From Scoring to Explanations: Evaluating SHAP and LLM Rationales for Rubric-based Teaching Quality Assessment

Anonymous ACL submission

Abstract

Automated scoring models are increasingly used to assign rubric-based quality ratings to complex language performances, including classroom transcripts, yet they typically provide little insight into why a particular score is produced. We propose a general framework for sentence-level interpretability of rubric-based scoring that combines model-agnostic Shapley-value attributions with rationales generated by large language models (LLMs). Instantiated on the Quality of Feedback dimension of the CLASS framework using the NCTE corpus, the framework enables systematic comparison of fine-tuned pretrained language models (PLMs) and prompted LLMs on both scoring performance and explanation faithfulness. Across 6k annotated transcript segments, fine-tuned PLMs outperform LLMs in prediction accuracy but exhibit label compression toward mid-scale scores. Deletion-based tests show that SHAP identifies sentences that reliably drive model predictions, producing typically larger and more coherent prediction shifts than LLM-generated rationales. Cross-model analyses further reveal that SHAP attributions transfer robustly across architectures, whereas LLM rationales exert limited and inconsistent influence. Overall, the findings demonstrate that SHAP provides more faithful and transferable explanations for rubric-based scoring, and that the proposed framework offers a principled basis for evaluating both scoring models and their explanations in high-stakes educational settings and other rubric-based language assessment tasks.

1 Introduction

Rubric-based scoring models are increasingly used to automatically evaluate open-ended language tasks, from student essays and peer feedback to clinical notes and classroom transcripts. In these settings, models assign scalar scores on multi-level rubrics that inform teaching, evaluation, and policy decisions, yet most systems provide little insight into why a particular score was assigned. This is especially problematic in high-stakes educational contexts, where stakeholders such as teachers must be able to understand, trust, and contest automated judgments—requirements that are now explicitly reflected in emerging regulatory frameworks such as the EU AI Act (European Parliament and Council of the European Union, 2024). Therefore, a central challenge emerges: **how can we trust rubric-based scores produced by opaque, black-box models such as large language models (LLMs) when their internal decision-making is inaccessible and their explanations may be unfaithful?** Recent work suggests that free-form explanations produced by LLMs can be persuasive without faithfully reflecting the underlying computation (Turpin et al., 2023; Ye and Durrett, 2022), raising thus a critical question for explainable NLP: **which parts of a text truly drive a model's rubric-based score, and how can we evaluate whether an explanation has captured them?**

High-quality feedback is central to teachers' professional growth, yet providing consistent and individualized feedback is resource-intensive and prone to inconsistency. Recent work demonstrates that automated feedback tools can enhance teachers' uptake of student ideas by as much as 24% (Demszky et al., 2024), showing the promise of NLP-based approaches for supporting teacher development. Classroom teaching quality is thus a prototypical example of a high-stakes, rubric-based judgment where opaque scores are insufficient, and explanations are critical.

Automated scoring of teaching quality dimensions has likewise proven feasible (Hou et al., 2024; Fütterer et al., 2026), but scoring alone does not provide insight into the reasoning behind a model's evaluation. Moving from *what* (the score) to *why* (the reasoning) is essential for generating actionable feedback and fostering user trust. However, whereas LLMs can generate rich, sentence-level

rationales, a growing body of work shows that such explanations often fail to reflect the model's actual decision process(Turpin et al., 2023; Ye and Durrett, 2022). Despite the rapid progress of LLMs and transformer-based scoring systems, their decision-making processes often remain opaque.

To bridge this gap, we investigate explainable NLP methods that can reveal which parts of classroom dialogue most strongly influence automated teaching quality assessments. Specifically, we propose a unified framework for sentence-level interpretability of rubric-based teaching quality scores, comparing model-agnostic feature attribution using SHAP (Lundberg and Lee, 2017) with LLM-based reasoning to identify aspects of teacher–student interaction that contribute to high- or low-quality feedback. Our study focuses on the *Quality of Feedback* dimension, evaluating the quality of the feedback given by teachers to their students in the classroom, within the *Instructional Support* domain of the Classroom Assessment Scoring System (CLASS) framework (Pianta et al., 2008), as providing feedback is a core teaching practice and exhibits a balanced label distribution in our dataset.

We evaluate fine-tuned transformer-based models and LLMs on the NCTE dataset (Demszky and Hill, 2023), containing elementary mathematics classroom transcripts annotated by expert observers. Beyond comparing model performance, our work examines the faithfulness and consistency of different explanation methods by systematically removing sentences highlighted as important, either by SHAP or by LLM-generated rationales, and measuring how these removals change predictions. We further introduce a cross-model evaluation protocol where explanations generated for one model family are used to perturb inputs for the other, allowing us to study whether explanations transfer across architectures or remain model-specific.

