, triangulate findings where possible, and document decisions clearly through ATLAS.ti.

7 Ethical Considerations

Fairness is a core ethical concern in educational assessment, particularly when deploying automated systems that may influence learner outcomes. AES models risk amplifying existing biases in training data, especially if rater disagreement, socio-cultural variation, or language proficiency differences are not adequately accounted for. Our work aims to address this by modelling rater disagreement directly, promoting transparency and interpretability, and supporting more equitable scoring practices in diverse educational contexts.

References

Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors. 2024. *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.

Aniket Ajit Tambe and Manasi Kulkarni. 2022. Automated Essay Scoring System with Grammar Score Analysis. In *2022 Smart Technologies, Communication and Robotics (STCR)*, pages 1–7.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection. *arXiv preprint*. ArXiv:2106.15896.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic Text Scoring Using Neural Networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated Essay Scoring in the Presence of Biased Ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, New Orleans, Louisiana. Association for Computational Linguistics.

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41, Atlanta, Georgia. Association for Computational Linguistics.

Stefano Bannò, Hari Krishna Vydana, Kate M. Knill, and Mark J. F. Gales. 2024. Can GPT-4 do L2 analytic assessment? *arXiv preprint*.

Virginia Braun and Victoria Clarke. 2021. *Thematic Analysis: A Practical Guide*. SAGE. Google-Books-ID: mToqEAAAQBAJ.

Gavin Brown. 2010. The Validity of Examination Essays in Higher Education: Issues and Responses. *Higher Education Quarterly*, 64:276–291.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: the criterion online writing service. *AI Magazine*, 25(3):27–36.

Hongwen Cai. 2015. Weight-Based Classification of Raters and Rater Cognition in an EFL Speaking Test. *Language Assessment Quarterly*, 12(3):262–282. Publisher: Routledge _eprint: https://doi.org/10.1080/15434303.2015.1053134.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give Me More Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. In *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 85–112. Springer Netherlands, Dordrecht.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28.

Richard A. Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. pages 41–48. Elsevier.

Jing Chen, James Fife, Isaac Bejar, and André Rupp. 2016. Building e-rater® Scoring Models Using Machine Learning Methods. *ETS Research Report Series*, 2016.

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping Contrastive Learning for Cross-prompt Automated Essay Scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

Minsoo Cho, Jin-Xia Huang, and Oh-Woog Kwon. 2024. Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading. *ETRI Journal*, 46(1):82–95.

Martin Chodorow and Jill Burstein. 2004. Beyond Essay Length: Evaluating e-raters®'s Performance on TOEFL® Essays. *ETS Research Report Series*, 2004(1).

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Liane Colonna. 2024. Teachers in the loop? An analysis of automatic assessment systems under Article 22 GDPR. *International Data Privacy Law*, 14(1):3–18.

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2258–2269, Online. Association for Computational Linguistics.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2024. The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *International Journal of Learner Corpus Research*. Status: forthcoming.

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.

Ramesh Dadi and Suresh Sanampudi. 2023. A Multitask Learning System for Trait-based Automated Short Answer Scoring. *International Journal of Advanced Computer Science and Applications*, 14.

David Danks and Alex John London. 2017. Regulating Autonomous Systems: Beyond Standards. *IEEE Intelligent Systems*, 32(1):88–91. Conference Name: IEEE Intelligent Systems.

Paul Deane. 2013. On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1):7–24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805.

Yuning Ding, Omid Kashefi, Swapna Somasundaran, and Andrea Horbach. 2024. When Argumentation Meets Cohesion: Enhancing Automatic Feedback in Student Writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17513–17524, Torino, Italia. ELRA and ICCL.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Automated Essay Scoring Using Grammatical Variety and Errors with Multi-Task Learning and Item Response Theory. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 316–329, Mexico City, Mexico. Association for Computational Linguistics.

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.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

David L. Donoho, Arian Maleki, Inam Ur Rahman, Morteza Shahram, and Victoria Stodden. 2009. Reproducible Research in Computational Harmonic Analysis. *Computing in Science & Engineering*, 11(1):8–18. Conference Name: Computing in Science & Engineering.

Dan Douglas. 1997. *Theoretical underpinnings of the Test of Spoken English revision project*. TOEFL monograph series ; MS-9. Educational Testing Service, Princeton, N.J.

