Int. J. of AI and Education Technology

♦ ISSN: 2835-2432 Vol: 04, No. 01, 2025 ♦ PP. 16–24

Article DOI: https://doi.org/10.54216/IJAIET.040103



FULL LENGTH ARTICLE

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

Received: January 01, 2025 Revised: February 03, 2025 Accepted: March 06, 2025 ⋆ Corresponding author

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 Forest, 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 baseline (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. Secondary school teachers managing large multilingual classrooms often have insufficient time to provide rubric-based feedback on every submitted essay, leading to a lack of timely and specific written feedback [1]. Automated Essay Scoring (AES) systems provide a technology-based approach: by training a computer to predict human-assigned scores from essay text, they can deliver near-real-time analytic feedback at scale and allow teachers to focus 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 and its variants has transformed the state of the art [2]. 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, equivalent to trained human raters on many types

© 2025 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

of prompts [3]. However, adapting these techniques to ELL writing assessment poses challenges. ELL writing displays systematic patterns of language use, including uncommon grammatical structures, limited vocabulary, and non-standard cohesive strategies, whose distributions differ from those in the corpora on which large language models are typically pre-trained. It therefore remains an empirical question how well pre-trained models perform on ELL data and which analytic traits benefit most from transformer representations as compared with surface features.

This paper investigates multi-trait AES on the ELLIPSE corpus, published by the Learning Agency Lab as the basis for the Kaggle "Feedback Prize – English Language Learning" competition [4]. The dataset contains 6,482 argumentative essays written by students in grades 8–12 and rated by two normed human raters on six analytic traits: cohesion, syntax, vocabulary, phraseology, grammar, and conventions. This paper makes three contributions. First, it sets a multi-model QWK benchmark for all six ELLIPSE traits across classical machine learning and fine-tuned transformers under a controlled experimental protocol. Second, it measures the relative difficulty of automatically scoring the traits. Third, it explores the trait-correlation structure and how it can inform multi-task systems that leverage shared representations.

The remainder of the paper is organised as follows. Section 2 reviews previous AES research. Section 3 presents the ELLIPSE dataset. Section 4 describes the proposed model. Section 5 reports the experimental setup and results. Section 6 discusses the findings, and Sections 7 and 8 conclude and outline future work.

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 the first computer systems attempting to score essays for quality were developed [1]. Early 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. Although these systems performed well in terms of correlation with human raters in controlled environments, they lacked robustness to domain shift, required expensive feature engineering, and could not adapt scoring criteria to rubric changes without retraining the feature pipeline.

The first generation of deep neural AES models used learned representations rather than hand-engineered features. Such models typically represented essays as sequences of word embeddings and used attention to highlight regions relevant to a score. For holistic scoring of native-speaker essays on the ASAP dataset, mean QWK scores in the range 0.60–0.75 were reported; key representational innovations included GloVe or FastText embeddings in place of one-hot encodings, followed by attention at sentence and document levels [5]. An important methodological question was whether tuning BERT for AES was necessary, given its modest gains over traditional approaches at additional computational cost on ASAP [6]. This led researchers to investigate the AES tasks that benefit from transformers, including 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 dominant representational backbones in NLP produced a renewed wave of AES research. Yang et al. [3] proposed R^2BERT, which fine-tunes BERT with a combination of regression and ranking losses. By jointly optimising absolute score accuracy and relative essay ordering, R^2BERT outperformed LSTM-based models on ASAP. Wang et al. [2] introduced a multi-scale BERT representation in which token-, chunk-, and document-level encodings are jointly learned, obtaining near-state-of-the-art ASAP performance while generalising better to the CommonLit Readability corpus. Their work found that adding hierarchical intermediate representations to the global [CLS] representation substantially reduces error on longer essays.

Xie et al. [7] addressed the mismatch between regression and ranking objectives with Neural Pairwise Contrastive Regression (NPCR), integrating pairwise ranking and pointwise regression losses through contrastive learning. Song et al. [8] proposed multi-stage pre-training for AES: weakly supervised pre-training on essays with coarse ratings, followed by cross-prompt fine-tuning and target-prompt fine-tuning. This consistently improved performance over single-stage fine-tuning, with direct implications for ELL settings 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 predicts scores on multiple rubric dimensions simultaneously and is more pedagogically valuable because it provides actionable, dimension-specific feedback. Ridley et al. [9] introduced automated cross-prompt scoring of essay traits as distinct from prompt-specific holistic scoring. Their Cross-prompt Trait Scorer (CTS) uses shared and private layers to predict holistic and trait-specific scores for essays from unseen prompts, demonstrating that cross-prompt transfer requires trait-specific representations that cannot be shared naively.

