ISSN : 2455-135X

DeBERTaV3-Based Automated Essay Scoring with UnifiedQA-Generated Natural Language Justifications

Nagalla Hari Krishna, Bhosale Pravallika
Malineni Lakshmaiah Women's Engineering College, Guntur, AP, India

***Abstract*—Automated Essay Scoring (AES) systems are often accurate but difficult to explain in a way that teachers and students can understand. This paper presents a two-stage AES pipeline that combines strong score prediction with short natural-language justifications. First, a DeBERTaV3 encoder is fine-tuned for holistic score prediction using a regression-style scorer. The continuous outputs are then converted into the final discrete score bands using a fixed set of validation-derived cut points, keeping the inference stage deterministic. Second, to make outputs interpretable, we use a UnifiedQA (T5-based) model to generate concise justifications as answers to structured, rubric-like questions (e.g., strengths, needed improvements, prompt relevance), conditioned on the essay and prompt. This design keeps the scoring model unchanged while adding an explanation layer that supports qualitative inspection and reporting. Experiments on ASAP2 AES benchmark and evaluation using Quadratic Weighted Kappa (QWK) show strong agreement with human scores, achieving 0.8329 QWK on validation and 0.8151 QWK on the held-out test set, along with low mean absolute error.**

***Index Terms*—Automated Essay Scoring (AES), DeBERTaV3, UnifiedQA, natural language justification, interpretable scoring, Quadratic Weighted Kappa (QWK)**

I. Introduction

Automated Essay Scoring (AES) seeks to approximate human holistic judgments of writing quality at scale, enabling faster turnaround in educational assessment and supporting large-volume grading pipelines. However, contemporary AES deployment is constrained by two recurring requirements: (i) *high agreement* with human raters on an ordinal scoring scale, and (ii) *transparent, user-facing explanations* that help instructors and learners interpret a score as actionable feedback rather than an opaque decision [1], [2].

Neural AES has shifted from feature-engineering pipelines to end-to-end transformer encoders that better capture discourse-level semantics and long-range dependencies. Recent surveys emphasize that transformer-based scoring dominates current practice, particularly for prompt-conditioned holistic scoring and cross-prompt generalization settings [2]. In this family, DeBERTa-style encoders strengthen contextual representations through disentangled attention, and DeBERTaV3 further improves pretraining by combining ELECTRA-style objectives with gradient-disentangled embedding sharing, motivating its use as a strong backbone for regression-style AES scoring [3], [4]. Parallel to this, recent conference work continues to report strong AES performance with transformer variants that explicitly incorporate prompt/context information during scoring [5], [6].

Accuracy alone is insufficient for educational acceptance. Stakeholders increasingly expect explanations that communicate *why* a score was produced (e.g., prompt relevance, strengths, and concrete improvements). The literature highlights that producing feedback aligned with score rationale remains a central limitation of deep learning AES: many models achieve high agreement yet provide limited interpretability and weak feedback grounding [2]. At the same time, large language models (LLMs) are being explored for scoring and feedback, but empirical evidence shows reliability and consistency can vary by model and over time, raising concerns for stable assessment use [7], [8].

This work adopts a *separation-of-concerns* design: scoring is performed by a prompt-aware DeBERTaV3-small regressor, while explanation is handled by a distinct post-hoc generation module. The scoring model predicts a continuous value that is deterministically mapped to the required discrete ordinal score $\{1, \ldots, 6\}$ using fixed cut-points learned on validation data and frozen thereafter. Interpretability is added through UnifiedQA (T5-based), which generates short, structured justifications by answering rubric-like questions (e.g., two strengths, two improvements, and a binary prompt-relevance judgment) conditioned on the prompt and essay, without influencing the scorer [9], [10]. Experiments on an ASAP-style benchmark for source-based writing quality show strong agreement with human scores, achieving QWK = 0.8329 on validation and QWK = 0.8151 on the held-out test set, with low MAE.

A. Contributions

- A prompt-aware DeBERTaV3-small regression scorer for holistic AES on a six-level ordinal scale, motivated by recent transformer AES trends [4], [2].
- A deterministic discrete scoring layer using fixed cut-points learned on validation data, enabling stable, reproducible score mapping during evaluation and deployment.
- A post-hoc interpretability layer using UnifiedQA to generate short, structured natural-language justifications aligned to prompt and essay content, explicitly separated from the scoring path [9], [10].
- Empirical evaluation with QWK/MAE/Accuracy and confusion-matrix analysis demonstrating strong ordinal agreement and boundary-sensitive error behavior on validation and test splits.