George Engelhard. 1994. Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, 31(2):93–112. Publisher: [National Council on Measurement in Education, Wiley].

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv preprint*. ArXiv:1506.02142 [stat].

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Steve Graham, Debra McKeown, Sharlene Kiuhara, and Karen R. Harris. 2012. A Meta-Analysis of Writing Instruction for Students in the Elementary Grades. *JOURNAL OF EDUCATIONAL PSYCHOLOGY*, 104(4):879–896. Num Pages: 18 Place: Washington Publisher: Amer Psychological Assoc Web of Science ID: WOS:000310861600001.

Sylviane Granger. 2003. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly*, 37(3):538–546. Publisher: [Wiley, Teachers of English to Speakers of Other Languages, Inc. (TESOL)].

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English. Version 2. Handbook and CD-ROM*.

Sylviane Granger, Maïté Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *International Corpus of Learner English. Version 3*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. *A Survey on LLM-as-a-Judge*.

Pedro Guerra, Adriano Veloso, Wagner Meira Jr, and Virgilio Almeida. 2011. *From bias to opinion: A transfer-learning approach to real-time sentiment analysis*. Pages: 158.

Majdi H. Beseiso. 2021. Essay Scoring Tool by Employing RoBERTa Architecture. In *International Conference on Data Science, E-learning and Information Systems 2021*, DATA'21, pages 54–57, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint*. ArXiv:2006.03654.

Dan Hendrycks and Kevin Gimpel. 2018. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv preprint*. ArXiv:1610.02136 [cs].

Yann Hicke, Tonghua Tian, Karan Jha, and Choong Hee Kim. 2023. Automated Essay Scoring in Argumentative Writing: DeBERTeachingAssistant. *arXiv preprint*. ArXiv:2307.04276 [cs].

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating Multiple Aspects of Coherence in Student Essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 185–192, Boston, Massachusetts, USA. Association for Computational Linguistics.

Mohammad Hossin and Sulaiman M.N. 2015. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5:01–11.

Eduard Hovy. 2010. Annotation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, page 4, Uppsala, Sweden. Association for Computational Linguistics.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22:13–36.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Jiaxin Huang, Xinyu Zhao, Chang Che, Qunwei Lin, and Bo Liu. 2024. Enhancing Essay Scoring with Adversarial Weights Perturbation and Metric-specific AttentionPooling. *arXiv preprint*. ArXiv:2401.05433 [cs].

Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao. 2019. O2U-Net: A Simple Noisy Label Detection Approach for Deep Neural Networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3325–3333. Conference Name: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) ISBN: 9781728148038 Place: Seoul, Korea (South) Publisher: IEEE.

Darryl M. Hunter, Richard M. Jones, and Bikkar S. Randhawa. 1996. The Use of Holistic versus Analytic Scoring for Large-Scale Assessment of Writing.

Canadian Journal of Program Evaluation, 11(2):61–86.

Brian Huot. 1990a. The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research*, 60(2):237–263. Publisher: [Sage Publications, Inc., American Educational Research Association].

Brian Huot. 1990b. Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know. *College Composition and Communication*, 41(2):201–213. Publisher: National Council of Teachers of English.

Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. Automated language essay scoring systems: a literature review. *PeerJ. Computer Science*, 5:e208.

Shin'ichiro Ishikawa. 2020. *Aim of the ICNALE GRA Project: Global Collaboration to Collect Ratings of Asian Learners' L2 English Essays and Speeches from an ELF Perspective.*

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge, London.

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.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-Decoder Models Can Benefit from Pre-trained Masked Language Models in Grammatical Error Correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online. Association for Computational Linguistics.

Dulakshi Santhusitha Kumari Karunasingha. 2022. Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, 585:609–629.

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Student Essays. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3994–4004, Florence, Italy. Association for Computational Linguistics.

Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art. pages 6300–6308.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv preprint*. ArXiv:1412.6980.

John Kingston. 2018. Artificial Intelligence and Legal Liability. *arXiv preprint*. ArXiv:1802.07782.

Klaus Krippendorff. 2004. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research*, 30(3):411–433.

Barbara Kroll, editor. 1990. *Second Language Writing (Cambridge Applied Linguistics): Research Insights for the Classroom*. Cambridge Applied Linguistics. Cambridge University Press, Cambridge.

Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. Accountable Algorithms.