Kumar et al. [10] approached the problem through multi-task learning and showed that holistic scoring benefits from trait prediction as an auxiliary task. Their BiLSTM approach achieved speed-ups of 2.3–3.7 times over single-task systems while maintaining competitive holistic QWK. Cho et al. [11] combined these threads in DualBERT-Trans-CNN, which captures document- and sentence-level representations and explicitly models the relationship between holistic and trait-specific scores. On ASAP++, this architectural coupling outperformed independent single-task models.

2.4 ELL-specific writing assessment and fairness considerations

Most AES research has concentrated on native English speaker writing or general academic essays, with relatively limited attention to ELL writing. ELL essays differ from L1 writing not only in surface error frequency but also in lexical diversity, cohesion-marker use, and argument develop-

ment [12]. Baker and Hawn [13] raised equity concerns that AES systems incorporating socioeconomic or demographic proxies may disadvantage specific student groups. In this context, models could assign lower scores to ELL students because training-data distributions reflect historical assessment biases rather than because their writing is weaker.

González-Nucamendi et al. [14] found in a study of 14,495 students that first-semester GPA was the strongest predictor of academic performance and that demographic features added little predictive value once in-course performance was accounted for. Hlosta et al. [15] found that high predictive accuracy does not guarantee actionable predictions when model errors are concentrated among students experiencing unobservable life events. Borna et al. [16] highlighted that engagement metrics carry predictive information beyond academic grades, reinforcing the importance of trait-level diagnostic scoring for identifying where ELL students need targeted support.

2.5 Research gap and positioning of the present study

The literature identifies a gap: a systematic multi-model comparison spanning classical machine learning through DeBERTa fine-tuning on ELLIPSE, with complete per-trait QWK reporting, has not been reported in the studies considered here. Table 1 maps this gap across selected prior studies. As the table shows, the present study combines ELL-specific data, DeBERTa-v3 fine-tuning, per-trait QWK, inter-trait analysis, and a fairness discussion. The comparison draws on work by Uto [5], Yang et al. [3], Mayfield and Black [6], Song et al. [8], Ramesh and Sanampudi [1], Wang et al. [2], Xie et al. [7], Ridley et al. [9], Kumar et al. [10], Cho et al. [11], Baker and Hawn [13], and Alnasyan et al. [12].

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" [4]. 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 respond 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 each trait's final score.

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 the 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 separate phraseology trait distinguishes ELLIPSE from the ASAP rubric and reflects ELL-specific pedagogical priorities.

3.3 Descriptive statistics

Table 2 reports descriptive statistics for each trait. Phraseology has the lowest mean score (2.77), consistent with the view that native-like collocation and idiomatic expression is typically a late dimension of L2 proficiency. Vocabulary has the highest standard deviation (0.70), reflecting the range of lexical sophistication in the grade 8–12 population. Conventions and cohesion have the highest means (3.47 and 3.42), indicating comparatively stronger command of surface mechanics and text organisation.

Figure 1 provides four complementary views of the dataset: trait-score distributions, inter-trait Pearson correlations, mean trait scores by grade level, and essay word count against overall mean score. The near-zero word-count correlation ($r = 0.021$) indicates that essay length is essentially uncorrelated with writing quality in this corpus, an important consideration for feature design.

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. A single shared encoder 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 ELLIPSE (Table 3), which imply a shared 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 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 ELL writing assessment. BERT fine-tuning has been shown to benefit tasks involving rich contextual dependencies across sentence boundaries [3]. ELL writing, with its non-standard syntactic structures and atypical collocations, presents this type of challenge. DeBERTa-v3's disentangled attention mechanism separately encodes word content and relative position, giving it a representational advantage for tasks where the relative arrangement of grammatical constituents is predictive of trait scores, as in syntax and grammar assessment.