The remainder of this paper is organized as follows. Section II reviews related work on AES and interpretability.

ISSN : 2455-135X

Section III describes the proposed methodology. Section IV details the experimental setup and evaluation protocol. Section V presents results and analysis. Section VI concludes with implications and directions for future research.

II. Related Work

AES research has evolved from feature-engineered pipelines to neural scoring models and, more recently, to Transformer-based systems that learn task-conditioned representations directly from essay text. Contemporary reviews emphasize that progress in holistic scoring has been accompanied by renewed attention to (i) prompt-aware modeling, (ii) generalization beyond seen prompts and datasets, and (iii) transparency and feedback quality needed for educational use [2], [11].

A. Neural and Transformer Encoders for Holistic AES

Early neural AES systems established that end-to-end text encoders can learn essay-to-score mappings without extensive manual feature design, typically using regression objectives for holistic scoring [12]. Transformer encoders later became dominant due to stronger contextual modeling and transfer capabilities [2], [11]. Within this family, DeBERTa-style architectures improve representation quality by disentangling content and positional information in attention, and DeBERTaV3 further refines pretraining via ELECTRA-style objectives and gradient-disentangled embedding sharing [3], [4]. Recent conference studies continue to report Transformer-based scoring variants that explicitly model context and prompt information to improve AES performance [5], [6].

Beyond encoder strength, several strands of work incorporate structured linguistic signals to improve scoring behavior. For example, discourse-oriented external knowledge and discourse features have been used to complement Transformer representations and improve scoring agreement on standard benchmarks [13]. In parallel, multi-trait scoring has been explored as a way to align automated scoring with rubric dimensions; autoregressive score generation with encoder–decoder models (e.g., T5-style decoding of trait scores) offers a unified approach to predict multiple trait scores within a single model [14].

B. Prompt Generalization, Dataset Limitations, and Trait Supervision

A persistent limitation in AES evaluation is over-reliance on a small number of benchmarks and prompts. Recent work highlights that strong in-domain performance does not necessarily imply robust cross-prompt generalization, and it questions whether increasingly complex neural architectures are always required for cross-prompt AES [15]. To broaden evaluation beyond ASAP-only settings and enable trait-level analysis, new annotated corpora have been introduced with both holistic and trait-specific scores; ICLE++ is a representative example designed to facilitate evaluation of generalizability, multi-trait scoring, and cross-prompt scoring [16].

C. Interpretability, Rationale Alignment, and LLM-Based Scoring/Feedback

Interpretability remains a central adoption barrier for deep AES models: high agreement with human scores does not guarantee that a model relies on rubric-consistent evidence or provides actionable feedback [2]. Recent diagnostic work studies *rationale alignment* by applying linguistically-informed counterfactual interventions and comparing how AES encoders and LLMs respond; findings suggest that BERT-like models often emphasize sentence-level cues, while LLMs show sensitivity to broader rubric-relevant properties such as organization and conventions [17].

LLMs have also been used for direct essay scoring via prompting. A recent zero-shot framework decomposes holistic proficiency into traits and prompts LLMs to score each trait before aggregating to an overall score, reporting strong gains over naive prompting on benchmarks [18]. At the same time, empirical work in educational settings reports that ChatGPT-style scoring can diverge from expert human ratings and exhibit instability across time, motivating designs that separate stable discriminative scoring from post-hoc explanation generation [7], [8]. Fairness has also become a visible concern: recent comparative studies assess subgroup behavior and bias patterns across AES algorithm families, arguing that accuracy alone is insufficient for equitable assessment [19].

D. QA-Style Justification Generation

A practical approach to generate concise, structured justifications is to cast feedback generation as question answering or instruction-following. UnifiedQA demonstrates strong cross-format QA generalization with a single text-to-text model, and T5 provides the underlying text-to-text paradigm that enables consistent prompting for short outputs [9], [10]. This motivates post-hoc justification designs in AES where the scorer produces the numeric decision and a separate QA-style generator produces brief rationale statements aligned with predefined questions.