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.

Vivekanandan Kumar and David Boulanger. 2020. Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. *Frontiers in Education*, 5. Publisher: Frontiers.

Thomas Landauer, Darrell Laham, and Peter Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. *Automated essay scoring: A cross-disciplinary perspective*, pages 87–112.

Yejin Lee, Seokwon Jeong, Hongjin Kim, Tae-il Kim, Sung-Won Choi, and Harksoo Kim. 2023. NC2T: Novel Curriculum Learning Approaches for Cross-Prompt Trait Scoring. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 2204–2208, New York, NY, USA. Association for Computing Machinery.

Yong-Won Lee, Claudia Gentile, and Robert Kantor. 2008. Analytic Scoring of Toefl® Cbt Essays: Scores from Humans and E-Rater®. *ETS Research Report Series*, 2008(1):i–71.

Yong-Won Lee, Claudia Gentile, and Robert Kantor. 2010. Toward Automated Multi-trait Scoring of Essays: Investigating Links among Holistic, Analytic, and Text Feature Scores. *Applied Linguistics*, 31(3):391–417.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. 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.

Shikun Li, Shiming Ge, Yingying Hua, Chunhui Zhang, Hao Wen, Tengfei Liu, and Weiqiang Wang. 2020. Coupled-View Deep Classifier Learning from Multiple Noisy Annotators. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:4667–4674.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv preprint*. ArXiv:1711.05101.

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.

Tom Lumley. 2002. Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, 19(3):246–276. Publisher: SAGE Publications Ltd.

Pranava Madhyastha and Rishabh Jain. 2019. On Model Stability as a Function of Random Seed. *arXiv preprint*. ArXiv:1909.10447.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building Better Open-Source Tools to Support Fairness in Automated Scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain. Association for Computational Linguistics.

Joseph P. Magliano and Arthur C. Graesser. 2012. Computer-based assessment of student-constructed responses. *Behavior Research Methods*, 44(3):608–621.

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).

Elijah Mayfield and Alan W Black. 2020. Should You Fine-Tune BERT for Automated Essay Scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162. Association for Computational Linguistics.

Philipp Mayring. 2014. *Qualitative content analysis - theoretical foundation, basic procedures and software solution.*

Jonathan Monk. 2016. Revealing the iceberg: Creative writing, process & deliberate practice. *English in Education*, 50(2):95–115. Publisher: Routledge _eprint: https://doi.org/10.1111/eie.12091.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110. Place: Cambridge, MA Publisher: MIT Press.

Carol Myford and Edward Wolfe. 2003. Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of applied measurement*, 4:386–422.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, New York, NY, USA. Association for Computing Machinery.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774.

Cecilia Ovesdotter Alm. 2011. Subjective Natural Language Problems: Motivations, Applications, Characterizations, and Implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Ellis B. Page and Dieter H. Paulus. 1968. The Analysis of Essays by Computer. Final Report. Technical report. ERIC Number: ED028633.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv preprint*. ArXiv:1912.01703.

Trena M. Paulus. 2023. Using Qualitative Data Analysis Software to Support Digital Research Workflows. *Human Resource Development Review*, 22(1):139–148. Publisher: SAGE Publications.

Silviu Paun and Dirk Hovy, editors. 2019. *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*. Association for Computational Linguistics, Hong Kong.

Karl Pearson. 1896. VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318.

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

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

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

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

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.

Susan Phillips. 2007. *Automated Essay Scoring: A Literature Review*. Society for the Advancement of Excellence in Education. Google-Books-ID: EA7qTX6YOIYC.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical Modality Tagging from Rule-based Annotations and Crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea. Association for Computational Linguistics.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained Models for Natural Language Processing: A Survey. *Science China Technological Sciences*, 63(10):1872–1897. ArXiv:2003.08271.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Joy M. Reid. 1993. *Teaching ESL writing*. Englewood Cliffs, N.J. : Regents/Prentice Hall.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. *arXiv preprint*. ArXiv:1707.09861.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv preprint*. ArXiv:1602.04938.

Jessica Richardi. 2022. *What Is Classical Education? Using Curriculum Theory to Define A Classical Approach to K-12 Schooling*. Ph.D. thesis, University of Rhode Island, Kingston, RI.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated Cross-prompt Scoring of Essay Traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13745–13753. Number: 15.

