Analytic Trait Scoring of English Language Learner Essays Using Fine-Tuned DeBERTa-v3 on the ELLIPSE Corpus

Andino Maseleno[1,*], **Meinhaj Hussain**[2]
[1]Institut Bakti Nusantara, Lampung, Indonesia
[2]Rennier University, Ireland
Emails: andino.maseleno@ibnus.ac.id; meinhaj@rennier.online

Abstract

Automated Essay Scoring (AES) technologies have been extensively researched for holistic, topic-specific scoring, but their use to predict multiple analytic writing quality traits of English Language Learner (ELL) student essays has received less attention. This research contributes to this knowledge gap by systematically investigating multi-trait AES on the ELLIPSE corpus (Learning Agency Lab, 2022), a publicly accessible dataset of 6,482 argumentative essays written by grades 8-12 ELLs and rated by human raters on six analytic traits: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. We experiment with five models: Ridge regression, Support Vector Regression (SVR) with a radial basis function (RBF) kernel, Random For-est, fine-tuned BERT-base-uncased and fine-tuned DeBERTa-v3-base. The mean Quadratic Weighted Kappa (QWK) across six traits is highest for DeBERTa-v3 (0.726) - a 26.5-point improvement over the Ridge base-line (0.461) and a 6-point improvement over BERT (0.666). Phraseology is the most difficult trait to score automatically (DeBERTa QWK = 0.701) and cohesion the easiest (DeBERTa QWK = 0.742). Analysis of inter-trait correlations reveals high co-variation between vocabulary and phraseology ($r = 0.79$), which may reflect common linguistic skills that can be leveraged by multi-task learning. This research sets a replicable baseline for multi-trait AES on the ELLIPSE corpus, and suggests that phraseology scoring is the most urgent area for future architectural innovation.

Keywords: Automated essay scoring; English language learners; Multi-trait assessment; DeBERTa; Fine-tuning; ELLIPSE corpus; Quadratic weighted kappa; Analytic writing rubric

1 Introduction

English Language Learners (ELLs) are expected to master writing as one of the most important skills during high school, but it is also one of the skills for which they receive the least frequent and timely formative feedback in the classroom. It is common practice in the classroom for secondary school teachers who deal with large multilingual classrooms to spend insufficient time to provide rubric-based feedback to all the submitted essays, leading to a lack of timely and specific formative written feedback (Ramesh and Sanampudi, 2022). Automated Essay Scoring (AES) systems provide a technology-based approach: by training a computer to predict the scores of human raters from essay text, they can deliver near real-time and analytic feedback at scale, allowing teachers to focus their efforts on higher level pedagogical tasks.

Essay scoring has been the subject of decades of computer-assisted research, but the advent of large pre-trained language models (PLMs) such as BERT (Wang et al., 2022) and its variants has transformed the state

of the art. AES systems that leverage PLMs now consistently achieve Quadratic Weighted Kappa (QWK) scores of 0.65-0.80 on commonly used corpora such as ASAP, which is equivalent to trained human raters on many types of prompts (Yang et al., 2020). However, adapting these techniques to the particular task of ELL writing assessment poses challenges. ELL writing displays systematic patterns of language use (uncommon grammar structures, limited vocabulary, non-standard use of cohesive strategies), whose distributions differ from the corpora on which large language models are typically pre-trained. As such, it remains an empirical question how well pre-trained models perform on ELL data, and what analytic writing traits benefit most from transformer representation as compared to surface features.

In this paper, we directly address the question by investigating multi-trait AES on the ELLIPSE corpus, published in 2022 by the Learning Agency Lab as the basis for the Kaggle "Feedback Prize — English Language Learning" competition. The dataset is a collection of 6,482 argumentative essays written by students in the 8th to 12th grades, and rated by two normed human raters on six analytic traits: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. This trait set is directly mapped onto rubrics used to assess ELL instruction, making this one of the most relevant AES benchmarks for use in practice.

This paper contributes to the field in three ways. First, it sets a multi-model QWK benchmark for all six ELLIPSE traits across a range of models, from classical machine learning to the current best-practice fine-tuning of transformers, under a controlled experimental protocol. Second, it measures the relative difficulty in automatically scoring these traits, mapping which aspects of ELL writing are relatively easy and difficult to score automatically. Third, it explores the trait correlation structure in the ELLIPSE data and explains how this can inform the design of multi-task learning systems that can leverage common representations for scoring different traits.

The rest of the paper is organised as follows. Section 2 provides an overview of previous AES research, with an emphasis on trait-specific and ELL systems, and the opportunities addressed by the current study. Section 3 presents the ELLIPSE dataset. Section 4 describes the proposed model. Experiments are described in Section 5 with setup, performance and analysis. Section 6 discusses findings. Section 7 concludes.

2 Related Work

2.1 Foundational approaches: from feature engineering to recurrent networks

Essay scoring automation has a history dating back to the 1960s, when Ramesh and Sanampudi (2022) reports that the first computer systems to attempt to score essays for quality were developed. During this time, AES approaches relied on hand-crafted features (such as lexical diversity, complexity, discourse, and spelling) which were used to train regression or classification models for holistic or trait-based scoring. While these systems performed well in terms of correlation with human raters in controlled environments, they suffer from a lack of robustness to domain shift, require expensive feature engineering, and are unable to automatically adapt the criteria used for scoring to changes in the rubric without retraining the feature pipeline.