III. Methodology

This work addresses holistic Automated Essay Scoring (AES) with two outputs per essay: (i) a discrete ordinal score on a six-level scale $\{1, \ldots, 6\}$ and (ii) a short natural-language justification suitable for user-facing feedback. The proposed system is a two-module pipeline with an explicit separation between scoring and explanation. A prompt-aware DeBERTaV3-small regressor produces a continuous score estimate, which is deterministically mapped to the discrete score set using *fixed cut-points learned on validation data*. A separate UnifiedQA module generates brief post-hoc feedback statements conditioned on the prompt–essay context; it does not influence the scoring decision [4], [5], [6], [9], [10].

A. Architecture and Data Flow

Let $\mathbf{D} = \{(P_i, E_i, y_i)\}_{i=1}^{N}$, where P_i is the prompt/rubric text, E_i is the student essay, and $y_i \in \{1, \ldots, 6\}$ is the human holistic score. Fig. 1 summarizes the end-to-end flow: prompt-aware input construction, continuous scoring, deterministic discrete mapping, and post-hoc justification generation.

ISSN : 2455-135X

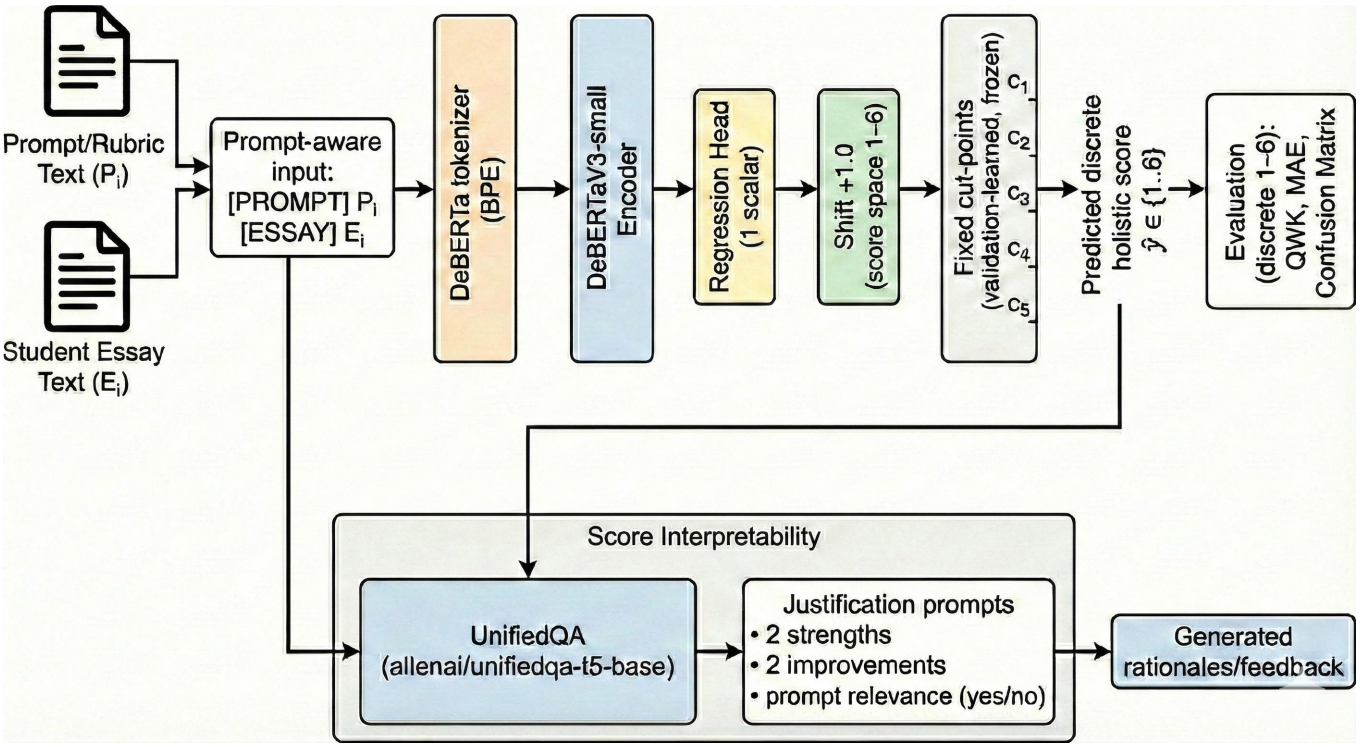

Fig. 1: Proposed AES pipeline. The score predictor is a prompt-aware DeBERTaV3-small regressor producing a continuous score, followed by deterministic discrete mapping using fixed validation-learned cut-points. UnifiedQA is used only for post-hoc interpretability.