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 preprint*. ArXiv:2008.01441.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 859–866, Reykjavik, Iceland. European Language Resources Association (ELRA).

D Royce Sadler. 2009. Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education - ASSESS EVAL HIGH EDUC*, 34:159–179.

Veronica Schmalz and Alessio Brutti. 2022. Automatic Assessment of English CEFR Levels Using BERT Embeddings. pages 293–299.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, pages 462–472, New York, NY, USA. Association for Computing Machinery.

Takumi Shibata and Masaki Uto. 2022. Analytic Automated Essay Scoring based on Deep Neural Networks Integrating Multidimensional Item Response Theory.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking Interpretability in the Era of Large Language Models. *arXiv preprint*. ArXiv:2402.01761.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv preprint*. ArXiv:1206.2944.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, USA. Association for Computational Linguistics.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical Chaining for Measuring Discourse Coherence Quality in Test-taker Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 950–961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

C. Spearman. 1987. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 100(3/4):441–471. Publisher: University of Illinois Press.

Kun Sun and Rong Wang. 2024. Automatic Essay Multi-dimensional Scoring with Fine-tuning and Multiple Regression. *arXiv preprint*. ArXiv:2406.01198 [cs].

Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Yi Tay, Minh Phan, Luu Tuan, and Siu Hui. 2017. SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

The Learning Agency Lab. 2023. The Feedback Prize: A Case Study In Assisted Writing Feedback Tools Working Paper.

Grigorios Tsoumakas and Ioannis Katakis. 2009. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3:1–13.

Kanishka Tyagi, Chinmay Rane, Harshvardhan, and Michael Manry. 2022. Chapter 4 - Regression analysis. In Rajiv Pandey, Sunil Kumar Khatri, Neeraj kumar Singh, and Parul Verma, editors, *Artificial Intelligence and Machine Learning for EDGE Computing*, pages 53–63. Academic Press.

University of Cambridge Local Examinations Syndicate. 2001. *FCE – First Certificate in English Handbook*. UCLES: Cambridge.

Masaki Uto. 2021. A review of deep-neural automated essay scoring models. *Behaviormetrika*, 48(2):459–484.

Masaki Uto and Maomi Ueno. 2018. Empirical comparison of item response theory models with rater's parameters. *Heliyon*, 4(5):e00622.

Shixiao Wang. 2024. DeBERTa with hats makes Automated Essay Scoring system better. *Applied and Computational Engineering*, 52:45–54.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

David Williamson, Xiaoming Xi, and F. Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31:2–13.

Joshua Wilson and Rod D. Roscoe. 2020. Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy. *Journal of Educational Computing Research*, 58(1):87–125. Publisher: SAGE Publications Inc.

Congying Xia, Chen Xing, Jiangshu Du, Xinyi Yang, Yihao Feng, Ran Xu, Wenpeng Yin, and Caiming Xiong. 2024. FOFO: A Benchmark to Evaluate LLMs' Format-Following Capability. *arXiv preprint*. ArXiv:2402.18667.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, pages 293–305, New York, NY, USA. Association for Computing Machinery.

Jin Xue, Xiaoyi Tang, and Liyan Zheng. 2021. A Hierarchical BERT-Based Transfer Learning Approach for Multi-Dimensional Essay Scoring. *IEEE Access*, 9:125403–125415. Conference Name: IEEE Access.

Taichi Yamashita. 2024. An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics*, 3(3):100133.

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.

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of Automated Text Scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223, Denver, Colorado. Association for Computational Linguistics.

Haoran Zhang and Diane Litman. 2018. Co-Attention Based Neural Network for Source-Dependent Essay Scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 399–409, New Orleans, Louisiana. Association for Computational Linguistics.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.*, 15(2):20:1–20:38.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint*. ArXiv:2306.05685 [cs].

Wentao Zhong. 2024. Effectiveness of finetuning pre-trained BERT and deBERTa for automatic essay scoring. *Applied and Computational Engineering*, 52:87–95.

