Faithfulness Beyond Plausibility: Auditing Human Explanations in Educational Assessment

Ria Talsania Dhruv Shah Sudhir Dhage (Senior Member, IEEE)
Department of Computer Engineering
Sardar Patel Institute of Technology
{ria.talsania23, dhruv.shah22, sudhir_dhage}@spit.ac.in

Abstract

When rubric-based feedback tools explain a grade, students and instructors assume those explanations reflect how the score was actually determined. Yet it remains unclear whether explanation components such as rubric assignments and evidence spans reflect how scores are constructed or primarily serve as post-hoc justifications. This gap has direct implications for automated essay scoring and rubric-based feedback tools, where explanation reliability is often assumed but rarely evaluated.

We introduce a knowledge graph framework that represents human tutor grading traces as structured objects, enabling controlled counterfactual testing of explanation components. Using 400 grading traces from 10 expert human tutors evaluating 100 narrative essays, we define a reconstruction-based diagnostic to measure how explanation components contribute to score interpretation, independent of prediction. Our results reveal a consistent asymmetry: removing rubric-level information leads to substantial changes in reconstructed scores, while removing evidence spans has minimal impact. This suggests that rubric structure is central to score interpretation, whereas cited evidence spans may function primarily as post-hoc justifications. We further observe tutor-specific variation in grading behavior. These findings highlight the need for explanation mechanisms that better align with scoring processes, ensuring that feedback provided to learners is both interpretable and functionally relevant.

1 Introduction

When a human tutor assigns a score to a student essay and writes that the opening lacked narrative tension or that the plot development was underdeveloped, the student reads these comments as the reason for the score. Yet in practice, holistic grading does not work this way. Human tutors typically form an overall judgment first, then construct a structured justification afterward (Messick, 1989). Explanation components accompanying a grade, rubric assignments, evidence spans, and written comments may therefore not reflect how the score was actually determined, but instead represent a post-hoc narrative selected to be consistent with a judgment already made.

This distinction matters for educational practice. As rubric-based feedback tools and automated essay scoring systems become more widely deployed (Ke and Ng, 2019; Zhang et al., 2020), explanations are increasingly used by students to understand their performance, by instructors to calibrate their grading, and by institutions to validate assessment quality (Madnani et al., 2018). If explanation components are structurally disconnected from scoring behavior, feedback built on these explanations risks misleading learners about what actually determines their grades.

Existing approaches have largely treated this as a prediction problem: rubric-based AES systems explain holistic scores by predicting rubric-level scores from linguistic features (Ke and Ng, 2019; De Vrindt et al., 2025), while faithfulness research evaluates explanations by perturbing inputs and observing changes in model predictions (Jacovi and Goldberg, 2020), assuming explanation signals act as causal drivers of decisions (Jain and Wallace, 2019; Serrano and Smith, 2019). Fairness analyses have further shown that scoring behavior varies systematically across student subgroups (Kwako and Ormerod, 2024; Schaller et al., 2024). Yet a complementary question remains unaddressed: how do individual evaluators structure their justifications, and do those justifications align with their scoring behavior? We address this gap through the following contributions:

1. A knowledge graph framework for auditing human grading explanations as structured, testable objects.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 621–631

2. Diagnostics for rubric attention, evidence usage, intra-tutor consistency, and rubric sensitivity, revealing evaluator-specific patterns not recoverable from scores alone.

3. The *Explanatory Final Score* (EFS), a reconstruction-based diagnostic quantifying evidential support for grading judgments without any predictive model.

4. A counterfactual validation demonstrating a structural asymmetry: rubric nodes are load-bearing for score reconstruction while evidence spans are not.

2 Related Work

Automated essay scoring (AES) systems have evolved from feature-engineered approaches (Attali and Burstein, 2006) to neural models (Taghipour and Ng, 2016; Dong and Zhang, 2017), with recent work emphasizing interpretability through rubric alignment or feedback generation. Rubric-based approaches have sought to provide more interpretable scoring by grounding predictions in explicit assessment criteria (Ke and Ng, 2019; Zhang et al., 2020), with recent work extending this to multi-trait scoring across multiple rubric dimensions simultaneously (Do et al., 2023; He et al., 2022). While these approaches improve transparency relative to holistic scoring models, they remain outcome-oriented: rubric predictions are validated by their correlation with human scores rather than by whether they reflect the structural reasoning of human evaluators. Our framework departs from this paradigm by treating rubric-linked evidence as a structural artifact of human reasoning rather than a predictive signal. This concern is directly relevant to rubric-based scoring systems currently being developed for educational practice, including the BEA 2026 shared task on rubric-based short answer scoring, where the reliability of rubric-linked explanations remains an open question.