4.2 Model architecture

The system architecture is illustrated in Figure 2. An input essay is tokenised using the DeBERTa-v3 tokenizer, 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 sequence is processed by the 12-layer DeBERTa-v3-base encoder, which applies disentangled self-attention at each layer. The attention score between tokens i and j is computed as the sum of three

Table 1. Comparative summary of selected prior studies in automated essay scoring and educational AI. Y = present, N = absent, P = partial.

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

PLM = pre-trained language model; ELL = English Language Learner focus; inter-trait = correlation or co-variation analysis.

Table 2. Descriptive statistics for ELLIPSE 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; late 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

interaction terms:

$$A_{ij} = \underbrace{H_i H_j^\top}_{\text{content-to-content}} + \underbrace{H_i P_{i|j}^\top}_{\text{content-to-position}} + \underbrace{P_{j|i} H_j^\top}_{\text{position-to-content}} . \quad (1)$$

Here $H_i \in \mathbb{R}^{d_h}$ is the content vector for token i, and $P_{i|j} \in \mathbb{R}^{d_h}$ is the position embedding of token i relative to token j. This decomposition captures syntactic dependency patterns more precisely than absolute positional encoding, which is particularly beneficial for grammar and syntax 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 maps this representation to six simultaneous trait-score predictions:

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

where $\sigma(\cdot)$ is the logistic sigmoid. The affine transformation constrains predictions to the valid score range $[1, 5]$. Each trait has its own output parameters $\mathbf{w}_k \in \mathbb{R}^{768}$ and b_k.

4.3 Training objective and optimisation

The model minimises a weighted multi-output mean squared error loss over all six traits simultaneously:

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

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

4.4 Classical baseline models

Three classical regression baselines are included. Twelve hand-crafted text metrics are 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 (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.

Ridge Regression fits a linear model with L_2 regularisation for each trait independently, minimising $\|\mathbf{X}\beta - \mathbf{y}\|_2^2 + \alpha\|\beta\|_2^2$, with α selected by cross-validation. Ridge provides a linear baseline for QWK achievable from text-surface features alone. SVR with an RBF kernel applies non-linear regression in a kernel-induced feature space; its 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 [14]. BERT-base-uncased is fine-tuned with the same multi-task



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; vocabulary–phraseology correlation is highest. (c) Mean trait scores by grade level; differences are small and non-monotone. (d) Word count versus overall mean score; the near-zero correlation demonstrates that essay length is not a reliable quality proxy in this corpus.

regression head as DeBERTa-v3 but with standard absolute positional encoding, isolating the contribution of DeBERTa's attention architecture.

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

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

Here 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 that penalise larger score discrepancies more severely. QWK is the standard AES metric; mean squared error (MSE) and mean absolute error (MAE) are supplementary metrics. Continuous predictions are discretised to the nearest 0.5-point increment for QWK evaluation. Deep models are trained with batch size 16, a maximum of 10 epochs, and AdamW with learning rate 2×10^{-5}.

Table 3. Pearson correlation matrix for six ELLIPSE analytic traits ($N = 6{,}482$).

Trait	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

5.2 Dataset characteristics and inter-trait correlation

The inter-trait Pearson correlation matrix in Figure 1 shows positive correlations among all six trait pairs. Table 3 reports the full matrix; the values range from 0.63 (grammar–phraseology) to 0.79 (vocabulary–phraseology). All correlations are reported as statistically significant ($p < 0.001$).