A Analytic AES Datasets

Table 1 records the main public datasets of analytically scored essays. We compare them along seven dimensions:

1. **Essay Types**: the types of essays present in the corpus—argumentative (A), response to reading (R), narrative or creative (N), comment (C), suggestion (S) and letter (L);
2. **Writers' Information**: the language and academic levels of the essay writers;
3. **No. of Essays**: the total number of essays present in the corpus;
4. **Analytic Traits**: the linguistic dimensions (different from holistic) on which the essays have been graded;
5. **No. of Raters**: the number of individual raters (i.e., awarded marks) per essay;
6. **Multi-marks Available?**: whether those raw marks have been made publicly available (Yes), as opposed to only the aggregate scores (No); and
7. **Score Ranges**: the score range of the essays for a given dimension.

A.1 ICLE++

The International Corpus of Learner English (ICLE) is a corpus of essays written by upper-intermediate and advanced non-native English learners. The first version of the corpus, released in 2002, contained 2.5 million words produced by learners from 11 L1s (Granger, 2003). The corpus has since grown to 5.7 million words from 25 L1s (Granger et al., 2020). Concurrently, the Human Language Technology Research Institute in the University of Texas at Dallas, USA, contributed to the corpus by annotating subsets of it along several traits (Persing et al., 2010; Persing and Ng, 2013, 2014, 2015; Ke and Ng, 2019).

This effort culminated in the release of the ICLE++ dataset[6], which includes the annotation of 1,006 ICLE essays with both holistic scores and ten analytic scores (see Table 1). For the precise definitions of these traits, refer to Li and Ng (2024). This particular sample of essays was chosen in response to 10 specific prompts, chosen to be well-represented in multiple languages, to support as much L1 diversity as possible. In this annotation, each essay was graded by two different annotators, and disagreement were resolved through open discussion. The raw multi-mark scores have recently been released.

A.2 ASAP++

The Automated Student Assessment Prize (ASAP) dataset was introduced as part of the "The Hewlett Foundation: Automated Essay Scoring" Kaggle competition in 2012[7] and has since been widely used in AES research, both for prompt-specific (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017; Tay et al., 2017) and cross-prompt (Phandi et al., 2015; Cummins et al., 2016; Jin et al., 2018; Ridley et al., 2020) tasks. The original dataset contains eight different essay sets, one for each of the eight prompts considered, for a total of 12,980 essays written by native English speaking children between grades 7 and 10.[8] Marking guidelines and rubrics specific to each prompt were provided, and all essays were holistically marked by two (or three) independent human raters. Additionally, the essays for Prompts 7 and 8 were analytically scored by two markers: the multi-marks can be found in the original dataset. Subsequently, Mathias and Bhattacharyya (2018) provided single-marked analytic scores for the remaining six prompts to form the ASAP++ dataset.[9]

A.3 CELA

The Chinese EFL Learners' Argumentation (CELA) dataset[10] is a collection of 144 argumentative essays written by undergraduate students in non-English majors in China first introduced by Xue et al. (2021). Participants were asked to write a 300-word essay in response one single prompt. Subsequently, two expert raters independently scored the essays both holistically and along five analytic sub-scales (Grammar, Lexicon, Global and Local Organisation, and Supporting Ideas). The final dataset only records the average score of the two rater scores for each essay trait,

[6] The annotations are available via https://github.com/samlee946/ICLE-PlusPlus.

[7] The original dataset and annotation guidelines can be downloaded from https://www.kaggle.com/c/asap-aes/data.

[8] According to the K-12 (from kindergarten to 12th grade) curriculum (Richardi, 2022)

[9] These can be downloaded from https://lwsam.github.io/ASAP++/lrec2018.html.

[10] The dataset is available at https://github.com/gzutxy/CELA.

Table 1: Comparison of known analytic AES corpora.

Corpora	Essay Types	Writers' Information	No. of Essays	Analytic Traits (≠ Holistic)	No. of Raters	Multi-marks Available ?	Score Ranges
ICLE++	A	Non-native; undergraduate students	1,006	Prompt Adherence Thesis Clarity Argument Strength Development Organisation Coherence Cohesion Sentence Structure Vocabulary Technical Quality	2	Yes	1-4 (half-point increments)
ASAP++	A, R, N	US students; Grades 7-10	12,980	Content/Ideas Conventions Organisation Prompt Adherence Language Sentence Fluency Word Choice Voice Style	1-3	Partly	0-3, 0-4, and 1-6 (prompt-dependent; integer scales)
CELA	A	Non-native; undergraduate students in China	144	Grammar Lexicon Global Organisation Local Organisation Supporting Ideas	2	No	1-8 (integer scales)
ELLIPSE	A, N, C, S, L	Non-native; Grades 8-12	6,482	Cohesion Syntax Vocabulary Phraseology Grammar Conventions	2-3	Yes	1-5 (half-point increments)
ICNALE GRA	A	Asian English language learners Native English	136 4	Intelligibility Complexity Accuracy Fluency Comprehensibility Logicality Sophistication Purposefulness Willingness Involvement	80	Yes	0-10 (half-point increments)

not the raw multi-marks.