Most closely related to our work, De Vrindt et al. (2025) proposed explaining holistic scores from comparative judgment assessments by predicting rubric scores from linguistic features. While this provides transparency about what holistic scores represent, it remains model-centric and critically assumes that rubric scores causally explain holistic judgments. Our findings directly challenge this assumption: removing rubric-linked evidence spans leaves EFS and the reconstructed score stable, while removing rubric nodes causes both to collapse, indicating that rubric structure is load-bearing for explanatory reconstruction while individual spans are not. Rather than predicting rubric scores, our framework audits the structural properties of existing human explanations without assuming an underlying predictive decision function.

Prior work has examined explanations as communicative artifacts rather than causal traces (Miller, 2019), and educational assessment research emphasizes that grading involves nuanced judgment and selective evidence use beyond numerical scores (Madnani et al., 2018). Measurement theory similarly evaluates validity through interpretability and consequences of evaluation practices (Messick, 1989). Our framework enables structural auditing of tutor explanations aligned with this perspective.

Within educational measurement, the distinction between construct validity and surface plausibility is well established (Messick, 1989): an explanation may appear well-reasoned to a reader while failing to reflect the actual evaluative process that produced the score. Cognitive research on expert judgment further shows that evaluators often cannot accurately report the criteria driving their decisions (Nisbett and Wilson, 1977), constructing post-hoc narratives that are coherent but not causally accurate. Our framework operationalizes this insight structurally: rather than asking tutors to self-report their reasoning, we audit the relationship between their provided evidence and their scores directly through the knowledge graph.

Knowledge graphs have been used in NLP for structured reasoning and explainability (Bosselut et al., 2019; Ilievski et al., 2021; Guo et al., 2022), with educational applications modeling assessment criteria and learner knowledge (Liu et al., 2019). Unlike Ilievski et al. (2021), who use knowledge graphs as auxiliary reasoning inputs for downstream prediction, our graph encodes tutor behavior as the primary object of structural analysis rather than as a feature for a learned model. This design choice is reflected directly in our schema: the six node types and seven edge types are chosen to preserve the structural separation between holistic judgment and cited evidence, making counterfactual node removal tractable without reconstructing relational dependencies from scratch.

Graph-based representations have also been applied to model argumentation structure and evidence relations in educational contexts (Stab and Gurevych, 2017), where the goal is to identify how

claims are supported by textual evidence. Our use of knowledge graphs is complementary but distinct: rather than parsing argument structure from essay text, we encode evaluator behavior, linking rubric criteria, evidence spans, and scores as relational objects. Whereas argumentation graphs treat evidence relations as properties of the text, our graph treats them as properties of the evaluator, enabling auditing of explanation properties such as coverage, dependence, and consistency at the individual tutor level, without making assumptions about the underlying decision function.

Research on free-text explanation faithfulness is also directly relevant to our setting. Wiegreffe and Pinter (2019) and Hase et al. (2020) examine whether natural language explanations are faithful to underlying model decisions, distinguishing between explanations that are plausible to humans and those that are causally grounded in the decision process. Our work extends this concern to human evaluators: just as model-generated explanations may be post-hoc rationalizations rather than causal traces, tutor-provided evidence spans may function as justificatory artifacts rather than determinants of grading decisions.

Faithfulness-oriented approaches ask whether explanations change when inputs are modified (Jacovi and Goldberg, 2020; Wachter et al., 2017), with benchmarks such as ERASER (DeYoung et al., 2020) evaluating whether rationales are necessary for model predictions. Feature-attribution methods such as LIME and SHAP (Ribeiro et al., 2016; Lundberg and Lee, 2017) similarly identify input features that influence model outputs. Our setting differs fundamentally: we analyze explanations produced by human evaluators rather than model predictions, focusing on structural properties of justification rather than feature attribution.

While prior faithfulness research has focused on model-generated explanations, the criteria for evaluating faithfulness — causal grounding, structural consistency, and resistance to post-hoc rationalization — apply equally when the decision-maker is human. Human evaluators, like learned models, produce explanations that may be plausible without being causally accurate (Nisbett and Wilson, 1977; Miller, 2019). Our framework extends faithfulness auditing to this setting, treating human grading traces as behavioral objects subject to the same structural scrutiny applied to model outputs, while recognising that the absence of a learned decision function requires structural rather than perturbation-based evaluation criteria.

3 Knowledge Graph Construction

We represent each grading trace as a knowledge graph encoding the relational structure of a tutor's evaluative reasoning, connecting evaluators to essays, rubric criteria to textual evidence spans, and scores to justifications. Three properties make this representation essential. First, it operates exclusively on tutor annotations with no learned representations, ensuring observed patterns reflect actual tutor behavior rather than artifacts of a scoring function, distinguishing our framework from rubric prediction approaches (Ke and Ng, 2019; De Vrindt et al., 2025). Second, targeted counterfactual interventions are tractable: removing a rubric node automatically severs all associated ATTENDS_TO and EVIDENCE_FOR edges without reconstructing relational dependencies from scratch. Third, each grading trace extends a tutor's reasoning history without retraining, enabling explanation reliability to be evaluated relative to the individual evaluator, a necessary condition given that faithfulness is not a population-level property but an evaluator-specific one.