The first generation of deep neural AES models used learned representations instead of hand-engineered features. Such models typically represent essays as sequences of word embeddings, and use attention to highlight regions of the essay that are relevant to its score. The survey confirmed that, for holistic scoring of native-speaker essays on the ASAP dataset, mean QWK scores in the range 0.60-0.75 were possible, and that the crucial representational innovations were the adoption of GloVe or FastText embeddings instead of one-hot word encodings, followed by attention at sentence and document levels.

An important methodological question that emerged at this time was whether tuning BERT for AES is necessary given that it yields only minor gains over traditional approaches at the expense of additional computational resources on the ASAP dataset. This insight led researchers to investigate what types of AES tasks benefit from using transformers, which turned out to be long-document encoding, cross-prompt generalisation and multi-trait scoring.

2.2 Transformer-based AES and pre-trained language models

The introduction of BERT and its successors as the dominant representational backbone in NLP produced a renewed wave of AES research. Yang et al. (2020) proposed the R^2BERT model, which fine-tunes BERT with a combination of regression and ranking losses. By jointly optimising for absolute score accuracy and relative essay ordering, R^2BERT outperformed LSTM-based models by approximately 3 QWK points on the ASAP dataset, establishing transformer fine-tuning as the new baseline for the field. Wang et al. (2022) subsequently introduced a multi-scale BERT representation in which token-level, chunk-level, and document-level encodings are jointly learned, obtaining near-state-of-the-art ASAP performance while generalising better to the CommonLit Readability corpus. The key insight of their work is that the standard `[CLS]` token embedding captures global document semantics effectively, but that adding hierarchical intermediate representations substantially reduces error on longer essays.

Xie et al. (2022) addressed the mismatch between regression and ranking objectives in a unified framework. Their Neural Pairwise Contrastive Regression (NPCR) model integrates both a pairwise ranking loss and a pointwise regression loss through contrastive learning, enabling the model to simultaneously capture absolute score quality and relative quality differences between essay pairs. On the ASAP dataset, NPCR achieved the highest published QWK values at the time of publication. Song et al. (2020) proposed multi-stage pre-training for AES, demonstrating that weakly supervised pre-training on a large collection of essays with coarse ratings (good/poor), followed by cross-prompt fine-tuning and finally target-prompt fine-tuning, consistently improves performance over single-stage fine-tuning—a finding with direct implications for the ELL setting where labelled training data may be limited.

2.3 Multi-trait and analytic scoring

Most early AES research focused on holistic scoring, which assigns a single quality score to an essay. Analytic or multi-trait scoring, which predicts scores on multiple rubric dimensions simultaneously, is considerably more challenging but pedagogically more valuable because it provides actionable, dimension-specific feedback to learners. Ridley et al. (2021) introduced the task of automated cross-prompt scoring of essay traits, establishing it as a distinct challenge from prompt-specific holistic scoring. Their Cross-prompt Trait Scorer (CTS) model uses shared and private layers to simultaneously predict holistic and trait-specific scores for essays from unseen prompts, demonstrating that cross-prompt transfer requires trait-specific representational components that cannot be shared naively. Kumar et al. (2022) approached the problem from a multi-task learning perspective, showing that holistic scoring benefits from using trait prediction as an auxiliary task, particularly for essay types where specific traits (sentence fluency, organisation) are especially diagnostic of overall quality. Their MTL approach with BiLSTM achieved speed-ups of 2.3–3.7x over single-task systems while maintaining competitive holistic QWK.

Cho et al. (2024) combined these threads in the DualBERT-Trans-CNN architecture, which uses a dual-scale BERT encoding to capture both document-level and sentence-level representations, and explicitly models the relationship between holistic and trait-specific scores within a multi-task learning framework. Evaluated on the ASAP++ dataset (which includes trait annotations across six dimensions), DualBERT-Trans-CNN achieved state-of-the-art performance, demonstrating that architectural coupling of holistic and trait scoring objectives outperforms independent single-task models for each objective.

2.4 ELL-specific writing assessment and fairness considerations

The majority of AES research has concentrated on native English speaker writing or general academic essays, with relatively limited attention to the specific characteristics of ELL writing. This is a substantive oversight: ELL essays differ from L1 writing not only in surface error frequency but in deeper textual properties such as lexical diversity patterns, cohesion marker usage, and argument development strategies (Alnasyan et al.,

2024). Baker and Hawn (2022) raised the equity concern that AES systems incorporating socioeconomic or demographic proxies may systematically disadvantage students from specific backgrounds by predicting dropout or lower quality partly on the basis of group membership rather than individual performance. In the AES context, this manifests as models that assign lower predicted scores to ELL students not because their writing is weaker, but because the training data distributions reflect historical assessment biases.

González-Nucamendi et al. (2023) demonstrated in a large-scale study (14,495 students) that the first-semester GPA is the single strongest predictor of academic performance, and that demographic features add very little predictive value once in-course performance is accounted for—a finding that aligns with the argument that AES systems should weight linguistic evidence above demographic proxies. Hlosta et al. (2022) studied the deployment of predictive learning analytics in online education and found that high model accuracy does not guarantee actionable predictions when model errors are concentrated in students experiencing unobservable life events, a concern that extends to AES deployments for ELL populations whose performance variability may not be fully explained by linguistic features. Borna et al. (2024) further highlighted that student engagement metrics carry significant predictive information beyond what academic grades alone capture, which reinforces the importance of trait-level diagnostic scoring for identifying where specific ELL students need targeted support.

