Interpretable Automated Essay Scoring via Quantitative Bipolar Argumentation

Xiang Yin[1,*]

[1]Imperial College London, Exhibition Road, London, SW7 2AZ, UK

Abstract

Automated Essay Scoring (AES) has achieved strong predictive performance with modern neural and large language models, but these approaches are typically opaque and provide limited insight into how scores are derived. In this paper, we propose an interpretable AES framework based on Quantitative Bipolar Argumentation Frameworks (QBAFs), in which each essay is modelled as an argumentative structure composed of discourse units, and the support and attack relations between them. To capture complementary information beyond discourse structure, we further introduce feature-derived arguments into the QBAF, representing handcrafted essay-level statistics such as essay length and discourse-type diversity. Experiments show that the proposed approach achieves competitive performance compared with some standard baselines, while also providing structured and transparent explanations for the assigned scores. Overall, our results suggest that QBAFs provide a promising formalism for developing transparent and interpretable AES.

Keywords

Explainable AI, Quantitative Argumentation, Automated Essay Scoring

1. Introduction

Automated Essay Scoring (AES) [1] aims to automatically assess the quality of student essays and assign scores. Over the past decades, AES has typically been formulated as a supervised learning problem, with recent approaches predominantly based on neural models [2], including transformer-based and large language model (LLM) methods [3]. These approaches either directly encode the full essay text or leverage engineered features to predict human-annotated scores, achieving high agreement with ground-truth holistic scores on standard benchmark datasets.

Despite these advances, most state-of-the-art AES systems remain fundamentally black-box models. While modern AES systems can score essays with high predictive accuracy, the main remaining limitation is that they provide limited insight into how a particular score is derived and little actionable feedback on how an essay could be improved. As a result, teachers may find it difficult to trust the system's decisions, and students receive limited constructive feedback from the assigned scores. This limitation has been increasingly recognized in the literature [3, 4], highlighting the need for more transparent and explainable AES systems. Furthermore, from a regulatory perspective, AI systems used to evaluate learning outcomes are classified as high-risk under the EU AI Act (Annex III) [5], reflecting their potential impact on learners' educational and career opportunities.

To address these limitations, we revisit this problem from an argumentation perspective. A substantial body of work has shown that argumentation frameworks are well-suited to support explainability by exposing transparent reasoning processes and providing structured explanations (e.g. see [6] for an overview). Moreover, this perspective naturally aligns with essay scoring, as essays typically exhibit explicit argumentative structures, such as claims, counterclaims, and supporting evidence.

As a motivating example, Figure 1 illustrates a hierarchical quantitative bipolar argumentation framework (QBAF) [7] constructed from an essay, where each argument node corresponds to a discourse unit. The `Position` node at the top represents the overall stance, supported by `Claim` nodes and attacked by a `Counterclaim`, which are in turn supported and attacked by the `Evidence` and `Rebuttal` nodes,

The 26th International Workshop on Computational Models of Natural Argument (CMNA'26)

*Corresponding author.

xy620@ic.ac.uk (X. Yin)

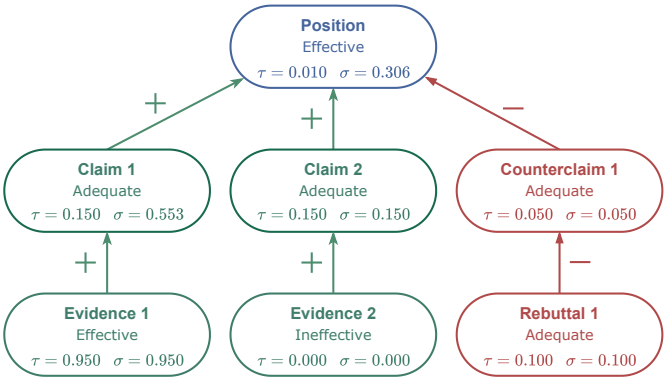

Figure 1: An example QBAF for Automated Essay Scoring (AES). Nodes denote argumentative discourse units extracted from an essay, while edges labelled "+" and "-" represent *support* and *attack*, respectively. Here, τ denotes the base score of each argument, reflecting its initial unit writing quality, and σ denotes its final QE strength after reasoning. Further details of QBAF construction are provided in Section 4.

respectively. Each node is assigned a *base score* based on its discourse type and content effectiveness label from the dataset (i.e., `Effective`, `Adequate`, or `Ineffective`) [1]. We then apply Quadratic Energy (QE) semantics [8] to evaluate the *strength* of the `Position`, which serves as the predicted essay score. In this work, we assume that essays are already segmented into discourse units with discourse types, effectiveness labels, and hierarchical parent relations; deriving these annotations from raw text is left to future work.

The main contributions of this paper are threefold.

- We propose an interpretable QBAF-based AES system grounded in the argumentative structure of essays (Section 4.2);
- We extend the QBAF with interpretable handcrafted features to capture complementary statistical information (Section 4.3);
- We empirically demonstrate that the proposed approach achieves competitive performance while providing structured explanations (Section 5).

The code is available at: https://github.com/XiangYin2021/QBAF4EDU.