A.4 ELLIPSE Corpus

The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus was released by the Vanderbilt University and the Learning Agency Lab[11] in 2022 for the "Feedback Prize - English Language Learning" Kaggle competition[4] (Crossley et al., 2024). The full dataset contains 6,482 essays written by English language learners between the 8th and 12th grade on 29 different prompts as part of state-wide standardised writing assessments in the 2018/19 and 2019/20 school years in the United States (US).[12]

All essays were independently marked by a minimum of two raters along six analytic dimensions Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions which are defined in Crossley et al. (2024, Table 1).[13], as well as a holistic score. All scores follow a 9-point Likert scale and range from 1.0 to 5.0 with increments of 0.5, where a maximal score in one of these dimensions signifies a native-like proficiency. Any disagreement between raters was adjudicated in a discussion between the two parties and both mean and raw scores have been published. Finally, the authors of the dataset conducted an MFRM analysis for the raters and essays and found the scores to be reliable (Crossley et al., 2024).

[11] See https://www.the-learning-agency-lab.com.

[12] The dataset can be downloaded from https://github.com/scrosseye/ELLIPSE-Corpus.

[13] These were identified by teaching and research advisory boards of experts in the fields of composition and ELL education as the principal components of language acquisition (The Learning Agency Lab, 2023).

Table 2: Best hyper-parameter settings for each of the different pre-trained models when fine-tuned on the CLC FCE corpus.

Model	No. of Parameters	No. of Epochs	Batch Size	Learning Rate	Weight Decay
`microsoft/deberta-v3-base`	184M	7	8	4.02e-5	8.98e-2
`roberta-base`	125M	6	8	2.02e-5	6.20e-2
`bert-base-cased`	109M	7	16	4.16e-5	2.87e-2
`bert-base-uncased`	109M	7	8	4.47e-5	4.28e-2
`distilbert-base-cased`	65.8M	4	8	6.87e-5	6.26e-2
`distilbert-base-uncased`	65.8M	6	16	3.32e-5	3.96e-2

A.5 ICNALE GRA

The Global Rating Archive (GRA) was developed as part of the International Corpus Network of Asian Learners of English (ICNALE) corpus (Ishikawa, 2020, 2023), a corpus of more than 15,000 samples of Asian ELLs' essays, monologues, and speeches. In particular, GRA includes 140 essays written to one single prompt on the topic of part-time jobs for college students. Of those essays, 136 were written by Asian ELLs representing ten different regions, and the remaining four were written by native English. Most uniquely, the essays were independently marked by 80 human raters both holistically, and analytically for 10 different essay traits. See Ishikawa (2020, 2023) for a detailed description of the corpus.

B Choosing DeBERTa

To motivate our choice of underlying baseline model (Section 4.2), we considered six variants of the pre-trained BERT model (Devlin et al., 2019), which have been successfully applied to AES in the past (Mayfield and Black, 2020; H. Beseiso, 2021; Schmalz and Brutti, 2022) and are particularly easy to use. Each was then fine-tuned on the seminal holistic AES dataset (Ke and Ng, 2019): the CLC FCE corpus (Yannakoudakis et al., 2011).[14] This dataset is a collection of 2,469 short essays written by English language learners (ELLs) from around the world who sat the Cambridge English for Speakers of Other Languages (ESOL) First Certificate in English examinations between 2000 and 2001. Essays were marked by an examiner with a 0–5 band score using the rubric from the University of Cambridge Local Examinations Syndicate (2001, p.19). Following Yannakoudakis et al. (2011), we mapped these scores to a 0–20 linear scale, ideal for a regression task. Table 2 shows a summary of the models we considered, their size (in number of parameters), and the best hyper-parameters values we obtained for each in the step-by-step method in Appendix C.4.