3.1 Dataset Overview

Our dataset comprises 400 grading traces from 10 expert tutors evaluating 100 narrative essays written by secondary school students. Essays were composed in response to a shared prompt and graded independently using a ten-criterion rubric: Opening, Plot, Characters, Description, Theme Depth, Language Style, Ending, Accuracy, Control Register, and Creativity, with rubric scores summing to a total mark out of 40.

Each grading trace contains three annotation types: rubric-level scores, free-form written comments, and span-level annotations linking contiguous text segments to specific rubric criteria, capturing not only the score assigned but which textual evidence was cited and under which criterion. Tutors did not grade all essays, reflecting realistic marking conditions; this partial assignment structure allows tutor-specific patterns to emerge without confounding cross-tutor comparison. While smaller than large-scale AES benchmarks, this scale is appropriate for a diagnostic framework paper and is comparable to prior structural analyses of human assessment behavior (De Vrindt et al., 2025).

3.2 Tutor-Specific Grading Profiles

Each tutor is associated with a grading profile comprising a stated grading focus and a rubric preference distribution indicating the relative importance assigned to each of the ten criteria. Declared priorities vary substantially across evaluators: some tutors concentrate weight on a single criterion such as Theme Depth or Description, while others distribute attention uniformly across all rubric dimensions. These weights are purely descriptive. They are not learned parameters and are never used to compute grades; they provide interpretive context for observed explanation behavior. Although all tutors apply the same rubric schema, differences in declared emphasis give rise to systematic variation in evidence selection and justification construction, motivating analysis at the individual evaluator level rather than across tutors.

3.3 Graph Schema

The schema is fixed across all tutors and essays, ensuring grading traces are represented in a consistent, comparable structure. It comprises six node types and seven edge types, each corresponding to a distinct component of the assessment process.

Node types. *Tutor*, *Essay*, and *Prompt* nodes represent evaluators, submissions, and the shared writing task respectively. *Rubric* nodes represent the ten evaluation criteria. *Span* nodes represent contiguous text regions annotated as evidence for a rubric criterion. *OverallEvaluation* nodes capture a tutor's complete assessment, including total mark and written feedback.

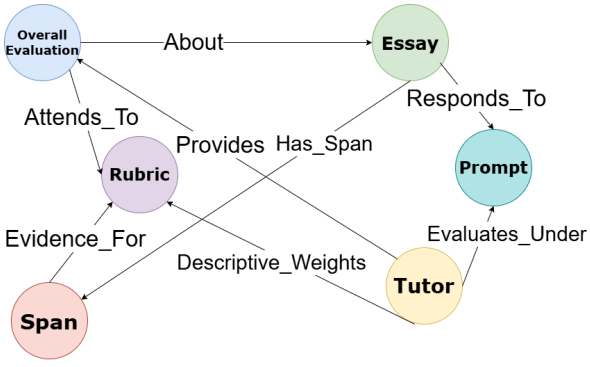

Figure 1: Knowledge graph schema illustrating the six node types and seven edge types used to represent a grading trace.

Edge types. Essays link to prompts via RESPONDS_TO and tutors to prompts via EVALUATES_UNDER. Assessments connect to evaluators via PROVIDES and to essays via ABOUT. Spans link to essays via HAS_SPAN and to rubric criteria via EVIDENCE_FOR. Rubric-level scores are stored on ATTENDS_TO edges between *OverallEvaluation* and *Rubric* nodes.

As illustrated in Figure 1, rubric-level scores and free-form comments are recorded together in a single *OverallEvaluation* node, while evidence spans are linked separately to rubric criteria via EVIDENCE_FOR, preserving the structural distinction between a tutor's holistic judgment and the textual evidence cited in support of it. This separation is what makes targeted counterfactual interventions tractable: removing a rubric node severs its ATTENDS_TO and EVIDENCE_FOR edges in a single operation, isolating the structural contribution of that criterion to the explanation without disturbing the holistic score recorded in the *OverallEvaluation* node.

4 Empirical Analysis of Tutor Explanation Structure

4.1 Local Explanation Diagnostics

We analyze structural properties of tutor explanations to understand how rubric criteria and textual evidence are used to justify grading decisions. We examine three diagnostics: rubric attention, evidence usage, and intra-tutor consistency.

Rubric Attention. We measure how frequently tutors ground explanations in textual evidence associated with each rubric. For tutor t and rubric r, rubric attention is defined as:

$$A(t, r) = |S_{t,r}|$$

where $S_{t,r}$ denotes the set of evidence spans linked to rubric r. Tutor 1 cites the highest attention for *Characters* ($A = 5$) compared to all other criteria ($A = 2$ each), while Tutor 2 concentrates on *Description* ($A = 4$) and *Theme Depth* ($A = 3$), with *Characters* receiving lower attention ($A = 2$). These differences persist across essays and are not recoverable from rubric scores alone.