Our work contributes to the growing body of research on explainable NLP and reliable LLM rationales in education and other rubric-based assessment settings by:

1. Proposing a general framework for sentence-level interpretability of rubric-based scoring models that combines model-agnostic Shapley-value attributions with LLM-generated rationales, instantiated on automated teaching quality assessment.

2. Comparing specialized fine-tuned models and LLM prompting for teaching quality scoring.

3. Evaluating the faithfulness of SHAP and LLM-based explanations through deletion-based tests, assessing how influential the identified sentences truly are for model predictions.

4. Introducing cross-model consistency analyzes, where LLM-selected sentences are removed and evaluated with fine-tuned models (and vice versa), to probe the alignment of explanation methods across architectures.

5. Discussing the implications of our findings for the design of transparent, actionable teacher feedback tools and, more broadly, for the use of LLM rationales as explanations in rubric-based educational assessments.

We address the following research questions (RQs):

- **RQ1:** How do fine-tuned transformer-based pretrained language models (PLMs) compare to prompted LLMs in predicting rubric-based teaching quality scores on the Quality of Feedback dimension?

- **RQ2:** How faithful and reliable are SHAP- and LLM-based sentence-level explanations in identifying influential parts of a classroom transcript, as measured by deletion-based changes in predictions?

- **RQ3:** To what extent do explanations transfer across model types, i.e., does removing sentences identified by one model meaningfully affect the predictions of another?

2 Related Work

2.1 Explainability Methods for NLP Models

Explainable NLP methods aim to identify input components that most strongly influence model predictions, a crucial requirement in educational contexts where transparency is essential. Model-agnostic approaches, e.g., LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), provide local feature attributions by approximating complex classifiers or computing Shapley values. Beyond NLP, SHAP has also been applied in other domains to improve model interpretability, including work that combines SHAP explanations with LLM-generated descriptions to enhance human-understandable rationales (Khediri et al., 2024). In

our work, we use SHAP at a sentence embedding level within a hierarchical PLM architecture, treating sentences as features to obtain document-level attributions that are both computationally tractable and directly actionable for feedback.

For neural text models, attention-based interpretations have been debated due to concerns about whether attention weights reflect causal importance (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019). To address these limitations, research increasingly distinguishes plausibility from faithfulness (Jacovi and Goldberg, 2020). Deletion- and perturbation-based evaluation, i.e., removing influential input elements and observing prediction changes, provides a more direct measure of explanation faithfulness (DeYoung et al., 2020). We adopt this perspective and extend it to a cross-model setting: explanations are evaluated not only with respect to their source model, but also by measuring how they perturb predictions of alternative architectures. Furthermore, our work contributes by applying sentence-level SHAP in a hierarchical PLM setting and evaluating its faithfulness through systematic deletion tests.

2.2 Faithfulness and Reliability of Model Explanations in LLMs

LLMs are increasingly producing free-form rationales and structured justifications for their predictions. However, a growing body of work suggests that these explanations may not accurately reflect the underlying computation. Chain-of-thought rationales can be unfaithful even while producing correct answers (Turpin et al., 2023), and explanations in few-shot prompts frequently exhibit inconsistencies or hallucinations (Ye and Durrett, 2022). Structured prompting can improve reliability, but challenges remain (Ayala and Bechard, 2024).

Recent approaches have proposed adapting Shapley-based methods to LLMs (Mohammadi, 2024), although computational constraints limit their practical use. Despite the increased adoption of LLMs in educational settings, the faithfulness of their rationales has not been systematically evaluated relative to established attribution methods such as SHAP, especially on long, naturalistic transcripts and multi-level rubrics. Our work addresses this gap by comparing LLM-generated sentence rankings against PLM-based SHAP attributions using matched deletion-based faithfulness tests and cross-model robustness analysis, thereby providing empirical evidence on when LLM rationales align with model behavior and when they diverge.

2.3 Automated Scoring and Teacher Feedback in Educational Settings

Recent efforts have explored automated approaches to analyze classroom instruction and support teacher learning. The NCTE dataset (Demszky and Hill, 2023) has facilitated large-scale research on evaluating teacher uptake of student ideas (Demszky et al., 2024), and on leveraging models such as ChatGPT for instructional scoring and feedback (Wang and Demszky, 2023). More broadly, researchers have begun to validate automated assessments of teaching quality dimensions using multimodal approaches, including audio, video, and text features, combining embeddings with LLM-generated scores (Hou et al., 2024, 2025a; Fütterer et al., 2026). In particular, Hou et al. (Hou et al., 2025b) evaluate LLM-based multimodal models for classroom assessment by comparing their predictions against human annotations, providing evidence that such models can approximate human judgments of teaching quality.