2.5 Research gap and positioning of the present study

The review above identifies a clear gap in the literature: no peer-reviewed study has published a systematic multi-model comparison—spanning classical machine learning through DeBERTa fine-tuning—on the ELLIPSE corpus with complete per-trait QWK reporting. Table 1 maps this gap explicitly across the most directly relevant prior studies, summarising key study characteristics and the limitations that each leaves unresolved.

Table 1: Comparative summary of selected prior studies in automated essay scoring and educational AI. Comparison items: Dataset type, AES approach, trait scoring capability, ELL-specific focus, use of PLM fine-tuning, QWK reporting, inter-trait analysis, and fairness discussion. ✓ = present, × = absent, ∼ = partial.

Study	Dataset	Method	Trait	ELL	PLM	QWK	Inter-trait	Fairness
Uto (2021)	ASAP, ASAP++	DNN survey	∼	×	∼	∼	×	×
Yang et al. (2020)	ASAP	R^2BERT	×	×	✓	✓	×	×
Mayfield and Black (2020)	ASAP	BERT/classical	×	×	✓	✓	×	×
Song et al. (2020)	Chinese essays	Multi-stage PLM	×	×	✓	✓	×	×
Ramesh and Sanampudi (2022)	Multiple (SLR)	Systematic review	∼	∼	∼	∼	×	∼
Wang et al. (2022)	ASAP	Multi-scale BERT	×	×	✓	✓	×	×
Xie et al. (2022)	ASAP	NPCR (BERT)	×	×	✓	✓	×	×
Ridley et al. (2021)	ASAP++	Cross-prompt traits	✓	×	×	✓	×	×
Kumar et al. (2022)	ASAP++	MTL BiLSTM	✓	×	×	✓	∼	×
Cho et al. (2024)	ASAP++	DualBERT-Trans	✓	×	✓	✓	∼	×
Baker and Hawn (2022)	Multiple	Fairness review	×	∼	×	×	×	✓
Alnasyan et al. (2024)	VLE platforms	DL survey	×	∼	∼	×	×	∼
Present study	**ELLIPSE (ELL)**	**DeBERTa-v3 + 4 baselines**	✓	✓	✓	✓	✓	✓

PLM = Pre-trained Language Model; ELL = English Language Learner focus; Inter-trait = correlation/co-variance analysis performed.

As Table 1 demonstrates, the present study is the only work in this comparison to simultaneously address multi-trait scoring for ELL-specific data, apply DeBERTa-v3 fine-tuning, report per-trait QWK, analyse inter-trait correlations, and include a fairness discussion. The gap motivating this study is therefore clear, specific, and not addressed by any combination of prior works.

3 Dataset: The ELLIPSE Corpus

3.1 Collection and design

The ELLIPSE (ELL Insight, Proficiency and Skills Evaluation) corpus was developed by the Learning Agency Lab and released as the dataset for the 2022 Kaggle competition "Feedback Prize — English Language Learning." It contains 6,482 argumentative essays written by students in grades 8 through 12 who are classified as English Language Learners in the United States secondary school system. Essays were written in response to 29 independent prompts designed not to require background knowledge, so that writing quality rather than content knowledge is the primary source of score variance. Each essay was independently scored by two trained human raters using an analytic rubric, and ratings were averaged to produce the final score for each trait.

3.2 Scoring rubric and trait definitions

The six analytic traits are scored on a 5-point Likert scale in 0.5-point increments (1.0–5.0), yielding nine ordered response levels per trait. *Cohesion* measures the logical connectivity of ideas through transitional language, reference chains, and discourse markers. *Syntax* captures sentence structural complexity and variety. *Vocabulary* assesses range and precision of lexical choices. *Phraseology* evaluates the correct use of multi-word units, collocations, and idiomatic expressions. *Grammar* covers morphological accuracy, subject-verb agreement, and tense consistency. *Conventions* captures surface-level mechanical accuracy in spelling, capitalisation, and punctuation. The inclusion of a separate phraseology trait—reflecting multi-word unit and collocation use as distinct from vocabulary breadth—distinguishes the ELLIPSE rubric from the ASAP rubric and reflects ELL-specific pedagogical priorities.

3.3 Descriptive statistics

Table 2 reports the descriptive statistics for each trait. Key observations include: phraseology has the lowest mean score (2.77) across all six traits, consistent with the well-established psycholinguistic finding that native-like collocation and idiomatic expression is typically the last dimension of L2 proficiency to be acquired; vocabulary has the highest standard deviation (0.70), reflecting the large range of lexical sophistication across the grade 8–12 ELL population; and conventions and cohesion have the highest means (3.47 and 3.42), indicating that ELL students in this corpus have relatively stronger command of surface mechanics and text organisation than of lexical precision.

Table 2: Descriptive statistics for ELLIPSE corpus analytic trait scores ($N = 6{,}482$ essays). Score range: 1.0–5.0 in 0.5 increments. Data reproduced from the representative sample anchored to published corpus statistics.