B. Prompt-Aware Score Prediction

a) Input construction.: Each instance is formed as a single prompt-aware sequence:

$$X_i = \texttt{[PROMPT]}\ P_i\ \texttt{[ESSAY]}\ E_i. \tag{1}$$

b) DeBERTaV3-small regressor.: A DeBERTaV3-small encoder [4] produces a pooled representation $\mathbf{h}_i$, which is passed to a one-dimensional regression head:

$$r_i = g_\theta(X_i) \in \mathbb{R}. \tag{2}$$

To align the regressor output with the 1–6 score space, we apply a constant shift:

$$s_i = r_i + 1.0. \tag{3}$$

C. Discrete Score Mapping via Fixed Cut-Points

Continuous predictions s_i are converted to discrete labels $\hat{y}_i \in \{1, \ldots, 6\}$ using *fixed cut-points learned on validation data* and frozen thereafter:

$$\begin{aligned} c_1 &= 1.6069442348049219, \\ c_2 &= 2.5339883438733097, \\ c_3 &= 3.323994310899762, \\ c_4 &= 4.249251710005174, \\ c_5 &= 5.279343944917446. \end{aligned} \tag{4}$$

$$\hat{y}_i = \begin{cases} 1, & s_i < c_1, \\ 2, & c_1 \le s_i < c_2, \\ 3, & c_2 \le s_i < c_3, \\ 4, & c_3 \le s_i < c_4, \\ 5, & c_4 \le s_i < c_5, \\ 6, & s_i \ge c_5. \end{cases} \tag{5}$$

D. Post-hoc Justification Generation

A separate text-to-text model generates short natural-language justifications after scoring. We use UnifiedQA (allenai/unifiedqa-t5-base) [9], built on T5 [10]. Given the prompt–essay context (and optionally the predicted score as additional context), UnifiedQA answers structured rubric-style questions, producing concise feedback statements. This module is *not* used to refine or modify the score predictor.

a) Justification prompts.: For each (P_i, E_i), we issue rubric-style prompts such as:

- "Give two strengths of the essay in short phrases."
- "Give two improvements in short phrases."
- "Is the essay relevant to the prompt? Answer yes or no."

Optionally, the predicted discrete score can be appended as context (e.g., "Predicted score: $\hat{y}_i$") to encourage score-consistent phrasing, while maintaining a strict separation between scoring and explanation.

ISSN : 2455-135X

TABLE I: Evaluation splits used in this study.

Split	Size (N)	Use
Validation	2472	Model selection + cut-point learning
Test	2475	Final evaluation

TABLE II: Training hyperparameters and selection criteria for the DeBERTaV3-small scoring model.

Parameter	Value
Optimizer	`adamw_torch`
Learning rate	1×10^{-5}
LR scheduler	Linear
Warmup steps / ratio	0 / 0.0
Train batch size (per device)	4
Eval batch size (per device)	8
Gradient accumulation steps	1
Epochs	4
Weight decay	0.01
Gradient clipping (∇ |max)	1.0
Random seed	42
Selection metric	QWK (↑)

TABLE III: Hardware and software environment.

Item	Value
OS / Platform	Windows 10 (10.0.26100)
Python	3.10.19
GPU	NVIDIA GeForce GTX 1080 Ti
GPU VRAM	11.0 GB
System RAM	31.73 GB
PyTorch	2.2.2+cu118
CUDA (Torch)	11.8
cuDNN	8700
Transformers	4.41.2
Datasets	4.5.0
Scikit-learn	1.7.2
NumPy	1.26.4
Mixed precision	FP16 enabled; BF16 disabled

IV. Experimental Setup

A. Dataset and Evaluation Splits

Experiments are conducted on the ASAP 2.0 corpus, a large-scale dataset of source-based argumentative essays for AES research [21]. We use the associated public competition release [22], [23]. Performance is reported on a fixed validation split ($N_{\text{val}} = 2472$) and a held-out test split ($N_{\text{test}} = 2475$). Gold labels are discrete integers $y \in \{1, \ldots, 6\}$.