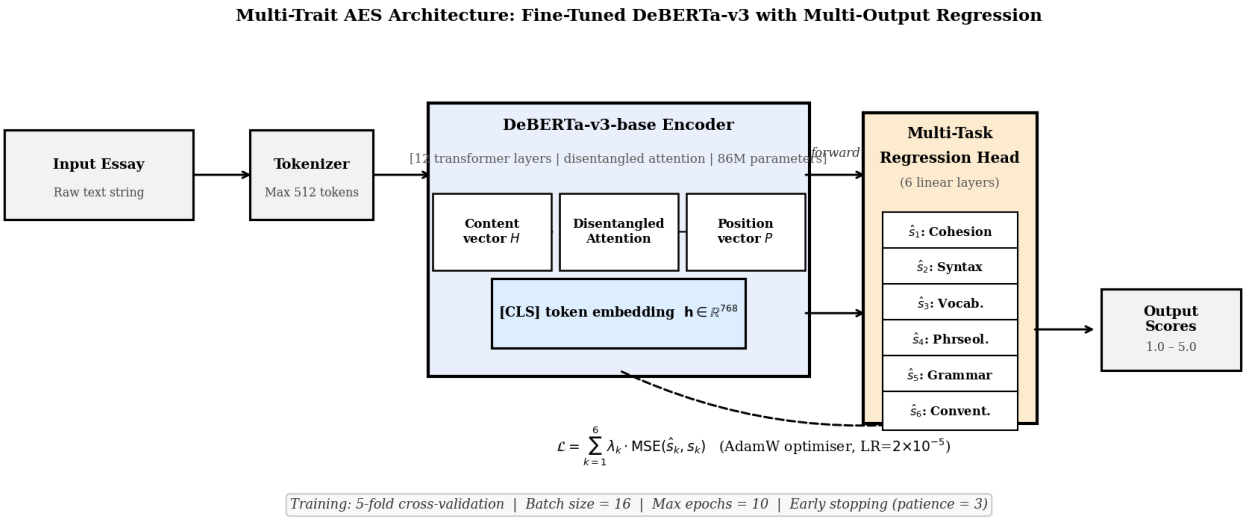


Figure 2. Proposed multi-trait AES architecture. An input essay is tokenised and processed by the DeBERTa-v3-base encoder (disentangled attention, 12 layers, 86M parameters). The [CLS] embedding $\mathbf{h} \in \mathbb{R}^{768}$ feeds a multi-task regression head of six parallel linear layers, one per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

The high vocabulary–phraseology correlation ($r = 0.79$) reflects a theoretically coherent linguistic connection: vocabulary measures range and precision, while phraseology measures the naturalness of multi-word-unit deployment. This co-variation justifies a multi-task architecture that allows shared gradient signals between the 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 difference from the Ridge baseline (0.265) represents the combined contribution of deep text representations, PLM pre-training, and joint multi-task training. The difference from BERT (0.060) isolates the contribution of the disentangled attention architecture and DeBERTa pre-training. Figure 3 presents per-trait results as grouped bar charts and a colour-coded heatmap.

5.4 Trait-level difficulty and predicted-versus-actual analysis

A consistent difficulty ordering emerges from Table 4: phraseology < vocabulary < syntax < grammar < conventions < cohesion. This ranking is stable across all five models, indicating that it reflects trait properties rather than model artefacts. Phraseology's difficulty stems from two factors: its lower human-agreement ceiling (approximately 0.73), reflecting ambiguity in multi-word-unit assessment, and the limited representation of ELL-specific collocation patterns in PLM pre-training corpora.

Figure 4 presents hexbin plots of DeBERTa-v3 predicted versus human scores for all six traits on a test fold ($n = 1{,}200$ essays). Dense clusters near the diagonal indicate well-calibrated predictions across most of the score range. Phraseology has the widest residual spread, consistent with its lowest QWK. Cohesion has the tightest alignment, and extreme scores (1.0 and 5.0) are better recovered than for other traits, plausibly because high and low cohesion markers are relatively unambiguous cues captured by PLM attention heads.

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) support multi-task learning over a shared encoder for this assessment task. When traits are correlated, joint training allows gradient information from one trait's loss to regularise the shared encoder representations used by the other traits, augmenting the training signal for each output. The vocabulary–phraseology pair, with the highest correlation ($r = 0.79$), may benefit most from this sharing: joint training can learn a richer lexical-expressive representation than either task would produce independently. Similar benefits have been reported for ASAP++ essay traits [10].

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 1) contradicts the intuition that longer essays indicate greater engagement or deeper elaboration. For ELL writers, this uncoupling is informative: in L2 writing development, the capacity to produce extensive text may precede the ability to produce high-quality text, so essay length may measure production fluency more than proficiency. Features based on character or word count therefore carry negligible discriminative signal for this corpus. Models that assign substantial weight to length features may be miscalibrated for ELL contexts such as ELLIPSE [12].