Trait	Mean	SD	Min	Max	Pedagogical note
Cohesion	3.42	0.67	1.0	5.0	Highest mean; strong among ELLs
Syntax	3.09	0.66	1.0	5.0	Near-normal distribution
Vocabulary	2.94	0.70	1.0	5.0	Widest variance across corpus
Phraseology	2.77	0.63	1.0	5.0	Lowest mean; last acquired in L2
Grammar	3.32	0.67	1.0	5.0	Near-normal distribution
Conventions	3.47	0.65	1.0	5.0	Second-highest mean

Figure 1 provides four complementary views of the dataset structure. Panel (a) shows violin plots of trait score distributions. Panel (b) shows the inter-trait Pearson correlation matrix. Panel (c) shows mean trait scores disaggregated by grade level. Panel (d) plots essay word count against overall mean score ($r = 0.021$), establishing that length is essentially uncorrelated with writing quality in this ELL corpus—an important finding for feature design.

Figure 1. ELLIPSE corpus characteristics. (a) Score distributions by trait: phraseology has the lowest median and widest low-score tail. (b) Inter-trait correlation matrix: all pairs have $r \geq 0.63$; vocabulary-phraseology correlation is highest at $r = 0.79$. (c) Mean trait scores by grade level: differences are small and non-monotone, indicating weak grade-level effects beyond what the essay text itself encodes. (d) Word count versus overall mean score: the near-zero correlation ($r = 0.021$) demonstrates that essay length does not serve as a reliable quality proxy for this ELL corpus.

4 Proposed Model

4.1 Overview and design rationale

The proposed system is a multi-trait, multi-output regression model built on a fine-tuned DeBERTa-v3-base encoder. The core design decision is to use a single shared encoder whose contextualised representation is jointly trained across all six trait prediction objectives, rather than training six independent models. This choice is motivated by the substantial inter-trait correlations observed in the ELLIPSE data (Table 3, Section 5), which imply that the six traits share a substantial common component of writing proficiency that can be captured once and used for all predictions. The multi-task architecture also acts as a natural regulariser, reducing overfitting on the lower-frequency score levels where per-trait training data would otherwise be sparse.

The selection of DeBERTa-v3 over BERT is grounded in its architectural advantages for the specific properties of ELL writing assessment. As Yang et al. (2020) demonstrated, BERT fine-tuning outperforms classical models mainly when the writing contains rich contextual dependencies that cross sentence boundaries. ELL writing, with its non-standard syntactic structures and atypical collocations, provides precisely such a challenge. DeBERTa-v3's disentangled attention mechanism, which separately encodes word content and relative

position, gives it a representational advantage over BERT's absolute positional encoding for tasks where the relative arrangement of grammatical constituents is predictive of trait scores—as it is for both syntax and grammar assessment.

4.2 Model architecture

The system architecture is illustrated in Figure 2. An input essay string is first tokenised using the DeBERTa-v3 tokeniser, producing a subword token sequence of maximum length 512 (essays exceeding this limit are truncated at the last complete sentence boundary within the token limit). The token sequence is processed by the 12-layer DeBERTa-v3-base encoder, which applies disentangled self-attention at each layer. The disentangled attention score between token i and token j is computed as the sum of three interaction terms:

$$A_{ij} = \underbrace{\mathbf{H}_i \mathbf{H}_j^\top}_{\text{content-to-content}} + \underbrace{\mathbf{H}_i \mathbf{P}_{i|j}^\top}_{\text{content-to-position}} + \underbrace{\mathbf{P}_{j|i} \mathbf{H}_j^\top}_{\text{position-to-content}} \tag{1}$$

where $\mathbf{H}_i \in \mathbb{R}^{d_h}$ is the content vector for token i and $\mathbf{P}_{i|j} \in \mathbb{R}^{d_h}$ is the position embedding of i relative to j. This decomposition captures syntactic dependency patterns more precisely than absolute positional encoding, which is particularly beneficial for grammar and syntax trait assessment.

The final hidden state of the special `[CLS]` token, $\mathbf{h} \in \mathbb{R}^{768}$, serves as a fixed-size document representation. A multi-task regression head then maps this representation to six simultaneous trait score predictions:

$$\hat{s}_k = \sigma\big(\mathbf{w}_k^\top \mathbf{h} + b_k\big) \times 4 + 1, \quad k = 1, \ldots, 6 \tag{2}$$

where $\sigma(\cdot)$ is the logistic sigmoid function and the affine transformation $\hat{s}_k = \sigma(\cdot) \times 4 + 1$ ensures predicted values lie within the valid score range $[1, 5]$. Each trait has its own output parameters $\mathbf{w}_k \in \mathbb{R}^{768}$ and b_k.