2. Related Work

AES has evolved from heuristic-based systems to feature-engineered machine learning models, and subsequently to deep neural approaches and, more recently, LLM-based methods [3]. Early heuristic approaches relied on manually designed rules and surface-level features (e.g., length and grammar), offering a degree of transparency but failing to capture deeper aspects such as coherence and argument quality. With the availability of annotated datasets, machine learning models were able to leverage a broader range of linguistic and semantic features, improving predictive performance. However, these approaches remained dependent on feature engineering and often struggled to generalize across prompts. Deep learning methods further advanced AES by learning distributed representations using architectures such as CNNs, RNNs, and Transformers (e.g., BERT [9]), achieving strong performance on benchmark datasets. Nevertheless, these models are typically opaque and difficult to interpret, raising concerns about whether they genuinely assess writing quality or rely on superficial or latent signals that are hard to explain. More recently, LLMs have enabled zero-shot and few-shot essay scoring via prompting [10], showing promise in capturing higher-level content. However, they remain sensitive to prompt design and often underperform fine-tuned models, leaving issues of reliability and interpretability unresolved.

[1] QBAF construction details are provided in Section 4.

Overall, despite substantial progress in predictive performance, most AES approaches treat scoring as a black-box prediction task. In contrast, our work emphasizes interpretability by making the underlying evaluation process transparent and understandable.

3. Preliminaries

A *QBAF* is a quadruple $\langle \mathscr{A}, \mathscr{R}^-, \mathscr{R}^+, \tau \rangle$ where $\mathscr{A}$ is a finite set of *arguments*; $\mathscr{R}^- \subseteq \mathscr{A} \times \mathscr{A}$ is a binary *attack* relation; $\mathscr{R}^+ \subseteq \mathscr{A} \times \mathscr{A}$ is a binary *support* relation; $\mathscr{R}^- \cap \mathscr{R}^+ = \emptyset$; $\tau : \mathscr{A} \to [0,1]$ is a *base score function*. A QBAF is *acyclic* if its underlying graph has no directed cycles. In our setting, the discourse structures provided by the dataset are inherently acyclic, as they are defined via hierarchical parent–child relations. Moreover, acyclic structures yield a clearer argumentative flow from supporting and attacking units to the final decision, making the resulting reasoning process easier to understand.

A *gradual semantics* σ is a function that evaluates a QBAF $\langle \mathscr{A}, \mathscr{R}^-, \mathscr{R}^+, \tau \rangle$ by ascribing values $\sigma(\alpha) \in [0,1]$ to every $\alpha \in \mathscr{A}$ as their *strength*. In the paper, we adopt the QE semantics [8] to evaluate the dialectical strength of arguments. QE semantics satisfies several desirable properties proven in [8], and has been used in several practical settings (e.g. [11, 12]). For acyclic QBAFs, the strength of an argument α is defined as: $\sigma(\alpha) = \tau(\alpha) + (1 - \tau(\alpha))\frac{(E_\alpha)^2}{1+(E_\alpha)^2}$ when $E_\alpha > 0$, and $\sigma(\alpha) = \tau(\alpha) - \tau(\alpha)\frac{(E_\alpha)^2}{1+(E_\alpha)^2}$ otherwise, where E_α denotes the *aggregation strength* of all arguments supporting and attacking α, defined as $E_\alpha = \sum_{\{\beta \in \mathscr{A} | (\beta,\alpha) \in \mathscr{R}^+\}} \sigma(\beta) - \sum_{\{\beta \in \mathscr{A} | (\beta,\alpha) \in \mathscr{R}^-\}} \sigma(\beta)$.

Example 1. *Consider the QBAF in Figure 1. The base scores are given as follows:* $\tau(\textit{Evidence 1}) = 0.950$, $\tau(\textit{Evidence 2}) = 0.000$, $\tau(\textit{Rebuttal 1}) = 0.100$, $\tau(\textit{Claim 1}) = 0.150$, $\tau(\textit{Claim 2}) = 0.150$, $\tau(\textit{Counterclaim 1}) = 0.050$, *and* $\tau(\textit{Position 1}) = 0.010$. *Since Evidence 1, Evidence 2, and Rebuttal 1 have no children, we have* $E_{\textit{Evidence 1}} = E_{\textit{Evidence 2}} = E_{\textit{Rebuttal 1}} = 0$, *and thus* $\sigma(\textit{Evidence 1}) = 0.950$, $\sigma(\textit{Evidence 2}) = 0.000$, *and* $\sigma(\textit{Rebuttal 1}) = 0.100$. *For Claim 1, we have* $E_{\textit{Claim 1}} = \sigma(\textit{Evidence 1}) = 0.950$, *thus* $\sigma(\textit{Claim 1}) = 0.5532$. *For Claim 2, we have* $E_{\textit{Claim 2}} = \sigma(\textit{Evidence 2}) = 0.000$, *thus* $\sigma(\textit{Claim 2}) = 0.150$. *For Counterclaim 1, we have* $E_{\textit{Counterclaim 1}} = -\sigma(\textit{Rebuttal 1}) = -0.100$, *thus* $\sigma(\textit{Counterclaim 1}) = 0.0495$. *Finally, for Position 1, we have* $E_{\textit{Position 1}} = \sigma(\textit{Claim 1}) + \sigma(\textit{Claim 2}) - \sigma(\textit{Counterclaim 1}) = 0.6537$, *thus* $\sigma(\textit{Position 1}) = 0.3064$.