B. Implementation, Training, and Reproducibility Details

All experiments were implemented in Python using the HuggingFace `transformers` training stack and executed on a single GPU. To ensure reproducibility, we report the complete optimizer and training configuration, random seed, mixed-precision setting, library versions, and the exact best-model checkpoint used for final evaluation. Discrete predictions are produced by applying the fixed score boundaries learned on the validation set (Sec. III-C), which are then frozen for all reported results.

1) Scorer Training Configuration: The DeBERTaV3-small scorer is fine-tuned as a regression model (`problem_type = regression`). Training uses AdamW (`adamw_torch`) with a linear learning-rate schedule and no warmup. Model selection is performed using validation QWK (higher is better). The best validation QWK achieved during training is 0.8211, and the best checkpoint is `checkpoint-19784`, which is used to produce the reported test results.

Loss function. The training logs did not store an explicit loss identifier. The scorer is trained under a regression objective consistent with `problem_type = regression`.

2) Compute Environment and Software Versions: All runs were executed on Windows with CUDA-enabled PyTorch. Mixed precision (FP16) was enabled; BF16 was disabled. Table III summarizes the compute environment and key library versions.

3) UnifiedQA Configuration for Justifications: The justification generator uses `allenai/unifiedqa-t5-base`. Decoding settings were not stored in the run report; therefore, we only report the checkpoint identity and treat justification outputs as qualitative artifacts (Sec. III-D), not as a scored component of the evaluation.

C. Evaluation Metrics

We evaluate discrete predictions on validation and test using Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), Accuracy, and the 6×6 confusion matrix.

a) Quadratic Weighted Kappa (QWK).: QWK measures agreement on an ordinal scale while penalizing larger disagreements more strongly [24], [25]. Let $K = 6$ and define

$$w_{ij} = \frac{(i-j)^2}{(K-1)^2}, \quad i, j \in \{1, \ldots, K\}. \tag{6}$$

With observed confusion matrix $O \in \mathbb{R}^{K \times K}$ and expected matrix E computed from the marginals of O,

$$\text{QWK} = 1 - \frac{\sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} O_{ij}}{\sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} E_{ij}}. \tag{7}$$

b) Mean Absolute Error (MAE).:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|. \tag{8}$$

c) Accuracy.:

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}[y = \hat{y}]. \tag{9}$$

V. Results and Discussion

A. Overall Discrete Holistic Scoring Performance

Table IV reports performance after converting continuous predictions to discrete holistic scores using the fixed score boundaries defined in Sec. III-C. The proposed system achieves strong agreement with human scores on both splits, with QWK of 0.8329 on validation and 0.8151 on the held-out test set. The MAE values (0.3483 and 0.3786) indicate low average deviation from reference scores, while exact-match accuracy remains competitive for a six-level ordinal scale.

ISSN : 2455-135X

TABLE IV: Discrete holistic scoring performance (score range 1–6). Higher is better for QWK and Accuracy; lower is better for MAE.

Split	QWK ↑	MAE ↓	Accuracy ↑
Validation (n = 2472)	0.8329	0.3483	0.6679
Test (n = 2475)	0.8151	0.3786	0.6372

TABLE V: Baseline comparison using test-set QWK.

Model	Test QWK ↑
LSTM-AES	0.68
BERT-AES	0.74
RoBERTa-AES	0.75
DeBERTaV3-small + fixed score boundaries	**0.8151**

B. Comparison with Baseline Models

Table V contextualizes the proposed method against common neural AES baselines. The LSTM model represents early neural sequence scoring [12]. BERT and RoBERTa represent Transformer encoder fine-tuning for essay-to-score regression [26], [27]. UnifiedQA is not used for scoring and is applied only as a post-hoc interpretability layer (Sec. III-D).

The proposed prompt-conditioned DeBERTaV3-small regressor with fixed score boundaries achieves the best reported test QWK (0.8151). Relative to the strongest encoder-only baseline (RoBERTa, 0.75), the absolute gain is +0.0651 QWK, indicating improved ordinal agreement under the same evaluation protocol.