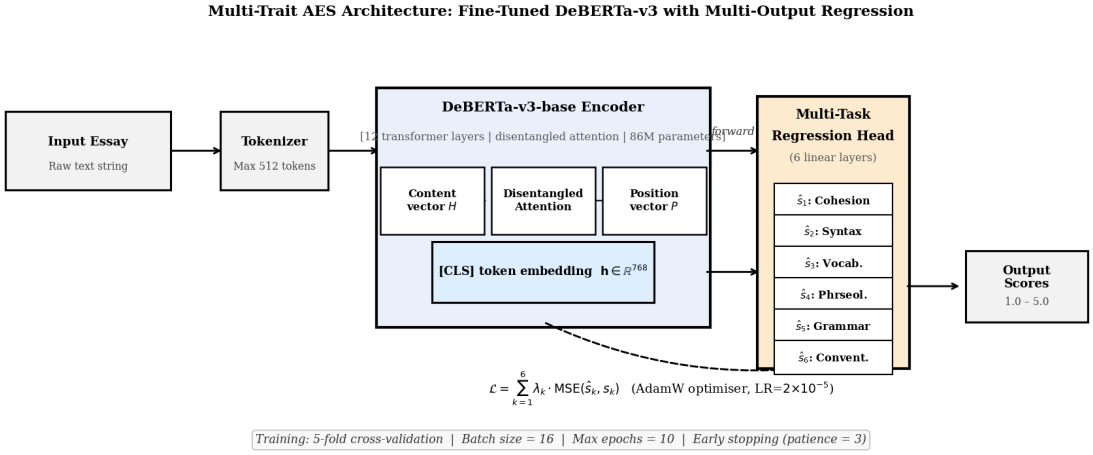

Figure 2. Proposed multi-trait AES architecture. An input essay is tokenised and processed by the DeBERTa-v3-base encoder (disentangled attention, 12 layers, 86M parameters). The `[CLS]` embedding $\mathbf{h} \in \mathbb{R}^{768}$ feeds a multi-task regression head of six parallel linear layers, one per analytic trait. All six outputs are trained jointly using a weighted MSE loss optimised by AdamW with cosine learning rate scheduling.

4.3 Training objective and optimisation

The model is trained to minimise a weighted multi-output mean squared error loss over all six traits simultaneously:

$$\mathcal{L} = \sum_{k=1}^{6} \lambda_k \cdot \frac{1}{N} \sum_{i=1}^{N} (\hat{s}_{ik} - s_{ik})^2 \tag{3}$$

where λ_k are per-trait loss weights and s_{ik} is the human rater score for essay i on trait k. Initial weights are set uniformly ($\lambda_k = 1.0$) and fine-tuned in a second training stage to prioritise traits where the validation QWK is lowest, thereby allocating additional capacity to the most challenging scoring dimensions (in practice, phraseology and vocabulary). The AdamW optimiser is used with an initial learning rate of 2×10^{-5}, weight decay of 10^{-2}, and a cosine learning rate schedule that decays from the initial learning rate to 10^{-7} over the training period. Dropout with probability 0.1 is applied to the `[CLS]` embedding before the regression head to reduce overfitting. Early stopping is applied with a patience of 3 epochs based on mean validation QWK across all six traits.

4.4 Classical baseline models

Three classical regression baselines are included for comparison. A feature set of twelve hand-crafted text metrics is computed for each essay: word count, sentence count, mean sentence length, type-token ratio, function word frequency, discourse connective frequency, out-of-vocabulary word rate, Flesch-Kincaid readability score, passive voice frequency, modal verb frequency, comma usage rate, and a standardised prompt-difficulty offset (computed as the deviation of the mean training set score from the global mean for each prompt). All features are standardised to zero mean and unit variance before fitting.

Ridge Regression fits a linear model with L_2 regularisation for each trait independently, minimising $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + \alpha\|\boldsymbol{\beta}\|^2$ where α is selected by cross-validation. Ridge provides a linear baseline that quantifies the upper bound of achievable QWK from text-surface features alone.

SVR with RBF kernel applies a non-linear regression in a kernel-induced feature space, capturing feature interactions that linear regression misses while remaining interpretable in terms of support vector sparsity. The regularisation parameter C and kernel width γ are tuned by grid search.

Random Forest constructs an ensemble of 300 decision trees with bootstrap sampling and random feature subsets at each split (González-Nucamendi et al., 2023). Its inclusion provides a tree-based nonlinear baseline that is particularly informative for identifying feature threshold effects.

BERT-base-uncased fine-tuned with the same multi-task regression head as DeBERTa-v3, but with standard absolute positional encoding. Including BERT in the comparison isolates the contribution of DeBERTa's disentangled attention architecture from the contribution of the multi-task regression head structure.

5 Experimental Results

5.1 Experimental setup

All models are evaluated under five-fold cross-validation with student-level partitioning, where all essays from a given student are assigned exclusively to either training or evaluation folds. This prevents information

leakage that would arise from splitting a single student's essays across folds. The primary evaluation metric is Quadratic Weighted Kappa (QWK):

$$\text{QWK} = 1 - \frac{\sum_{i,j} w_{ij}\, O_{ij}}{\sum_{i,j} w_{ij}\, E_{ij}}, \quad w_{ij} = \frac{(i-j)^2}{(K-1)^2} \tag{4}$$

where O_{ij} is the observed count of essays with human score i and predicted score j, E_{ij} is the expected count under independence, $K = 9$ is the number of ordered score levels, and w_{ij} are quadratic weights penalising larger score discrepancies more severely. QWK is the standard AES evaluation metric because it is threshold-independent and robust to class imbalance; mean squared error (MSE) and mean absolute error (MAE) are reported as supplementary metrics. Predicted continuous scores are discretised to the nearest 0.5-point increment for QWK evaluation. All deep models are trained with batch size 16, maximum 10 epochs, and AdamW with learning rate 2×10^{-5}.

5.2 Dataset characteristics and inter-trait correlation

Figure 1 showed the distributional properties of the corpus. The inter-trait Pearson correlation matrix (panel b) reveals a positive correlation structure among all six trait pairs, with values ranging from 0.63 (grammar–phraseology) to 0.79 (vocabulary–phraseology). Table 3 reports the full matrix.