Metrics. We use Quadratic Weighted Kappa (QWK) [13] as the main evaluation metric for AES. For predicted and ground-truth scores on the 1–6 scale, QWK is defined as $\text{QWK} = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$, $w_{i,j} = \frac{(i-j)^2}{(K-1)^2}$, where $K = 6$, and O and E are the observed and expected agreement matrices. QWK is suitable for AES because it respects the ordinal nature of essay scores and penalizes larger disagreements more strongly.

4. QBAF Construction

4.1. Dataset

Dataset selection. We use PERSUADE 2.0 [14], a widely used public dataset for AES, as it provides discourse-unit-level argumentative annotations together with explicit hierarchical attachment information between units. This makes it possible to construct essay-specific and interpretable QBAFs directly from annotated discourse units, without requiring an additional argument mining stage.

Dataset description. PERSUADE 2.0 contains over 25,000 argumentative essays written by U.S. students in grades 6–12. Each essay is assigned a holistic score on a discrete 1–6 scale, where higher scores indicate better overall writing quality. For each discourse unit, the dataset provides its type (`Lead`, `Position`, `Claim`, `Evidence`, `Counterclaim`, `Rebuttal`, and `Conclusion`), its effectiveness label (`Effective`, `Adequate`, or `Ineffective`), and hierarchical attachment information linking the unit to its parent within the essay structure.

Unit Type	Effectiveness	Parent	Content
Position	Effective	–	*... all students should participate in community service at least once over the school year.*
Claim 1	Adequate	Position	*I think that the town certainly needs a clean-up and this is a great way to do it.*
Claim 2	Adequate	Position	*I would be willing to help you, and maybe could be the assistant in the community service.*
Evidence 1	Effective	Claim 1	*...The town has some very disgusting neighborhoods, causing water and air pollution...*
Evidence 2	Ineffective	Claim 2	*I have done a community service before and know how to get it done properly.*
Counterclaim 1	Adequate	Position 1	*Of course, most students have busy schedules, and may not be able to participate in it on a certain day.*
Rebuttal 1	Adequate	Counterclaim 1	*As a community, we need to take care of it ourselves and therefore, all students should be required to take part at least once over the year.*

Table 1
A PERSUADE 2.0 [14] annotation example corresponding to the motivating QBAF in the Introduction.

Data preprocessing. In PERSUADE 2.0, each row corresponds to one annotated discourse unit rather than a complete essay. We therefore group all rows sharing the same `essay_id` to recover essay-level samples. Rows labeled as `Unannotated` are excluded, as they do not correspond to argumentative discourse units. We also retain only essays with exactly one `Position` unit, ensuring a unique central stance (98.19% of essays). One concrete dataset example corresponding to the motivating QBAF in the Introduction is shown in Table 1.

Data split. We use the official train/test split provided by the dataset. After preprocessing, this consists of approximately 15,000 training essays and 10,000 test essays. From the training set, we further construct a validation set of approximately 3,000 essays using a stratified random split based on holistic scores, ensuring similar score distributions between train-dev and validation data. This results in a final train-dev/validation/test split of 12,000/3,000/10,000 essays (48%/12%/40%).

QE strength mapping. Since QE semantics generates strength values in $[0, 1]$, we map to a discrete essay score in 1–6 via data-driven thresholds. For each score level $k \in \{1, \dots, 6\}$, we compute μ_k, the mean root strength over all essays in the train-dev split whose ground-truth score is k. We then define the decision thresholds as the midpoints between adjacent class means, i.e., $t_k = (\mu_k + \mu_{k+1})/2$ for $k = 1, \dots, 5$. A new essay is assigned score k if its strength falls in $[t_{k-1}, t_k)$, with $t_0 = 0$ and $t_6 = 1$.

4.2. Discourse-structure-based QBAF Construction

QBAF topology construction. We construct an acyclic QBAF for each essay from its annotated discourse structure. Each argument node corresponds to one discourse unit of type `Lead`, `Position`, `Claim`, `Evidence`, `Counterclaim`, `Rebuttal`, or `Conclusion`. Briefly, a Lead introduces the topic, the Position states the essay's main stance, Claims support this stance, Evidence provides supporting reasons or examples, Counterclaims present opposing views, Rebuttals respond to counterclaims, and Conclusions restate the overall position and claims. We treat the unique `Position` as the main standpoint of the essay and hence as the *topic argument*. Edges are recovered from the hierarchical annotations provided in the dataset. In particular, each discourse unit is connected to its annotated parent unit. The polarity of each edge is determined by the discourse type of the child unit: edges originating from `Counterclaim` and `Rebuttal` are attack edges, while edges originating from the remaining discourse types are support edges. In this way, the resulting QBAF preserves the essay-specific discourse structure annotated in the dataset. Not every essay contains all seven discourse types. Figure 1 shows an example QBAF constructed from one essay in the dataset.