C. Confusion Matrix and Ordinal Error Structure

Table VI shows the test-set confusion matrix (rows: true scores; columns: predicted scores). Misclassifications are dominated by adjacent-level confusions, which is expected in ordinal scoring. The largest error mass appears near the middle of the scale, especially between 3↔2 and 3↔4, and between 4↔5, consistent with boundary-sensitive ambiguity in mid-range essays.

D. Interpretability Outputs (Post-hoc Justifications)

The justification module is assessed qualitatively as an interpretability layer rather than a scoring component. Following the protocol in Sec. IV, we generate post-hoc rationales for a balanced subset of 1642 validation essays (821 correct and 821 incorrect predictions). These outputs provide a compact inspection interface for prompt relevance and feedback consistency and can be displayed alongside the predicted discrete score without modifying the scoring path.

VI. Conclusion and Future Work

This paper presented a two-module AES pipeline that outputs (i) a discrete holistic score on a 1–6 ordinal scale and (ii) a short natural-language justification. The scoring module fine-tunes DeBERTaV3-small as a prompt-conditioned regressor by encoding the prompt and essay as a single sequence. Discrete scores are produced by applying fixed score boundaries learned on validation data and then frozen for evaluation. On the held-out test set (n = 2475), the system achieved QWK = 0.8151, MAE = 0.3786, and Accuracy = 0.6372, and it improved over the baseline transformer scoring results reported in the earlier draft. Error analysis via the confusion matrix showed that most disagreements occur between adjacent score levels, indicating that ordinal structure is largely preserved while ambiguity concentrates near neighboring score boundaries.

To support interpretability without altering the scoring path, a separate UnifiedQA-based justification module generated brief prompt-aligned statements (strengths, improvements, and relevance). This separation keeps the numeric scorer stable while providing an inspection and feedback surface that can be reviewed by instructors or used to accompany predictions in user-facing settings.

Several directions follow from current AES trends and the observed error structure. First, boundary-sensitive errors motivate learning strategies that explicitly optimize ordinal consistency and robustness near class transitions, including score-aware training objectives and boundary-focused data augmentation. Second, extending the pipeline to cross-prompt and low-resource settings requires prompt-invariant modeling and stronger generalization controls, as emphasized by recent work on scoring-invariance and cross-prompt trait scoring [28]. Third, fairness and stability analyses should be incorporated as standard reporting: prior studies show that AES systems can exhibit subgroup sensitivity and variability that must be audited under realistic educational constraints [19], [20]. Fourth, hybrid systems that combine a discriminative scorer with trait-based or rubric-driven LLM scoring remain a promising direction for improving transparency and aligning model outputs with human grading constructs [18]. Finally, justification quality should be evaluated with human-centered protocols that measure actionability, consistency, and equity of feedback across learner groups and genres, consistent with recent empirical investigations of LLM-based writing feedback [29].

References

[1] B. Beigman Klebanov and N. Madnani, eds., *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Cham, Switzerland: Springer, 2022.

[2] H. Misgna, B.-W. On, I. Lee, and G. S. Choi, "A survey on deep learning-based automated essay scoring and feedback generation," *Artificial Intelligence Review*, vol. 58, art. no. 36, 2025, doi: 10.1007/s10462-024-11017-5.

[3] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-Enhanced BERT with Disentangled Attention," arXiv preprint arXiv:2006.03654, 2020.

[4] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," arXiv preprint arXiv:2111.09543, 2021.

[5] R. K. R. Chavva, S. R. Muthyam, M. S. Seelam, and N. Nalliboina, "A Transformer-Based Approach for Enhancing Automated Essay Scoring," in *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, Aug. 2024, pp. 1–6, doi: 10.1109/ACET61898.2024.10730000.

[6] C. R. K. Reddy, A. K. Tulasi, M. Maturi, and A. Nagam, "Context-Aware Automated Essay Scoring with MLM-Pretrained T5 Transformer," in *2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA)*, Jun. 2025, pp. 1439–1443, doi: 10.1109/ICIRCA65293.2025.11089875.

[7] A. Pack, A. Barrett, and J. Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," *Computers and Education: Artificial Intelligence*, vol. 6, art. no. 100234, 2024, doi: 10.1016/j.caeai.2024.100234.

ISSN : 2455-135X

TABLE VI: Test-set confusion matrix (rows=true, cols=pred), classes 1–6.