Table 3 shows the average performance of the different models for the best hyper-parameter setting in Table 2 across the five random seeds. DeBERTa (He et al., 2021) outperforms all of the other models across all five of our evaluation metrics (Appendix C.3), obtaining a record low RMSE score of 2.308 for the random seed `1002`. However, it is also the model that has the largest variance across different random seeds for RMSE, accuracy, precision and recall, which suggest that the model is not the most robust to random-seed instability (Madhyastha and Jain, 2019). Further, DeBERTa is more heavy-weight than the other models (i.e., it is larger in terms of number of parameters; Table 2), and thus, takes more time to train. But despite these limitations, we chose to use DeBERTa for the next part of the experiments because it unambiguously surpassed all the other candidates.

C Methodology

In this section, we describe the research methodology we plan to use for running our ML experiments. Note that this may be improved in the future. This same methodology was used in the experiment described in Appendix B.

C.1 Reproducibility

Ensuring the computational reproducibility of a project is very important both to allow others to build on the research and for its credibility: anyone should be able to obtain the same results if they use the exact data, models and code provided by the authors (Donoho et al., 2009). When it comes to ML, many model architectures and algorithms are by nature non-deterministic (Reimers and Gurevych, 2017). To overcome this, it is standard practice to set a random seed, making any subsequent "random" number deterministic.

[14] Note that at the time of running these experiments, the new corrected version of this dataset had not been published.

Table 3: Average performance of the different models on the CLC FCE test set using 0–20 scores as in Yannakoudakis et al. (2011) across the five random seeds (rounded to 3 decimal places) for the best hyper-parameter setting in Table 2 (Avg.). In (+), the difference between the average and the maximal value achieved for each metric for a particular seed, and in (−) the difference between the average and the minimal values. Together they show the variation of performance across the five seeds for a metric: the largest ranges are underlined for each metric.

Model		RMSE	Pearson	Spearman	Acc.	Prec.	Rec.	F1
`microsoft/deberta-v3-base`	Avg.	**2.705**	**0.690**	**0.680**	**0.152**	**0.134**	**0.135**	**0.115**
	+	0.477	0.025	0.034	0.040	0.042	0.023	0.037
	−	0.397	0.022	0.021	0.030	0.041	0.017	0.027
`roberta-base`	Avg.	2.927	0.252	0.257	0.137	0.009	0.069	0.017
	+	0.103	0.274	0.252	0.001	0.001	0.002	0.000
	−	0.045	0.326	0.269	0.004	0.000	0.002	0.001
`bert-base-cased`	Avg.	2.959	-0.022	-0.048	0.137	0.014	0.071	0.022
	+	0.076	0.351	0.364	0.007	0.010	0.004	0.010
	−	0.068	0.171	0.242	0.004	0.005	0.004	0.006
`bert-base-uncased`	Avg.	2.848	0.420	0.402	0.126	0.038	0.076	0.031
	+	0.151	0.110	0.153	0.015	0.033	0.023	0.018
	−	0.094	0.227	0.250	0.026	0.028	0.013	0.014
`distilbert-base-cased`	Avg.	2.949	0.305	0.363	0.135	0.027	0.078	0.031
	+	0.184	0.210	0.137	0.017	0.013	0.018	0.020
	−	0.238	0.270	0.065	0.013	0.017	0.008	0.014
`distilbert-base-uncased`	Avg.	3.953	0.183	0.098	0.122	0.009	0.069	0.015
	+	0.365	0.048	0.086	0.005	0.000	0.002	0.001
	−	0.267	0.087	0.056	0.003	0.001	0.002	0.000

```
random.seed(SEED)
set_seed(SEED)
torch.manual_seed(SEED)
torch.cuda.manual_seed_all(SEED)
np.random.seed(SEED)
os.environ['PYTHONHASHSEED']=str(SEED)
```

```
torch.backends.cudnn.deterministic = True
torch.backends.cudnn.benchmark = False
torch.use_deterministic_algorithms(True)
```

Figure 3: The code we use to set the random seed to the different Python packages needed in the experiments (top), and some additional lines needed to achieve consistent results with the `microsoft/deberta-v3-base` model in Appendix B.

We run the experiments with five different randomly chosen seeds[15] for better comparability and to ensure that the results we are seeing are not sub-optimal. See Figure 3 for the code we use to ensure the reproducibility of the results.