Node type	Effectiveness	Pearson	Initial BS	Local search set	Final BS
Position	Effective	0.204	0.010	{0.001, 0.010, 0.050}	0.010
Position	Adequate	-0.077	0.001	{0.001, 0.010, 0.050}	0.001
Claim	Effective	0.399	0.400	{0.350, 0.400, 0.450}	0.350
Claim	Adequate	0.258	0.200	{0.150, 0.200, 0.250}	0.150
Evidence	Effective	0.632	0.950	{0.900, 0.950, 0.990}	0.950
Evidence	Adequate	0.015	0.001	{0.001, 0.050, 0.100}	0.050
Counterclaim	Effective	0.247	0.300	{0.250, 0.300, 0.350}	0.250
Counterclaim	Adequate	0.170	0.100	{0.050, 0.100, 0.150}	0.050
Rebuttal	Effective	0.247	0.300	{0.250, 0.300, 0.350}	0.250
Rebuttal	Adequate	0.222	0.150	{0.100, 0.150, 0.200}	0.100
Lead	Effective	0.368	0.400	{0.300, 0.400, 0.500}	0.300
Lead	Adequate	0.140	0.100	{0.050, 0.100, 0.150}	0.150
Conclusion	Effective	0.452	0.500	{0.400, 0.500, 0.600}	0.600
Conclusion	Adequate	-0.034	0.001	{0.001, 0.050, 0.100}	0.100

Table 2
Pearson correlations, initial base scores, local search sets, and final base scores for each node type.

Base score assignment. Next, we assign base scores to discourse units to reflect their intrinsic quality prior to graph propagation. Since the dataset provides qualitative effectiveness labels for each discourse unit, namely `Effective`, `Adequate`, and `Ineffective`, we convert these annotations into quantitative base scores via a two-stage procedure. We impose two constraints throughout: (i) for each discourse type, `Effective` > `Adequate` > `Ineffective`; and (ii) all `Ineffective` nodes are assigned a base score of 0, reflecting that they contribute no positive prior quality before propagation, while still influencing the final score through the graph structure.

In the first stage, we compute the Pearson correlation between the frequency of each (`discourse type`, `effectiveness`) combination and the holistic ground-truth score, and use these correlations to define correlation-guided initial base scores. Intuitively, combinations with stronger positive correlation are assigned higher initial base scores, while weakly correlated or near-zero combinations are assigned lower scores. Since the `Position` node is the root of the QBAF, we further lower its initial base scores to avoid early saturation under QE semantics.

In the second stage, we refine these initial base scores via a local random search. For each tunable base score, we construct a local candidate set consisting of its initial base score and one lower and one higher neighbouring value, yielding three candidates per base score (for 14 tunable base scores, excluding `Ineffective`). Rather than exhaustively evaluating all combinations, we randomly sample 1,000 configurations on a 10% subset of train-dev (around 1,200 essays), retain the top 100 according to QWK under QE semantics, and select the best-performing configuration on the validation set. Using the resulting base-score configuration in the structure-only QBAF yields QWK values of 0.6720 on train-dev, 0.6761 on validation, and 0.6677 on the test set.

Table 2 reports the final selected base score configuration, together with the corresponding Pearson correlations, correlation-guided initial scores, and local search sets. A key observation is that the final base scores vary across discourse types. Although they generally follow the Pearson-based ranking, the final configuration is also shaped by the QBAF topology and the underlying QE semantics.

4.3. Feature-based QBAF Construction

Interpretable handcrafted features. The current QBAF topology captures the logical structure of an essay through discourse units and their relations. However, holistic score prediction also depends on summary statistics that are not explicitly represented by the graph structure. To incorporate such

No.	Feature	Description
Essay-level summary features		
1	essay_word_count	Total number of words in the essay.
2	total_argument_units	Number of annotated argumentative units.
3	num_discourse_types	Number of distinct discourse types.
4	structural_completeness	0/1/2 indicating whether the essay contains neither/one/both of Lead and Conclusion arguments.
5	ineffective_unit_ratio	Fraction of annotated units labeled Ineffective.
Discourse-unit-level features		
6	avg_unit_words	Average number of words per annotated discourse unit.
7	claim_child_count	Number of units with hierarchical_label=Claim.
8	position_child_count	Number of units with hierarchical_label=Position.
9	rebuttal_per_counterclaim	Average number of Rebuttal units per Counterclaim.

Table 3
The 9 handcrafted features used in the final QBAF.

information while preserving interpretability, we introduce a set of handcrafted features derived from the essay and represent them as additional arguments in the QBAF. These features capture complementary information beyond the QBAF topology without relying on black-box representations, and are grouped into two categories with different granularities, as shown in Table 3. The first group, *essay-level* features, captures global features of an essay, such as essay length and the total number of discourse units. The second group, *discourse-unit-level* features, summarizes statistical features of discourse units, such as average unit length and the number of children of the Position unit. These features are chosen to capture simple, interpretable, and pedagogically meaningful indicators of essay quality that are not fully reflected by discourse structure alone.