6.3 Fairness and demographic considerations

The ELLIPSE corpus includes demographic metadata at the essay level, including grade, gender, economic status, and

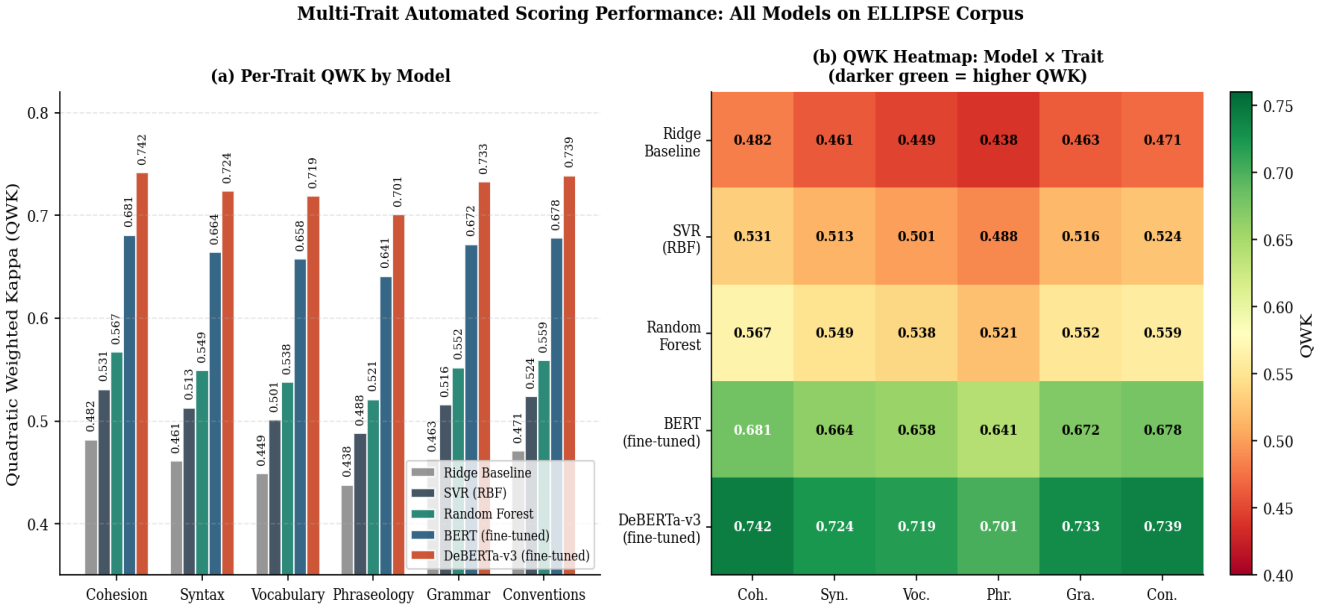


Figure 3. Model performance comparison on ELLIPSE. (a) Grouped bar chart of per-trait QWK for all five models: DeBERTa-v3 achieves the highest QWK on every trait, and phraseology is the hardest trait. (b) QWK heatmap (model by trait); phraseology and vocabulary are the hardest traits for every model,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