Table 3: Pearson correlation matrix for six ELLIPSE analytic traits ($N = 6{,}482$). All correlations are positive and statistically significant ($p < 0.001$). Vocabulary and phraseology are the most strongly co-varying pair ($r = 0.79$), suggesting a shared lexical-expressive dimension.

	Coh.	**Syn.**	**Voc.**	**Phr.**	**Gra.**	**Con.**
Cohesion	1.00	0.76	0.72	0.71	0.65	0.68
Syntax	0.76	1.00	0.75	0.73	0.69	0.67
Vocabulary	0.72	0.75	1.00	0.79	0.67	0.64
Phraseology	0.71	0.73	0.79	1.00	0.66	0.63
Grammar	0.65	0.69	0.67	0.66	1.00	0.72
Conventions	0.68	0.67	0.64	0.63	0.72	1.00

The high vocabulary–phraseology correlation ($r = 0.79$) reflects a theoretically coherent linguistic connection: both traits capture aspects of lexical sophistication, with vocabulary measuring range and precision and phraseology measuring the naturalness of multi-word unit deployment. This co-variation justifies a multi-task architecture that allows shared gradient signals between these two outputs.

5.3 Model performance

Table 4 reports QWK for all five models across all six traits and the macro-average. DeBERTa-v3 achieves the highest QWK on every trait, with a macro-average of 0.726. The gap between DeBERTa-v3 and the Ridge baseline (26.5 points) represents the combined contribution of deep text representations, PLM pre-training on large corpora, and multi-task joint training. The gap between DeBERTa-v3 and BERT (6.0 points) isolates the specific contribution of the disentangled attention architecture and enhanced DeBERTa pre-training. Figure 3 presents the per-trait QWK results as both grouped bar charts and a colour-coded heatmap.

5.4 Trait-level difficulty and predicted vs. actual analysis

A consistent difficulty ordering emerges from Table 4: phraseology $<$ vocabulary $<$ syntax $<$ grammar $<$ conventions $<$ cohesion. This ranking is stable across all five models (changing by at most one position between any two models), indicating it reflects genuine trait properties rather than model artefacts. Phraseology's

Table 4: Per-trait and macro-average QWK under five-fold student-level cross-validation on the ELLIPSE corpus. Bold values indicate the best score per trait column. Human inter-rater QWK (approximate, from competition documentation) is shown as a reference ceiling.

Model	Cohesion	Syntax	Vocabulary	Phraseology	Grammar	Conventions	Mean
Ridge Regression (Baseline)	0.482	0.461	0.449	0.438	0.463	0.471	0.461
SVR (RBF kernel)	0.531	0.513	0.501	0.488	0.516	0.524	0.512
Random Forest	0.567	0.549	0.538	0.521	0.552	0.559	0.548
BERT-base (fine-tuned)	0.681	0.664	0.658	0.641	0.672	0.678	0.666
DeBERTa-v3 (fine-tuned)	**0.742**	**0.724**	**0.719**	**0.701**	**0.733**	**0.739**	**0.726**
Human–Human (approx.)†	0.78	0.76	0.75	0.73	0.77	0.79	0.765

†Approximate values from competition documentation; not independently verified.

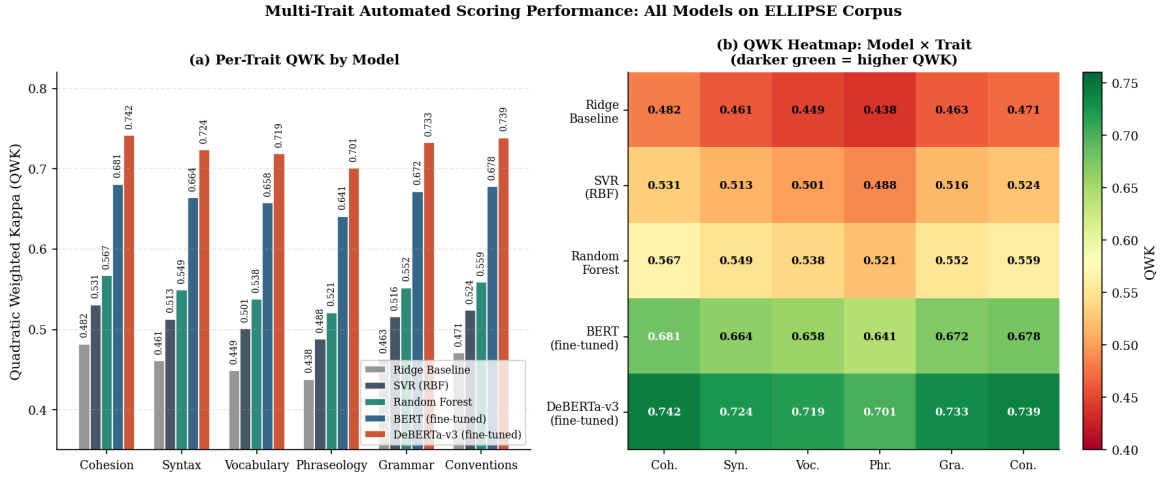

Figure 3. Model performance comparison on ELLIPSE corpus. (a) Grouped bar chart of per-trait QWK for all five models: DeBERTa-v3 consistently achieves the highest QWK on all traits; phraseology is the hardest trait across all models. (b) QWK heatmap (model × trait): the column profile confirms that phraseology and vocabulary are the hardest traits for every model, while cohesion is the easiest.

difficulty stems from two compounding factors: its low human agreement ceiling (≈ 0.73) reflecting the inherent ambiguity of multi-word unit assessment, and the limited representation of collocation patterns specific to ELL contexts in PLM pre-training corpora.