Base score assignment. To integrate the 9 interpretable handcrafted features into the QBAF, we first represent each feature as an argument node. For each such node, its base score is determined via a statistical analysis on the train-dev split. Specifically, the analysis is based on bin-level summary statistics. For low-cardinality discrete features, such as rebuttal_per_counterclaim, we compute the average essay score for each distinct value. For continuous features, we partition the train-dev data into quantile-based bins and compute the average essay score within each bin. The number of bins is empirically selected from $\{8, 12, 16\}$ based on validation QWK, with 12 bins giving the best performance. The resulting feature-score trends and bin-level mappings are illustrated in Figure 2. The resulting bin-level averages are then linearly normalized from the original score range 1–6 to $[0, 1]$. At test time, each feature value is mapped to its corresponding bin, and the normalized average score of that bin is used as the base score of the associated node.

QBAF topology enhancement. After assigning the base scores to the feature argument nodes, we connected each of them directly to the topic argument (Position), which determines the final essay score. Next, to assign edge polarity, we quantify the monotonic association between feature values and average essay scores. Specifically, we compute the Pearson correlation between bin indices (ordered by increasing feature value) and bin-level average scores, using quantile-based bins for continuous features and distinct values for discrete features. The polarity is then determined by the sign of the correlation: a positive correlation indicates a support feature, whereas a negative correlation indicates an attack feature (e.g., essay_word_count as support and ineffective_unit_ratio as attack). Figure 3 illustrates several resulting enhanced QBAFs, which augment the recovered argumentative structure with interpretable handcrafted features.

5. Evaluations

5.1. Baseline Methods

To test the QWK performance of the proposed QBAFs, we consider two groups of commonly used baselines in machine learning: white-box models and grey/black-box models. The grey/black-box

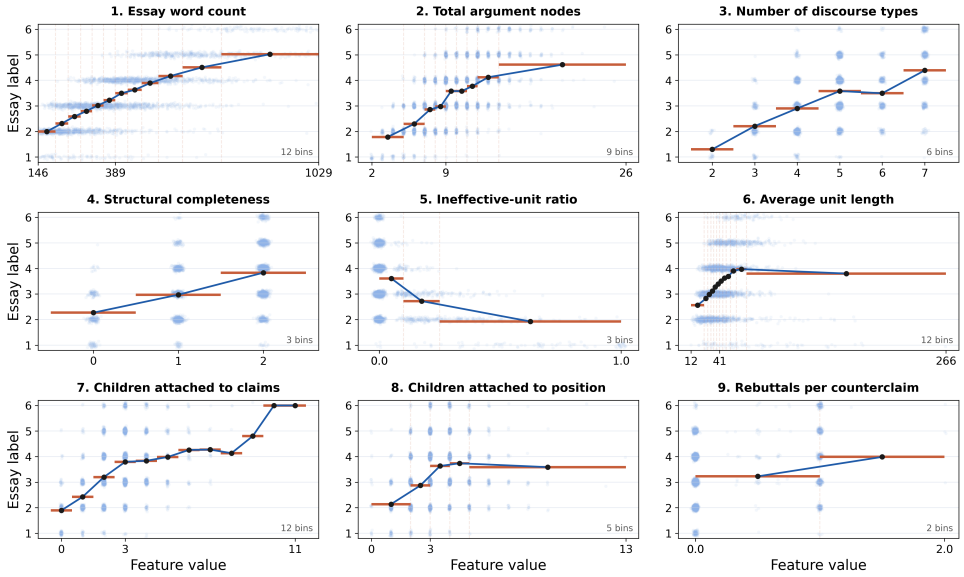

Figure 2: Scatter–bin plots of the 9 handcrafted features and their bin-level base score trends.

baselines represent mainstream predictive models in AES including Multi-Layer Perceptrons (MLPs), Random Forest, Gradient Boosting, and Linear Support Vector Machine (SVM). The white-box baselines include Logistic Regression, Decision Tree, and Gaussian Naive Bayes (NB). All baseline models are trained on the same 9 handcrafted features used in the final QBAF setting.

Baseline configurations. We standardize the input features for Linear SVM, Logistic Regression, and MLP using `StandardScaler`, whereas tree-based models and Gaussian NB use the raw features. Linear SVM (`LinearSVC`) uses $C = 2.0$, squared-hinge loss, and L_2 regularization. Logistic Regression uses $C = 4.0$, the `lbfgs` solver, and `max_iter=4000`. The Decision Tree uses the Gini criterion with `max_depth=7` and `min_samples_leaf=20`. Gaussian NB uses default priors with `var_smoothing=1e-9`. The MLP has a single hidden layer of size 64 with `ReLU` activation and the `Adam` optimizer (`max_iter=500`). The Random Forest uses 300 trees with `max_depth=16`, `min_samples_leaf=10`, and `max_features=sqrt`. Gradient Boosting uses `log_loss`, `learning_rate=0.1`, `n_estimators=100`, and `max_depth=2`. All randomized models use `random_state=42`, and other hyperparameters follow `scikit-learn` defaults unless specified.