Evidence Usage. We examine how evidence span density relates to rubric-level score magnitude. For tutor t, essay e, and rubric r, the normalized rubric score is:

$$\hat{s}(t, e, r) = \frac{a_{t,e,r}}{m_r}$$

where $a_{t,e,r}$ is the score assigned by tutor t to rubric r for essay e, and m_r is the maximum possible

score for that criterion. Span count does not increase monotonically with score magnitude. For Tutor 1 on Essay 2, *Characters* receives 4 spans yet a normalized score of only 0.33, while *Theme Depth* receives 1 span with a score of 1.0 on Essay 1. This pattern confirms that tutors use evidence interpretively rather than additively, with spans functioning as selective justificatory references rather than aggregated evidence.

Intra-Tutor Consistency. We examine whether tutors apply rubric criteria consistently across essays. For tutor t and rubric r, rubric variability is defined as:

$$V(t, r) = \max_e \hat{s}(t, e, r) - \min_e \hat{s}(t, e, r)$$

As shown in Table 1, *Accuracy* exhibits high consistency for both tutors (variance $= 0.17$), functioning as a stable constraint. *Theme Depth* shows zero variance for Tutor 2 yet variance of 0.67 for Tutor 1, reflecting evaluator-specific differences in how the same criterion is applied. Consistency patterns are therefore tutor-dependent and cannot be assessed against a shared standard.

Tutor	Rubric	Score Range	Consistency
Tutor 1	Opening	1.00	Low
Tutor 1	Description	1.00	Low
Tutor 1	Characters	0.67	Low
Tutor 1	Plot	0.60	Low
Tutor 1	Accuracy	0.17	High
Tutor 2	Creativity	0.75	Low
Tutor 2	Description	0.67	Low
Tutor 2	Characters	0.33	Medium
Tutor 2	Accuracy	0.17	High
Tutor 2	Theme Depth	0.00	High

Table 1: Rubric-level consistency patterns for Tutor 1 and Tutor 2, showing score range and consistency classification across essays.

4.2 Explanatory Score Reconstruction

To summarize how well a grading judgment is supported by explicit rubric-linked evidence, we introduce the *Explanatory Final Score* (EFS): a reconstruction-based diagnostic that aggregates rubric-level justification into a single measure in the same unit as the original assessment. EFS is not a predictive metric; it quantifies for a fixed grading judgment how much of that judgment is explicitly grounded in the rubric-linked evidence the tutor has provided.

For tutor t, essay e, and rubric r, let $S_{t,e,r}$ denote the set of evidence spans annotated for r. The proxy justification score is:

$$p(t, e, r) = \frac{1}{|S_{t,e,r}|} \sum_{s \in S_{t,e,r}} \frac{marks(r)}{max\_marks(r)}$$

When no spans are present, $p(t, e, r) = 0$, reflecting an absence of explicit justification for that criterion.

Aggregating proxy scores across essays yields rubric-level sensitivity:

$$\mathrm{Var}(t, r) = \mathrm{Var}_e(p(t, e, r))$$

Higher variance indicates a discriminative rubric; lower variance indicates a stable constraint. $\mathrm{SoftDep}(t, r)$ additionally flags whether justification collapses to zero for at least one essay when rubric-linked evidence is removed.

Combining sensitivity and soft dependence yields descriptive rubric weights:

$$w(t, r) = \mathrm{normalize}(\mathrm{Var}(t, r) + \mathrm{SoftDep}(t, r)),$$

$$\sum_r w(t, r) = 1$$

These weights are not used to recompute or adjust the tutor's original holistic score.

Explanatory Final Score. The weighted rubric contribution is:

$$C(t, e, r) = w(t, r) \cdot p(t, e, r)$$

The normalized EFS aggregates these contributions:

$$\mathrm{EFS}_{\mathrm{norm}}(t, e) = \sum_r C(t, e, r)$$

Rescaling to the full mark range gives:

$$\mathrm{EFS}_{40}(t, e) = 40 \times \mathrm{EFS}_{\mathrm{norm}}(t, e)$$

For Tutor 1's evaluation of Essay 2, this yields $\mathrm{EFS}_{40} = 36.6$, indicating that 36.6 of 40 marks are explicitly supported by rubric-linked evidence.

Worked Example. Explained marks per rubric are:

$$ExplainedMarks_r = p_r \times M_r$$

Aggregating with rubric weights yields:

$$\mathrm{EFS}_{\mathrm{raw}} = \sum_r w_r \times ExplainedMarks_r = 3.55$$

Normalizing and rescaling:

$$\mathrm{EFS}_{\mathrm{norm}} = 0.916 \qquad \mathrm{EFS}_{40} = 36.6$$

A low EFS indicates that a substantial portion of the tutor's judgment is not explicitly grounded in rubric-linked evidence, suggesting either that the tutor relied on holistic impression without citing supporting spans, or that the explanation components provided are structurally insufficient to reconstruct the grading decision. For feedback system designers, this signals that the explanation accompanying a grade may not be reliable enough to guide student revision.