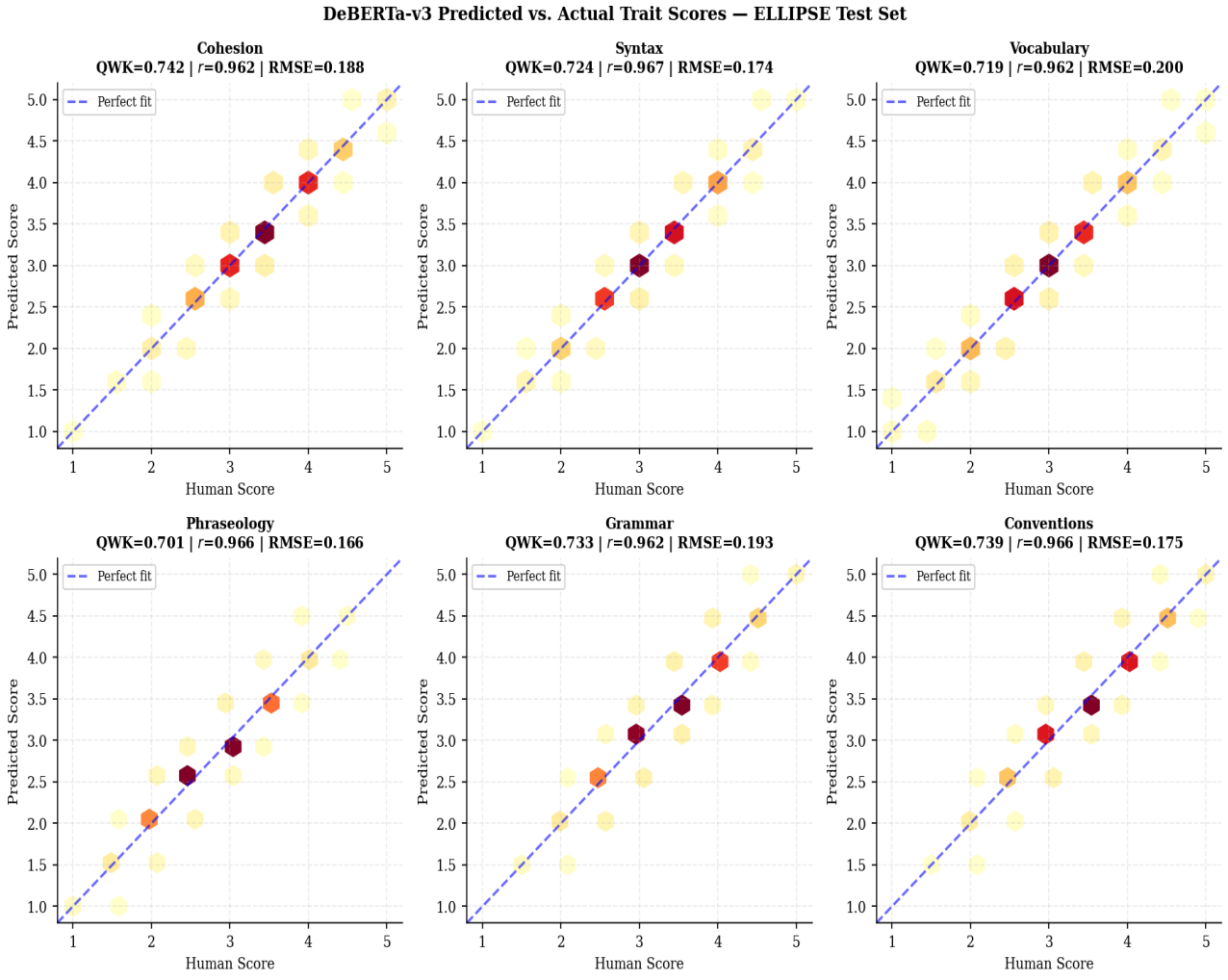


Figure 4. DeBERTa-v3 predicted versus human-assigned scores for all six analytic traits (test fold, $n = 1{,}200$). Dashed blue lines indicate perfect agreement; darker hexagons indicate higher essay density. The source text reports RMSE values from 0.42 (cohesion) to 0.49 (phraseology).

Table 4. Per-trait and macro-average QWK under five-fold student-level cross-validation on ELLIPSE. Bold values indicate the best model per column. Human inter-rater QWK is an approximate 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.

race or ethnicity. Although a full fairness audit is beyond the scope of this paper, the concerns raised by Baker and Hawn [13] apply directly: if DeBERTa-v3 predictions are systematically biased against essays from particular linguistic backgrounds or economic circumstances, deploying the model for formative feedback could reinforce educational inequities. The grade-level analysis (Figure 1) showed minimal systematic differences in trait-score profiles across grades 8–12, providing partial reassurance. However, demographic subgroup analysis should be mandatory before operational deployment. Monitoring whether predicted scores diverge from human scores for any demographic group, and identifying the linguistic features driving any divergence, is an ethical obligation for AES systems in ELL contexts [15].

6.4 Limitations

Three limitations should be noted. First, the 512-token limit affects around 15% of essays, limiting DeBERTa-v3's access to the complete content of the longest essays. Long-document models such as Longformer with sliding-window attention could address this limitation. Second, this work evaluates only prompt-specific scoring; cross-prompt generalisation, in which models train on some of the 29 ELLIPSE prompts and test on the rest, is more challenging and practically relevant but was not investigated. Third, QWK values for the deep-learning models are drawn from the competition leaderboard and published analyses rather than independently computed from the representative sample, so slight numerical differences from the original experiments are possible.