5.2. Results and Analyses

Table 4 reports the QWK performance across all baselines. Overall, the highest QWK values are achieved by black-box neural models, reaching around 0.84 on the test set. By contrast, the proposed QBAF remains competitive among white-box models, achieving comparable performance while outperforming Gaussian NB.

Argumentative explanations. Beyond predictive performance, the main advantage of QBAF lies in its intrinsic interpretability, stemming from its transparent structure and deterministic semantics. First, the recovered argumentative structure provides a clear reasoning backbone: support and attack relations explicitly capture how discourse units interact, while the base score of each argument reflects its local effectiveness. This allows learners and teachers to inspect the overall argumentative organization of an essay, for example whether it lacks counterclaim or rebuttal units, and whether individual discourse units are effective or need improvement. Second, the influence of handcrafted features to the final essay score is also transparent. Users can understand how features such as `number_of_discourse_types` and `structural_completeness` influence the final score by examining their relations to the topic argument and their base scores. More fine-grained argumentative explanation methods could further quantify how individual discourse units contribute to, or drive changes in, the final score [15, 16, 17, 12].

Baseline Model	Train-dev QWK	Validation QWK	Test QWK
White-box models			
Logistic Regression	0.8104	0.8047	0.8043
Decision Tree	0.8317	0.8063	0.8184
Gaussian NB	0.7899	0.7784	0.7925
QBAF (Ours)	0.7992	0.8015	0.8021
Grey/Black-box models			
Linear SVM	0.7631	0.7605	0.7557
Random Forest	**0.8592**	0.8276	0.8329
Gradient Boosting	0.8437	0.8277	0.8298
MLPs	0.8490	**0.8359**	**0.8394**

Table 4
QWK results of all models. The best value in each column is shown in bold.

In addition, the scoring process is deterministic. Under QE semantics, the same QBAF topology and base scores always produce the same final value, making the decision process fully reproducible. Argumentation has also been used to construct interpretable representations of conventional machine-learning models [18]. Moreover, faithfulness to the underlying model is an important desideratum for XAI [19]; here, the QBAF itself serves as both the scoring model and the basis for its explanations. Compared with other white-box models, the QBAF provides more essay-specific explanations. Logistic Regression explains predictions through the contributions of the 9 handcrafted features, while a Decision Tree provides a prediction path based on threshold tests over the same features, and Gaussian NB relies on their class-conditional likelihoods. Although these mechanisms are transparent, their explanations remain at the level of aggregate essay features. In contrast, the QBAF additionally preserves the essay-specific discourse structure, allowing explanations to identify individual discourse units, their effectiveness, and the support/attack relations through which they influence the Position. We therefore argue that QBAF offers a favourable trade-off between predictive performance and interpretability for AES tasks.

Illustrative Examples. As a qualitative illustration (Figure 3), we select six correctly predicted test essays covering all score levels from 1 to 6 [2]. In each sub-figure, the left-hand side shows the recovered argumentative structure, while the right-hand side shows the summary feature nodes.

These six examples qualitatively illustrate two possible patterns, rather than establishing general trends. First, lower-scoring essays tend to exhibit relatively simple argumentative structures, with fewer discourse units and shorter reasoning patterns centred around the `Position` node. By contrast, medium- and high-scoring essays display richer argumentative organization, including more `Claim` and `Evidence` nodes, and in some cases additional components such as `Counterclaim` and `Rebuttal`. For example, the score-4 essay already shows a more developed structure than the low-score cases, while the score-6 essay contains both multiple effective support units and an explicit counterclaim–rebuttal pattern. Second, the local quality of discourse units varies across score levels. Lower-score examples contain few or no `Effective` units, whereas higher-score essays include a larger proportion of locally effective components, such as effective `Claim`, `Evidence`, and `Conclusion` nodes. Overall, these examples illustrate how the QBAF representation can capture both the global argumentative organization and the local effectiveness of discourse units, providing an interpretable account of individual scoring cases.

5.3. Ablation Studies

Finally, to analyse the contribution of each component in the proposed QBAF, we conduct an ablation study comparing three variants: a *structure-only* QBAF, which uses only the recovered argumentative structure and discourse-unit base scores; a *feature-only* QBAF, which uses only the 9 interpretable handcrafted feature nodes; and the *full* QBAF, which combines both components. The test-set QWK

[2]The underlying text corresponding to the discourse-unit nodes in these six examples is available at: https://github.com/XiangYin2021/QBAF4EDU/blob/main/FIGURE3_EXAMPLE_NODE_CONTENTS.txt.

scores are: structure-only QBAF: 0.6677; feature-only QBAF: 0.7879; and full QBAF: 0.8021. These results reveal three clear patterns. First, the structure-only QBAF achieves relatively low performance, indicating that argumentative topology and local unit effectiveness alone are insufficient for accurate holistic score prediction. Second, the feature-only variant achieves substantially better performance, suggesting that handcrafted summary features provide stronger and more direct signals for essay-level scoring. Third, the full QBAF performs best overall, demonstrating that the structural and feature-based components are complementary rather than redundant. In particular, handcrafted features contribute most to predictive performance through strong global signals, while the recovered argument graph provides additional performance gains and, importantly, offers a transparent reasoning backbone for the scoring process.