5 Explainability Analysis

5.1 Aggregate Tutor Style via Rubric Emphasis

To characterize how tutor explanations diverge despite a shared rubric schema, we measure how explanatory evidence is distributed across rubric dimensions. For each tutor t and rubric r, rubric emphasis is defined as the average number of evidence spans per essay:

$$E(t, r) = \frac{1}{|E_t|} \sum_{e \in E_t} |S_{t,e,r}|,$$

where $S_{t,e,r}$ denotes the set of spans linked to rubric r in essay e, and E_t is the set of essays graded by tutor t.

Clear tutor-specific patterns emerge. Tutors with balanced preferences (e.g., Tutor 4) show relatively uniform explanatory emphasis across narrative rubrics, whereas tutors with focused preferences (e.g., Tutor 2) emphasize *Theme Depth* and *Characters*, closely aligning with declared rubric weights in the graph metadata.

Importantly, span density does not directly reflect rubric importance: concrete criteria (e.g., *Accuracy*) attract localized evidence, while abstract criteria rely on fewer, global references. Rubric emphasis therefore captures *how* tutors justify judgments rather than *how much* they value a criterion, providing a behaviorally grounded signal of tutor-specific explanation style.

5.2 Aggregate Rubric Sensitivity Across Tutors

While rubric emphasis captures *where* tutors focus their explanations, sensitivity analysis reveals *which* rubrics actually drive variation across submissions. For each tutor t and rubric r, we compute the standard deviation of baseline proxy justification scores across essays:

$$\sigma(t, r) = \mathrm{StdDev}_e(p(t, e, r))$$

Higher values indicate discriminative rubrics; lower values indicate stable constraints.

Opening reaches $\sigma = 1.00$ for Tutor 1 and *Plot* averages $\sigma = 0.74$ across tutors, indicating that narrative dimensions consistently play a discriminative role in tutor explanations. In contrast, *Accuracy* ($\sigma = 0.17$) and *Control Register* ($\sigma = 0.19$) exhibit consistently low sensitivity, functioning as stable floor constraints regardless of essay quality. Sensitivity remains tutor-dependent despite the shared rubric schema: *Creativity* is highly discriminative for some tutors ($\sigma = 0.81$, Tutor 3) but near-stable for others ($\sigma = 0.09$, Tutor 5), while *Theme Depth* shows moderate variance across most tutors ($\sigma = 0.52$–0.67), functioning as a shared evaluative anchor.

Together with rubric emphasis, sensitivity analysis disentangles *where* tutors focus their explanations from *which* criteria actively differentiate essays, two dimensions of explanation behavior that scores alone cannot reveal.

5.3 Tutor Variance and Structural Motivation

Before analyzing explanation structure, we first test whether evaluator identity contributes measurable variation to grading outcomes. If tutors behaved interchangeably after controlling for essay quality, modeling tutor-specific reasoning would be unnecessary. To evaluate this, we perform a two-way analysis of variance (ANOVA) that decomposes score variance into essay-level and tutor-level components.

Formally, we model the assigned score $y_{i,j}$ for essay i graded by tutor j as

$$y_{i,j} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j},$$

where μ denotes the global mean score, α_i captures the effect of essay i (representing essay quality differences), β_j captures the tutor-specific grading effect, and $\varepsilon_{i,j}$ represents residual variation.

As expected, essay identity strongly predicts score ($F_{49,124} = 13.32$, $p < 0.001$), confirming substantial variation in essay quality. More importantly, tutor identity also significantly affects scores ($F_{9,124} = 6.31$, $p < 0.001$). The corresponding effect size ($\eta^2 = 0.068$) indicates that approximately

6.8% of total score variance is attributable to evaluator identity after controlling for essay differences.

Although essay quality explains the majority of grading variance, the non-trivial contribution of tutor identity demonstrates structured evaluator-dependent behavior. This result shows that tutors are not interchangeable graders and motivates modeling tutor-specific explanation structure.

Source	Sum Sq	df	F	p-value
Essay ID	11792.67	49	13.32	4.68×10^{-31}
Tutor ID	1026.50	9	6.31	2.51×10^{-7}
Residual	2240.30	124	–	–

Table 2: Two-way ANOVA decomposition of score variance.

5.4 Baseline Modeling via Additive Span Regression

To evaluate whether rubric evidence can explain scores through simple additive relationships, we compare our structural framework against a linear regression baseline that predicts total score from rubric-linked span counts.