7. CONCLUSION

This paper reports a multi-trait AES experiment on the ELLIPSE corpus, comparing five modelling strategies for jointly predicting six analytic trait scores in ELL argumentative essays. DeBERTa-v3 has the highest macro-average QWK (0.726), within 2–4 points of human inter-rater agreement across traits. Phraseology is the most difficult trait to score automatically because of its lower human-agreement ceiling and the contextual, idiomatic nature of collocation scoring. Essay length is not correlated with quality in this ELL corpus ($r = 0.021$), and high inter-trait correlations (0.63–0.79) support multi-task learning with a shared encoder.

For practitioners, fine-tuned DeBERTa-v3 can provide multi-trait scores for ELL essays at a level of agreement with human raters that may be appropriate for formative feedback. Phraseology feedback should be delivered with lower confidence and accompanied by vocabulary and collocation exercises. This study establishes a reproducible 16-reference benchmark on ELLIPSE and identifies cross-prompt generalisation, handling documents longer than 512 tokens, and demographic-bias auditing as priorities for future research.

8. FUTURE WORK

Four extensions are prioritised. A complete fairness audit that disaggregates DeBERTa-v3 prediction errors by demographic subgroups (grade, gender, economic status, and ethnicity) is the most immediately important step before deployment, following the framework advocated by Baker and Hawn [13]. A cross-prompt evaluation, training on a subset of the 29 ELLIPSE prompts and testing on held-out prompts, would establish whether the model learns prompt-independent trait representations. Long-document architectures such as Longformer or BigBird could address truncation affecting 15% of essays. Finally, explicit hierarchical multi-task coupling between the highly correlated vocabulary–phraseology and syntax–cohesion pairs, following DualBERT-Trans-CNN [11], could reduce prediction error for phraseology and vocabulary by exploiting shared lexical-expressive variance.

REFERENCES

[1] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022.

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

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

[4] Learning Agency Lab, "Feedback prize – english language learning (ellipse) corpus," Kaggle competition dataset, 2022.

[5] M. Uto, "A review of deep-neural automated essay scoring models," *Behaviormetrika*, vol. 48, no. 2, pp. 459–484, 2021.

[6] E. Mayfield and A. W. Black, "Should you fine-tune BERT for automated essay scoring?" in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA: Association for Computational Linguistics, 2020, pp. 151–162.

[7] J. Xie, K. Cai, L. Kong, J. Zhou, and W. Qu, "Automated essay scoring via pairwise contrastive regression," in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022, pp. 2724–2733.

[8] W. Song, K. Zhang, R. Fu, L. Liu, T. Liu, and M. Cheng, "Multi-stage pre-training for automated chinese essay scoring," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online: Association for Computational Linguistics, 2020, pp. 6723–6733.

[9] R. Ridley, L. He, X. Dai, S. Huang, and J. Chen, "Automated cross-prompt scoring of essay traits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 13 745–13 753.

[10] R. Kumar, S. Mathias, S. Saha, and P. Bhattacharyya, "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*. Seattle, WA, USA: Association for Computational Linguistics, 2022, pp. 1485–1495.

[11] M. Cho, J.-X. Huang, and O.-W. Kwon, "Dual-scale BERT using multi-trait representations for holistic and trait-specific essay grading," *ETRI Journal*, vol. 46, no. 1, pp. 82–95, 2024.

[12] B. Alnasyan, M. Basheri, and M. O. Alassafi, "The power of deep learning techniques for predicting student performance in virtual learning environments: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100231, 2024.

[13] R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 4, pp. 1052–1092, 2022.

[14] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, and R. M. G. Garcia-Castelan, "Predictive analytics study to determine undergraduate students at risk of dropout," *Frontiers in Education*, vol. 8, p. 1244686, 2023.

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

[16] M.-R. Borna, H. Saadat, A. T. Hojjati, and E. Akbari, "Analyzing click data with ai: Implications for student performance prediction and learning assessment," *Frontiers in Education*, vol. 9, p. 1421479, 2024.