6. Conclusions

We proposed an interpretable approach to AES based on QBAFs. By representing essays as argumentative structures and evaluating them under QE semantics, the proposed framework provides a transparent and structured reasoning process for essay score prediction. We further enhanced the model with interpretable handcrafted features, enabling it to capture complementary statistical information beyond discourse structure. Experimental results on PERSUADE 2.0 show that the proposed approach achieves competitive performance against standard baseline models, while offering clear advantages in interpretability. Overall, this work suggests that argumentation-based models provide a promising foundation for transparent and interpretable AES.

Future Work. Future work includes extending the framework to an end-to-end setting from raw essay text, covering discourse unit identification, relation extraction, effectiveness estimation, and feature construction, potentially leveraging recent advances in argumentation with large language models [20, 21]. An important question is how errors in automatically inferred annotations propagate to the final essay score, and what level of annotation quality is required for the resulting scores and explanations to remain reliable and informative. Another direction is to explore a broader range of interpretable handcrafted features, such as richer statistics on the distribution of different discourse types, while developing compact and non-redundant feature sets. Finally, the framework could be extended to support interactive and contestable scoring. For example, students could ask why an essay received a particular score, which discourse units most influenced the score, or obtain counterfactual explanations [22] identifying how specific revisions could improve it. Moreover, if users identify problematic aspects of the scoring process, they could contest components of the QBAF [23], such as argument base scores, relations, or feature mappings, and trigger an updated evaluation. A further extension could consider multiple human or AI evaluators, where argumentative exchanges and post-hoc debate judgement may support principled resolution of competing assessments [24].

Acknowledgments

We thank the anonymous reviewers for their careful reading and detailed, constructive comments, which helped us improve the clarity and presentation of this work.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT 5.5 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