The regression model explains only a small fraction of grading variance ($R^2 = 0.160$, adjusted $R^2 = 0.116$), indicating weak explanatory power. Among all rubric dimensions, only *Description* ($p = 0.032$) and *Accuracy* ($p = 0.003$) emerge as statistically significant predictors of score. In contrast, several core narrative dimensions, including *Characters* ($p = 0.711$), *Creativity* ($p = 0.870$), and *Plot* ($p = 0.208$), show no reliable linear relationship with the final grade.

However, our structural analysis reveals substantial variability in these same dimensions. For example, the Rubric Sensitivity metric shows high variance for *Characters* ($\sigma = 0.80$) and *Theme Depth* ($\sigma = 0.67$), indicating that these criteria play discriminative roles across tutors and essays despite lacking additive span–score relationships.

This contrast highlights a limitation of additive evidence models: linear regression assumes that grading decisions arise from symmetric accumulation of evidence counts, whereas human evaluation reflects structured reasoning in which rubric dimensions interact and play different justificatory roles. Consistent with this, our additive span baseline explains only a small portion of grading variance ($R^2 \approx 0.16$). Because our setting contains no predictive model, gradient-based and perturbation-based attribution methods such as LIME, SHAP, and ERASER are inapplicable by definition. The additive span regression therefore serves as the strongest applicable comparator and its failure confirms that explanation structure in human evaluation cannot be captured through simple evidence accumulation alone, motivating explicit structural modeling.

5.5 Unsupervised Discovery of Tutor Grading Styles

To examine whether tutors exhibit systematic grading styles, we cluster tutors based on their rubric attention profiles. Each tutor t is represented as a vector of average rubric-linked span counts across rubric dimensions:

$$v_t = [s_{t,r_1}, s_{t,r_2}, \ldots, s_{t,r_{10}}]$$

where $s_{t,r}$ denotes the average number of spans cited by tutor t for rubric r per essay.

Hierarchical clustering using Euclidean distance reveals two distinct tutor groups (Figure 2). The separation is primarily driven by differences in attention to surface-level criteria.

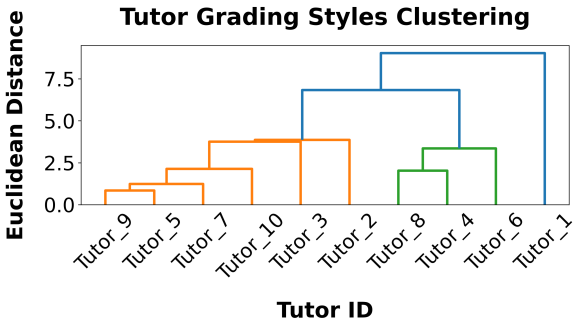

Figure 2: Hierarchical clustering of tutor grading styles based on rubric attention vectors.

Cluster 2 shows higher emphasis on *Accuracy*, *Language_Style*, and *Control_Register* (e.g., 1.68 vs. 0.68 spans per essay for Accuracy). In contrast, Cluster 1 distributes attention more evenly across narrative dimensions such as *Description* and *Theme_Depth*. These patterns correspond to two grading profiles: a *surface-focused* style emphasizing mechanical correctness and a *narrative-focused* style prioritizing interpretive criteria. Notably, these groupings emerge without supervision, indicating structured variation in how tutors justify their evaluations.

To assess robustness, we performed bootstrap resampling of essays (1000 iterations) and recomputed the clustering. Cluster agreement with the

original assignment was measured using the Adjusted Rand Index (ARI). The mean ARI was 0.73 (Figure 3), indicating moderate-to-strong stability and consistent recovery of the original clustering structure.

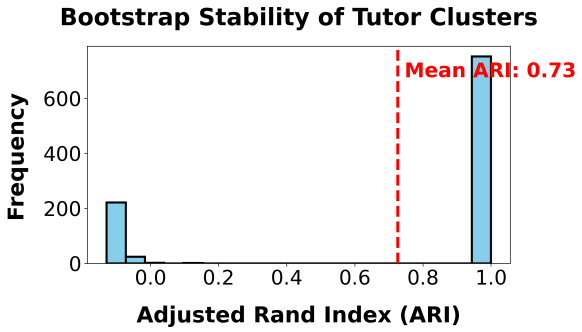

Figure 3: Bootstrap stability of tutor clusters measured using the Adjusted Rand Index (ARI).

Together, these results show that tutor grading styles form reproducible structural patterns rather than random variation. This finding supports our central premise that explanation faithfulness in assessment settings must be evaluated relative to tutor-specific reasoning behavior rather than assuming a uniform grading policy.

5.6 Structural Ablation

To evaluate which rubric dimensions act as structural anchors in tutor explanations, we perform a counterfactual structural ablation analysis. For each tutor, we first compute the baseline explanatory reconstruction score (EFS). We then remove one rubric node from the knowledge graph, recompute the EFS, and measure the percentage drop relative to the baseline. This procedure quantifies the *justificatory load* of each rubric dimension within the explanation structure.