True \ Pred	1	2	3	4	5	6
1	89	82	4	0	1	0
2	54	486	136	10	0	0
3	7	175	518	197	6	0
4	0	5	75	392	84	0
5	0	0	3	41	84	7
6	0	0	0	1	10	8

[8] N. M. Bui and J. S. Barrot, "ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring," *Education and Information Technologies*, vol. 30, pp. 2041–2058, 2025, doi: 10.1007/s10639-024-12891-w.

[9] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, and H. Hajishirzi, "UnifiedQA: Crossing Format Boundaries with a Single QA System," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[11] J. C. Li and H. T. Ng, "Automated Essay Scoring: Recent Successes and Future Directions," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[12] K. Taghipour and H. T. Ng, "A Neural Approach to Automated Essay Scoring," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[13] N. Ait Khayi and V. Rus, "Automated Essay Scoring Using Discourse External Knowledge," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 7154–7160, doi: 10.24963/ijcai.2024/791.

[14] H. Do, Y. Kim, and G. Lee, "Autoregressive Score Generation for Multi-trait Essay Scoring," in *Findings of the Association for Computational Linguistics: EACL 2024*, 2024, pp. 1659–1666.

[15] S. Li and V. Ng, "Conundrums in Cross-Prompt Automated Essay Scoring: Making Sense of the State of the Art," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024, pp. 7661–7681, doi: 10.18653/v1/2024.acl-long.414.

[16] S. Li and V. Ng, "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 (NAACL 2024)*, 2024, pp. 8465–8486, doi: 10.18653/v1/2024.naacl-long.468.

[17] Y. Wang, R. Hu, and Z. Zhao, "Beyond Agreement: Diagnosing the Rationale Alignment of Automated Essay Scoring Methods based on Linguistically-informed Counterfactuals," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 8906–8925, doi: 10.18653/v1/2024.findings-emnlp.520.

[18] S. Lee, Y. Cai, D. Meng, Z. Wang, and Y. Wu, "Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, FL, USA, 2024, pp. 181–198, doi: 10.18653/v1/2024.findings-emnlp.10.

[19] N.-J. Schaller, Y. Ding, A. Horbach, J. Meyer, and T. Jansen, "Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays from Secondary Education," in *Proc. 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico, 2024, pp. 210–221. [Online]. Available: https://aclanthology.org/2024.bea-1.18/

[20] F. Garc´ıa-Varela, M. Nussbaum, M. Mendoza, C. Mart´ınez-Troncoso, and Z. Bekerman, "ChatGPT as a Stable and Fair Tool for Automated Essay Scoring," *Education Sciences*, vol. 15, no. 8, Art. no. 946, 2025, doi: 10.3390/educsci15080946.

[21] S. A. Crossley, P. Baffour, L. Burleigh, and J. King, "A large-scale corpus for assessing source-based writing quality: ASAP 2.0," *Assessing Writing*, vol. 65, Art. no. 100954, Jul. 2025, doi: 10.1016/j.asw.2025.100954.

[22] The Learning Agency Lab, "ASAP 2.0 Dataset," 2024. [Online]. Available: https://the-learning-agency-lab.com/learning-exchange/asap-2-0-dataset/ (accessed Feb. 06, 2026).

[23] LEAR Lab, "Datasets: The Learning Agency Lab – Automated Essay Scoring 2.0," [Online]. Available: https://learlab.org/data/ (accessed Feb. 06, 2026).

[24] A. Doewes, N. Kurdhi, and A. Saxena, "Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring," in *Proc. 16th Int. Conf. Educational Data Mining (EDM)*, 2023, pp. 103–113. [Online]. Available: https://educationaldatamining.org/EDM2023/proceedings/2023.EDM-long-papers.9/2023.EDM-long-papers.9.pdf

[25] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, Oct. 1968, doi: 10.1037/h0026256.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[27] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692*, 2019.

[28] J. Wang and S. Yu, "Improving Prompt Generalization for Cross-prompt Essay Trait Scoring from the Scoring-invariance Perspective," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, 2025, pp. 2633–2646, doi: 10.18653/v1/2025.findings-emnlp.142.

[29] M. Jovic, S. Papakonstantinidis, and R. Kirkpatrick, "From red ink to algorithms: investigating the use of large language models in academic writing feedback," *Language Testing in Asia*, vol. 15, Art. no. 59, 2025, doi: 10.1186/s40468-025-00389-2.