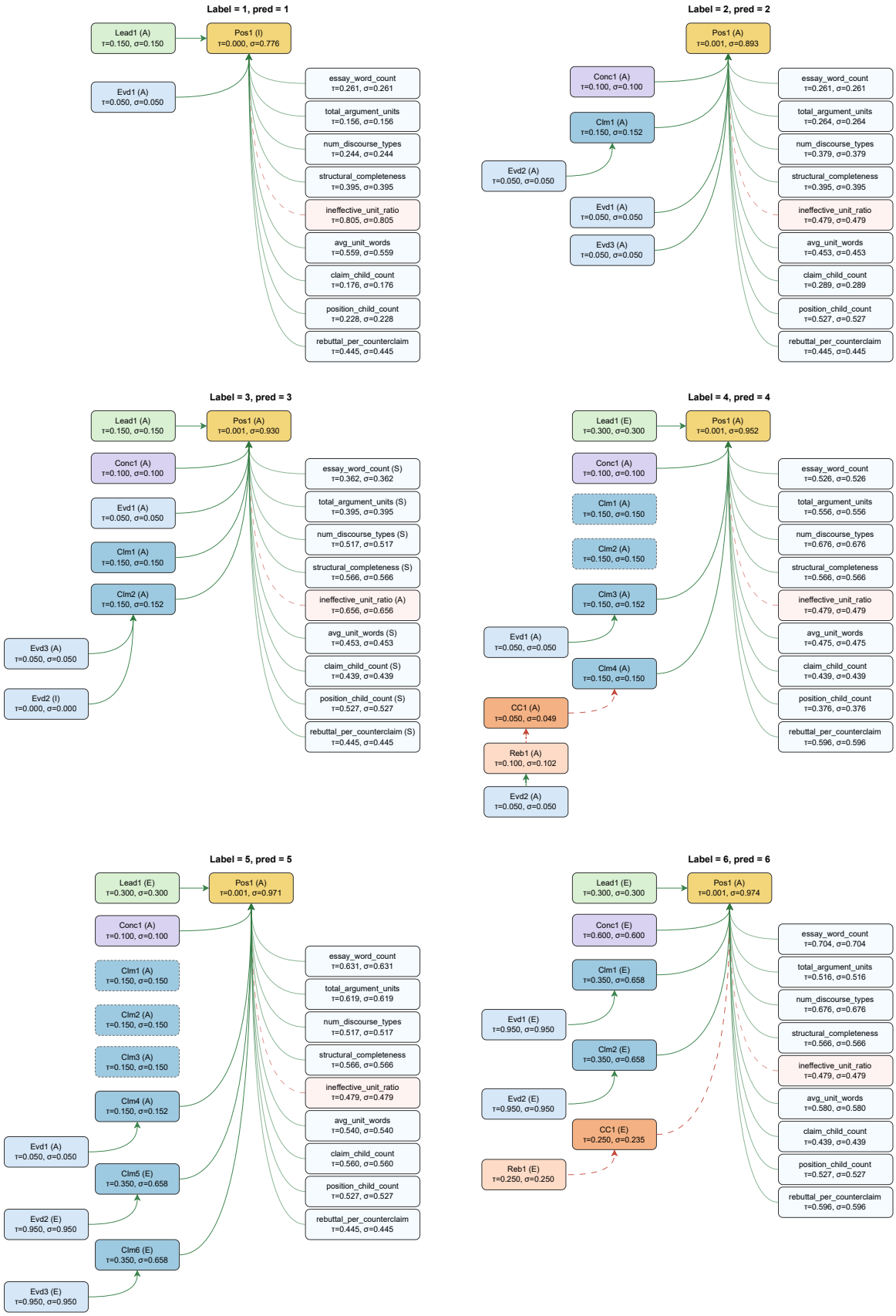

Figure 3: Six representative correctly predicted test essays, one for each score level from 1 to 6. In each sub-figure, the left-hand side shows the recovered argumentative structure and the right-hand side shows the handcrafted features. For visual simplicity, support/attack relations are shown by solid and dashed edges, respectively, without explicit "+/-" labels. τ denotes the base score of an argument and σ denotes its final QE strength.

References

[1] B. B. Klebanov, N. Madnani, Automated essay scoring, Springer Nature, 2022.

[2] K. Taghipour, H. T. Ng, A neural approach to automated essay scoring, in: Proceedings of the 2016 conference on EMNLP, 2016, pp. 1882–1891.

[3] S. Li, V. Ng, Automated essay scoring: A reflection on the state of the art, in: Proceedings of the 2024 Conference on EMNLP, 2024, pp. 17876–17888.

[4] D. Ramesh, S. K. Sanampudi, An automated essay scoring systems: a systematic literature review, Artificial intelligence review 55 (2022) 2495–2527.

[5] European Union, Regulation (eu) 2024 of the european parliament and of the council laying down harmonised rules on artificial intelligence (ai act), 2024. Annex III: High-risk AI systems.

[6] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative xai: A survey, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4392–4399.

[7] P. Baroni, M. Romano, F. Toni, M. Aurisicchio, G. Bertanza, Automatic evaluation of design alternatives with quantitative argumentation, Argument & Computation 6 (2015) 24–49.

[8] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: 16th International Conference on Principles of Knowledge Representation and Reasoning (KR), 2018.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[10] J. Atkinson, D. Palma, An llm-based hybrid approach for enhanced automated essay scoring, Scientific Reports 15 (2025) 14551.

[11] C. Al Anaissy, N. Maudet, Explaining online debate evolution under bipolar gradual argumentation semantics, in: 18th International Conference on Agents and Artificial Intelligence, volume 2, 2026, pp. 1528–1539.

[12] T. Kampik, X. Yin, N. Potyka, F. Toni, Strength change explanations in quantitative argumentation (2026).

[13] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit., Psychological bulletin 70 (1968) 213.

[14] S. A. Crossley, Y. Tian, P. Baffour, A. Franklin, M. Benner, U. Boser, A large-scale corpus for assessing written argumentation: Persuade 2.0, Assessing Writing 61 (2024) 100865.

[15] X. Yin, N. Potyka, F. Toni, Argument attribution explanations in quantitative bipolar argumentation frameworks, in: European Conference on Artificial Intelligence (ECAI), volume 372, 2023, pp. 2898–2905.

[16] X. Yin, N. Potyka, F. Toni, Explaining arguments' strength: Unveiling the role of attacks and supports, in: International Joint Conference on Artificial Intelligence (IJCAI), 2024.

[17] T. Kampik, N. Potyka, X. Yin, K. Cyras, F. Toni, Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis, Int. J. Approx. Reason. 173 (2024) 109255. URL: https://doi.org/10.1016/j.ijar.2024.109255. doi:10.1016/J.IJAR.2024.109255.

[18] N. Potyka, X. Yin, F. Toni, Explaining random forests using bipolar argumentation and markov networks, in: AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 9453–9460.

[19] N. Potyka, X. Yin, F. Toni, Towards a theory of faithfulness: Faithful explanations of differentiable classifiers over continuous data, arXiv preprint arXiv:2205.09620 (2022).

[20] G. Freedman, A. Dejl, D. Gorur, X. Yin, A. Rago, F. Toni, Argumentative large language models for explainable and contestable claim verification, in: T. Walsh, J. Shah, Z. Kolter (Eds.), AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA, AAAI Press, 2025, pp. 14930–14939. URL: https://doi.org/10.1609/aaai.v39i14.33637. doi:10.1609/AAAI.V39I14.33637.

[21] L. Chen, X. Yin, F. Toni, Latent debate: A surrogate framework for interpreting llm thinking, arXiv preprint arXiv:2512.01909 (2025).

[22] X. Yin, N. Potyka, F. Toni, CE-QArg: counterfactual explanations for quantitative bipolar argumentation frameworks, in: Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, 2024, pp. 697–707.
[23] X. Yin, N. Potyka, A. Rago, T. Kampik, F. Toni, Contestability in Edge-Weighted Quantitative Bipolar Argumentation Frameworks, in: Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning, 2026, pp. 676–687. URL: https://doi.org/10.24963/kr.2026/64. doi:10.24963/kr.2026/64.
[24] X. Yin, A. Dejl, A. Rago, L. Chen, F. Toni, A theory of post-hoc debate judgement, arXiv preprint arXiv:2608.19002 (2026).