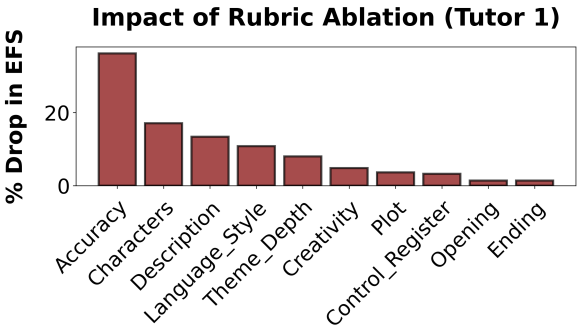

Figure 4: Structural ablation impact for Tutor 1.

Figure 4 illustrates the ablation impact for Tutor 1. Removing *Accuracy* produces the largest collapse in explanatory reconstruction (36.2%), indicating that a substantial portion of the tutor's justification depends on surface-level correctness. Secondary contributions arise from *Characters* (17.1%), *Description* (13.4%), and *Language Style* (10.8%), while dimensions such as *Opening* and *Ending* produce negligible structural impact ($\approx 1.4\%$).

To understand global patterns across evaluators, we repeat the ablation procedure for all tutors and visualize the distribution of impacts per rubric(Figure 5). The results reveal a hierarchical justificatory structure. Narrative dimensions consistently act as structural anchors, with *Description* producing the largest average collapse (33.65%), followed by *Theme Depth* (23.21%) and *Characters* (19.52%).

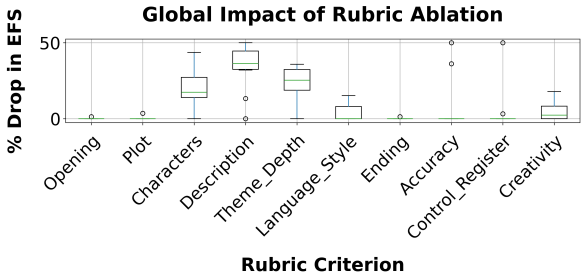

Figure 5: Distribution of structural ablation impact across tutors.

In contrast, surface-level criteria such as *Accuracy* show moderate mean impact (8.62%) but extremely high variability ($\sigma = 18.46$, maximum 50%), indicating that they function as dominant anchors for some tutors but play little role for others. Finally, rubrics such as *Opening*, *Ending*, and *Plot* exhibit uniformly minimal impact (typically $< 1\%$), suggesting limited justificatory centrality.

Together, these results demonstrate that tutor explanations exhibit both shared hierarchical structure and strong evaluator-specific dependencies. Certain rubric dimensions function as global justificatory anchors, whereas others act as divergence markers that differentiate tutor reasoning styles.

5.7 Counterfactual Validation of Structural Hierarchy

To test whether the Explanatory Final Score (EFS) captures meaningful justificatory structure rather than acting as a simple aggregation, we perform a counterfactual rubric-removal validation. Based on the global ablation analysis, rubrics are partitioned into *high-load* (Description, Theme Depth,

Characters) and *low-load* (Opening, Ending, Plot) groups.

For each tutor–essay pair, we measure the change in EFS (ΔEFS) when individual rubrics are removed from the explanation graph. If EFS reflects hierarchical justificatory dependence, removal of structurally central rubrics should produce substantially larger degradation than removal of peripheral ones.

Consistent with this expectation, removing high-load rubrics produces large collapses in explanatory reconstruction (mean $\Delta\text{EFS} \approx 25\%$), whereas removal of low-load rubrics has negligible effect (mean $\Delta\text{EFS} < 1\%$). This difference is highly significant ($p < 10^{-12}$) with a large effect size. This result demonstrates that EFS behaves as a structural sensitivity metric rather than a descriptive aggregate, as it responds asymmetrically to counterfactual removal of justificatory components.

6 Conclusion

This paper examined explanation faithfulness in settings where decisions are produced by human evaluators rather than learned models. By representing tutor grading behavior as explicit evidence–rubric–score relations in a knowledge graph, we treat explanations as structured behavioral traces rather than post-hoc rationalizations. This representation enables analysis of explanation properties such as emphasis, alignment, consistency, and dependence without assuming an underlying predictive decision function. Our analyses show that human grading decisions are holistic, while explanations exhibit stable tutor-specific structure. Evidence spans function primarily as justificatory artifacts rather than causal determinants of scores, revealing a limitation of score-perturbation–based faithfulness metrics in human evaluation settings.

These findings suggest that faithfulness for human-aligned NLP systems should be evaluated structurally and behaviorally rather than solely through outcome sensitivity. LLMs optimised to produce globally plausible explanations may generate rubric-grounded justifications that nonetheless fail to reflect the structural reasoning of a specific evaluator. Future work should apply the EFS diagnostic to audit LLM-generated feedback against human grading traces, providing a structural ground truth for alignment before deployment in real assessment pipelines. Although demonstrated on essay scoring, the framework generalises to other expert evaluation settings such as medical diagnosis and code review.