C.2 Hyper-parameter Optimisation

The process of hyper-parameter optimisation consists of finding the set of optimal hyper-parameters (parameters whose values control the learning process of an ML model; Goodfellow et al., 2016, Chapter 8). We use the Bayesian hyper-parameter optimisation algorithm (Snoek et al., 2012) as implemented by Comet ML[16], a search algorithm that is based on distributions and balances exploitation/exploration to make decisions about which hyper-parameter values to try next. This approach achieves optimal results with many less trials. Figure 4 shows the configuration details that we use (i.e., objective function, hyper-parameters considered and value ranges).

[15] Specifically, the random seeds `1601`, `2911`, `1044`, `1002`, and `2510` were used in the experiments of Appendix B.

[16] See https://www.comet.com/docs/v2/guides/optimizer/configure-optimizer/ for more details.

```
{
 "algorithm": "bayes",
   "spec": {
       "maxCombo": 40,
       "objective": "minimize",
       "metric": "eval_rmse",
       "minSampleSize": 100,
       "retryLimit": 20,
       "retryAssignLimit": 5,
   },
   "parameters": {
       "batch_size": {"type": "discrete", "values": [8, 16, 32]},
       "learning_rate": {"type": "double", "min": 1e-7, "max": 1e-4},
       "num_train_epochs": {"type": "integer", "min": 2, "max": 8},
       "weight_decay": {"type": "double", "min": 0.0, "max": 0.1}
   },
}
```

Figure 4: Extract of the Comet ML `Optimizer` configuration file used in experiments.

C.3 Evaluation and Reporting

Within the field of AES, the evaluation of scoring systems is traditionally carried out by comparing a systems' predicted scores to the gold standard labels for a held-out validation set of essays using a series of metrics (Williamson et al., 2012; Yannakoudakis and Cummins, 2015). Specifically, we report:

1. the Root Mean Squared Error (RMSE) (Tyagi et al., 2022);
2. the correlation between the predicted and gold standard scores with both the Pearson (Pearson, 1896) and Spearman Rank correlation coefficients (Spearman, 1987);
3. as well as the main classification metrics (precision, recall, accuracy and F1-score; Hossin and M.N, 2015) by rounding model predictions to the closest grade class (e.g., ELLIPSE uses a 1.0 to 5.0 scale with increments of 0.5; Section A.4).

C.4 Step-by-step Method

Having introduced the individual components of the experimental methodology, we now give below the step-by-step process we use to train, evaluate and test our models:

1. Start by running the Bayesian Hyper-parameter Optimisation algorithm for every one of the five random seeds. Given a random seed:
   (a) we use stratified data sampling to randomly split the dataset of essays into three parts using the `train_test_split()` function of the `scikit-learn`[17] Python library using a ratio of 70/15/15% for the training, validation and test sets respectively to limit sampling error;
   (b) then at each step of the algorithm (the total number of steps is given by the `maxCombo` field in Figure 4 which we set to 40), a different set of hyper-parameters values (Section C.2) is considered. With each, a model is trained from scratch on the training set, and then evaluated using the RMSE on the validation set to inform the next set of hyper-parameters the optimiser will try.
2. From step 1, retain the set of hyper-parameter settings that achieved the best results on the validation set in terms of the RMSE metric across the five random seeds, and round the

[17] For the documentation, see https://pypi.org/project/scikit-learn/.

learning rate and weight decay values to 3 significant figures.

3. Finally, re-run the experiments for five all seeds with the setting obtained in step 2 and report the maximum, minimum and average of every evaluation metric mentioned in Section C.3 across the five seeds on the test set.

Note that for the training and testing of models, we use the Trainer[18] interface. By default, Trainer implements the AdamW stochastic gradient descent optimisation method, an Adam algorithm (Kingma and Ba, 2017) with weight decay fix, as introduced by Loshchilov and Hutter (2019). Using AdamW optimisation has become the standard, and models trained with it generally yield better results than those trained without (Loshchilov and Hutter, 2019). Further, we use each model's default regression training loss, which is typically the Mean Square Error (MSE), implemented with the `MSELoss()` function from the PyTorch library[19] (Paszke et al., 2019). Finally, Trainer is set up such that model weights are saved after each training epoch and only the best model is loaded at the end of training with regards to the RMSE metric.

[18] See https://huggingface.co/docs/transformers/main_classes/trainer for a full documentation.

[19] The library can be access from https://pypi.org/project/torch/.