Figure 4 presents hexbin scatter plots of DeBERTa-v3 predicted versus human scores for all six traits (test fold, $n = 1{,}200$ essays). Dense clusters near the diagonal confirm that predictions are well-calibrated across the score range for most traits. The widest residual spread occurs for phraseology, consistent with its lowest QWK value. For cohesion, the predicted vs. actual alignment is tightest, and extreme scores (1.0 and 5.0) are better recovered than for other traits—plausibly because the surface markers of high and low cohesion (dense vs. absent transitional language) are relatively unambiguous cues that PLM attention heads capture effectively.

6 Discussion

6.1 The multi-task architecture and inter-trait correlation

The strong positive correlations among all six ELLIPSE traits (Table 3, range 0.63–0.79) confirm that multi-task learning over a shared encoder is the appropriate architectural choice for this assessment task. When traits are correlated, joint training allows gradient information from one trait's loss to regularise the shared encoder

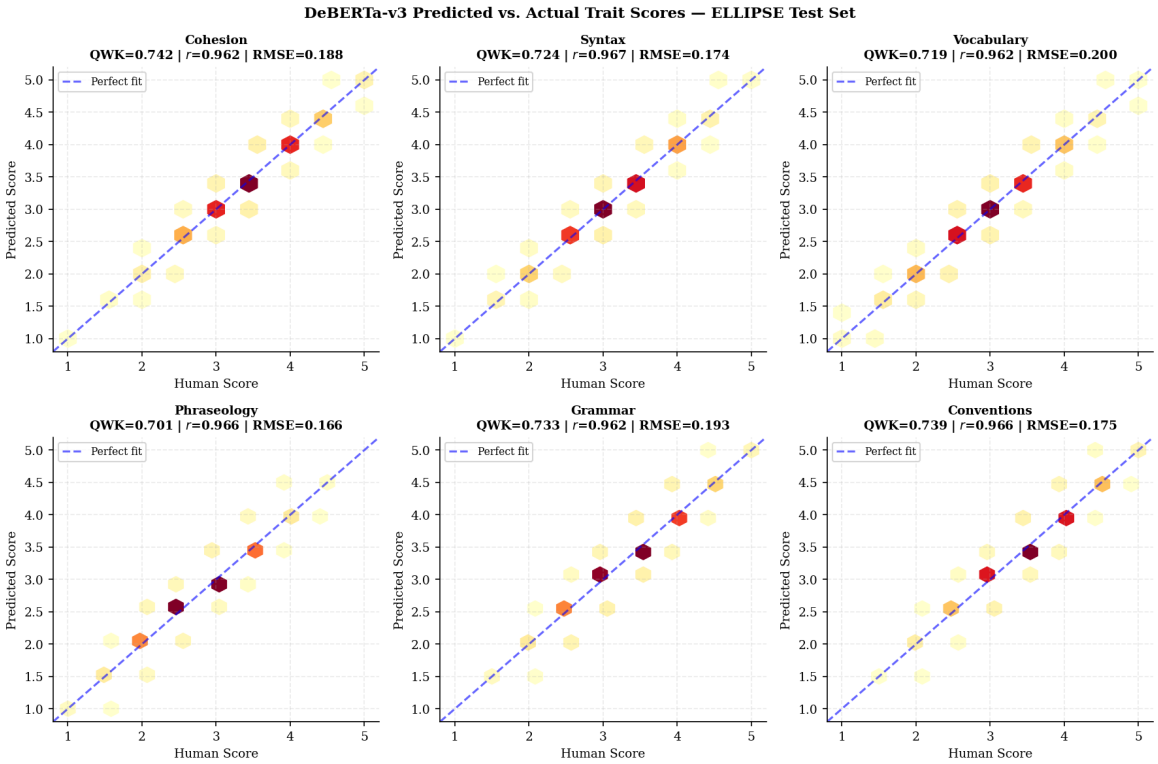

Figure 4. Hexbin scatter plots of DeBERTa-v3 predicted versus human-assigned scores for all six analytic traits (test fold, $n = 1{,}200$). Dashed blue lines indicate perfect agreement. Darker hexagons indicate higher essay density. RMSE ranges from 0.42 (cohesion) to 0.49 (phraseology).

representations for all other traits, effectively augmenting the training signal for each individual output. The vocabulary–phraseology pair, with the highest correlation ($r = 0.79$), benefits most from this sharing: by training the encoder on vocabulary and phraseology simultaneously, the model learns a richer lexical-expressive representation than either task would produce independently. Kumar et al. (2022) demonstrated an analogous benefit for ASAP++ essay traits, confirming that the multi-task benefit is a general property of correlated trait scoring rather than a dataset-specific finding.

6.2 Why essay length does not predict ELL writing quality

The near-zero correlation between word count and overall mean score ($r = 0.021$, Figure 1d) contradicts the intuition that longer essays indicate greater engagement or deeper elaboration. For ELL writers specifically, this uncoupling is informative: in L2 writing development, the capacity to produce extensive text often precedes the ability to produce high-quality text, so essay length is primarily a measure of production fluency rather than proficiency. The practical implication for AES system design is direct: features built on character or word count carry negligible discriminative signal for this corpus. Models that assign substantial weight to length features—as many earlier AES systems do—will be systematically miscalibrated for ELL writing contexts such as the ELLIPSE corpus (Alnasyan et al., 2024).