For feedback system designers, these findings suggest a concrete design principle: explanation components should be evaluated not by their plausibility but by their structural alignment with scoring behavior. A system whose explanations are rubric-grounded but evidence-light may still produce reliable feedback, whereas a system whose explanations cite abundant evidence but lack rubric structure may mislead learners about what actually determined their grade. EFS provides a diagnostic for making this distinction, enabling designers to audit explanation reliability before deployment.

Limitations

Our analysis is conducted on a controlled dataset of 100 essays evaluated by 10 tutors (400 grading traces). While sufficient for explainability diagnostics, the dataset remains smaller than large-scale AES benchmarks, though comparable to prior structural analyses of human assessment behavior (De Vrindt et al., 2025). Future work should evaluate generalizability across larger datasets, diverse institutions, and grading tasks.

Second, the study focuses on narrative essay grading under a fixed rubric schema. Although the proposed knowledge graph representation is task-agnostic, empirical validation on other expert evaluation settings, such as peer review, code assessment, medical decision support, remains future work.

Third, our analyses examine structural properties of explanations rather than causal determinants of grading decisions. The counterfactual experiments evaluate how explanation structure responds to evidence removal while keeping scores fixed, and therefore should be interpreted as diagnostics of justificatory dependence rather than causal mechanisms of human judgment.

Finally, the Explanatory Final Score (EFS) is introduced as a reconstruction metric that measures how much of a tutor's judgment is supported by explicit rubric-linked evidence. While our counterfactual validation shows that EFS captures hierarchical explanation structure, further work is needed to evaluate how well this metric generalizes across datasets and how it should be interpreted in real-world assessment systems.

Ethics Statement

This work analyzes human grading explanations in educational assessment. All grading data used in this study consists of tutor annotations collected under controlled research conditions. No personally identifiable information about students or tutors is disclosed, and all evaluators participated with awareness that their grading traces would be used for research purposes.

Our framework is intended as a diagnostic tool for auditing explanation structure rather than as an automated grading or evaluator ranking system. We caution against deploying EFS or related metrics as performance indicators for individual tutors without appropriate human oversight, as doing so risks reducing complex evaluative judgment to a single numeric signal.

Finally, while our knowledge graph representation could in principle be applied to align LLM-generated feedback with specific tutor styles, such applications should be designed with care to avoid reinforcing evaluator biases or creating systems that replicate inequitable grading patterns at scale.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of ACL*.

Michiel De Vrindt, Anaïs Tack, Renske Bouwer, Wim Van den Noortgate, and Marije Lesterhuis. 2025. Explaining holistic essay scores in comparative judgment assessments by predicting scores on rubrics. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, and Byron C. Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of ACL*.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Fei Dong and Yue Zhang. 2017. Automated essay scoring using transformer models. In *Proceedings of IJCAI*.

Xin Guo, Yun Liu, and Wenwu Zhang. 2022. Graphxai: Towards explainable graph neural networks. In *Proceedings of NeurIPS*.

Peter Hase, Mohit Bansal, Been Kim, and Samuel Gershman. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367. Association for Computational Linguistics.

Yaqiong He, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2022. Automated chinese essay scoring from multiple traits. In *Proceedings of COLING 2022*.

Filip Ilievski, Daniel Garijo, and Hans Chalupsky. 2021. Facilitating explainable ai with knowledge graphs. *Semantic Web Journal*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of ACL*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of NAACL*.

Zixuan Ke and Hwee Tou Ng. 2019. Enhancing automated essay scoring with rubric-guided learning. In *Proceedings of ACL*.

Alexander Kwako and Christopher Ormerod. 2024. Can language models guess your identity? analyzing demographic biases in AI essay scoring. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 78–86.

Qi Liu, Zhenya Huang, and Yu Yin. 2019. Knowledge tracing: A review of recent developments. *IEEE Transactions on Learning Technologies*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

Nitin Madnani, Aoife Cahill, and Brian Riordan. 2018. Automated scoring: Beyond the numbers. *Educational Measurement: Issues and Practice*.

Samuel Messick. 1989. Validity. In Robert L. Linn, editor, *Educational Measurement*, pages 13–103. American Council on Education.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Richard E Nisbett and Timothy D Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Nils-Jonathan Schaller, Yuning Ding, Andrea Horbach, Jennifer Meyer, and Thorben Jansen. 2024. Fairness in automated essay scoring: A comparative analysis of algorithms on german learner essays from secondary education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 210–221.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of ACL*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. volume 43, pages 619–659. MIT Press.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of EMNLP*.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box. In *Proceedings of the International Conference on Artificial Intelligence, Ethics, and Society*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20. Association for Computational Linguistics.

Yuhao Zhang, Jin Zhao, and Xiaoyong Liu. 2020. Feedback generation for automated essay scoring. In *Proceedings of EMNLP*.