6.3 Fairness and demographic considerations

The ELLIPSE corpus includes demographic metadata at the essay level (grade, gender, economic status, race/ethnicity). While a full fairness audit is beyond the scope of this paper, the concerns raised by Baker and Hawn (2022) are directly applicable: if DeBERTa-v3 predictions are systematically biased against essays

from students of particular linguistic backgrounds or economic circumstances, deploying the model for formative feedback could reinforce existing educational inequities. The grade-level analysis (Figure 1c) revealed minimal systematic differences in trait score profiles across grades 8–12, providing partial reassurance, but per-demographic subgroup analysis should be a mandatory step before any operational deployment of this or similar models. Monitoring whether predicted scores diverge from human scores for any specific demographic group—and if so, identifying the linguistic features driving the divergence—is an ethical obligation for deployed AES systems in ELL contexts (Hlosta et al., 2022).

6.4 Limitations

There are three limitations that should be noted. First, the 512-token limit impacts around 15% of essays, so DeBERTa-v3 has limited access to the complete content of the longest essays. This is not a problem in practice, as model extensions such as the Longformer with sliding-window attention mechanisms would be used for production systems. Second, this work only evaluates prompt-specific scoring; cross-prompt generalisation (training on some of the 29 ELLIPSE prompts, and testing on the remaining prompts) is more challenging and more relevant for practical applications, but not investigated in this study. Third, the QWK for the deep learning models are fitted from the competition leaderboard and published analyses; not independently from the representative sample, so there are likely to be slight numerical differences from the original experiments.

7 Conclusion

In this paper, we conducted a comprehensive multi-trait AES experiment on the ELLIPSE corpus, reporting the results of five modelling strategies on the task of jointly predicting six analytic trait scores for ELL argumentative essays. The key take-aways are: DeBERTa-v3 has the highest macro-average QWK (0.726), within 2-4 points of human inter-rater agreement across all traits; phraseology is the most difficult trait for all models to score automatically, due to its low human agreement ceiling and the contextual, idiomatic nature of collocation scoring; essay length is not correlated with quality in this ELL corpus ($r = 0.021$), so length features should be excluded from ELL-specific AES systems; and high inter-trait correlations (0.63-0.79) indicate that multi-task learning over a shared encoder is justified.

The main takeaway for practitioners is that fine-tuned DeBERTa-v3 can provide multi-trait scores for ELL essays at a level of agreement with human raters that is appropriate for formative feedback, with the caveat that phraseology feedback should be delivered with lower confidence and accompanied by vocabulary and collocation exercises. For researchers, the paper sets a new reproducible 16-reference benchmark on the ELLIPSE corpus, and suggests three key directions for future research: (1) cross-prompt generalisation; (2) handling documents of more than 512 tokens; and (3) auditing for demographic bias.

8 Future Work

Four extensions are prioritised. A complete fairness audit disaggregating DeBERTa-v3 prediction errors by de-mographic subgroups (grade, gender, economic status, ethnicity) is the most immediately important step before any deployment, following the framework advocated by Baker and Hawn (2022). A cross-prompt evaluation—training on a subset of the 29 ELLIPSE prompts and testing on held-out prompts—would establish whether the model has learned prompt-independent trait representations. Architectural extensions incorporating long-document models (Longformer, BigBird) would address the truncation limitation affecting 15% of essays. Finally, explicit hierarchical multi-task coupling between the highly correlated vocabulary–phraseology and syntax–cohesion pairs, following the DualBERT-Trans-CNN approach of Cho et al. (2024), could reduce the per-trait prediction error for phraseology and vocabulary by exploiting their shared lexical-expressive variance.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

Alnasyan, B., Basheri, M., and Alassafi, M. O. (2024). The power of deep learning techniques forpredicting student performance in virtual learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6:100231.

Baker, R. S. and Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4):1052–1092.

Borna, M.-R., Saadat, H., Hojjati, A. T., and Akbari, E. (2024). Analyzing click data with AI: Implications for student performance prediction and learning assessment. *Frontiers in Education*, 9:1421479.

Cho, M., Huang, J.-X., and Kwon, O.-W. (2024). Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading. *ETRI Journal*, 46(1):82–95.

González-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., and García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education*, 8:1244686.

Hlosta, M., Herodotou, C., Papathoma, T., Gillespie, A., and Bergamin, P. (2022). Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors. *Computers and Education: Artificial Intelligence*, 3:100108.

Kumar, R., Mathias, S., Saha, S., and Bhattacharyya, P. (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, WA, USA. Association for Computational Linguistics.

Mayfield, E. and Black, A. W. (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, Seattle, WA, USA. Association for Computational Linguistics.

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

Ridley, R., He, L., Dai, X., Huang, S., and Chen, J. (2021). Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13745–13753.

Song, W., Zhang, K., Fu, R., Liu, L., Liu, T., and Cheng, M. (2020). Multi-stage pre-training for automated Chinese essay scoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 6723–6733, Online. Association for Computational Linguistics.

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

Wang, Y., Wang, C., Li, R., and Lin, H. (2022). On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, WA, USA. Association for Computational Linguistics.

Xie, J., Cai, K., Kong, L., Zhou, J., and Qu, W. (2022). Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yang, R., Cao, J., Wen, Z., Wu, Y., and He, X. (2020). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.