In parallel, NLP has long been applied to educational assessment tasks, including essay scoring and discussion analysis. Recent work shows that LLMs and PLMs can approximate human rubric-based judgments across multiple writing dimensions (Seßler et al., 2025), while neural models have been used to assess the quality of classroom discussions or participation (Tran et al., 2023). LLMs are increasingly used to provide pedagogically aligned feedback to learners (Meyer et al., 2024). Our work situates itself within this line of research, but shifts the focus from *predictive performance* to *explanation quality*. More specifically, we compare fine-tuned PLMs and instruction-tuned LLMs for scoring a specific CLASS dimension (i.e., Quality of Feedback). More importantly, we introduce a general framework for evaluating sentence-level explanations for rubric-based scores.

3 Methods

We cast teaching quality assessment as a general rubric-based text scoring problem with model-agnostic sentence-level explanations and instantiate this framework using both PLMs and instruction-tuned LLMs.

3.1 Dataset

To instantiate our interpretability framework in an educational setting, we use the NCTE

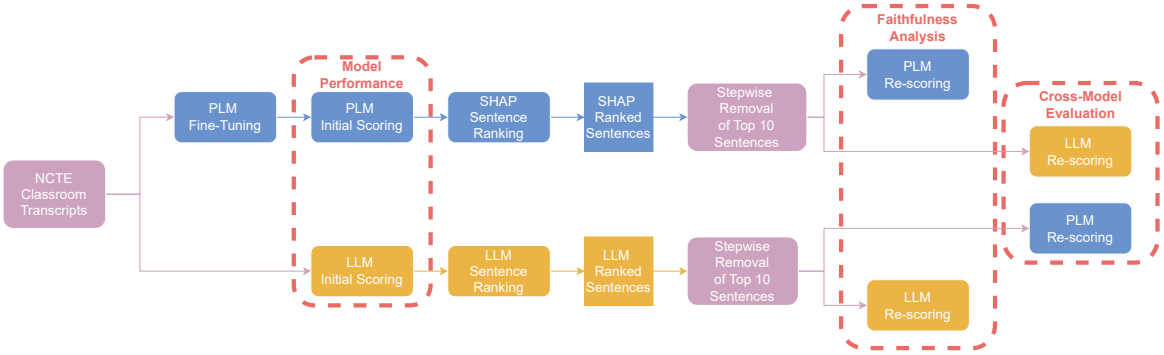

Figure 1: Overview of the proposed framework. The top branch (blue) shows the PLM pipeline, including fine-tuning, scoring, SHAP-based sentence ranking, and sentence removal with re-scoring. The bottom branch (yellow) depicts the corresponding LLM pipeline with prompted scoring and ranking. The three experimental settings are indicated by dotted red boxes.

dataset (Demszky and Hill, 2023), comprising over 1,600 transcripts of 45–60 minute elementary mathematics lessons. Lessons are segmented into 15-minute units, producing 6,005 segments annotated using the CLASS framework (Pianta et al., 2008), a tool that measures the quality of interactions between teachers and students to assess teaching. CLASS comprises three domains—*Emotional Support*, *Classroom Management*, and *Instructional Support*—covering thirteen dimensions. We focus on the *Quality of Feedback* (QoF) dimension within the Instructional Support domain.

QoF measures the extent to which teachers provide meaningful feedback, scaffold student thinking, prompt metacognition, elaborate on student responses, and clarify misunderstandings. Each segment receives a QoF rating on a 1–7 scale, where 1–2 indicates low-quality or absent feedback, 3–5 reflects moderate or inconsistent feedback, and 6–7 represents consistently high-quality feedback.

QoF is chosen for three reasons: (1) compared to other CLASS dimensions, its label distribution is less skewed (mean of 4.21, standard deviation of 1.13, with 81% of ratings in the 3–5 range), making it more suitable for supervised learning; (2) QoF is a core instructional practice strongly linked to student learning gains (Hattie and Timperley, 2007); and (3) many indicators of feedback quality (e.g., probing questions, elaborations, scaffolding) manifest clearly in text transcripts (Demszky and Hill, 2023), making QoF particularly appropriate for sentence-level interpretability analysis.

We adopt an 80/20 data split, resulting in 4.775 segments for training and 1,230 for testing. Transcripts belonging to the same class were kept in the same split, and we stratified the data based on label distribution. The training split is used exclusively for fine-tuning PLMs, whereas the test split is used to evaluate both PLMs and LLMs, as well as to conduct all interpretability experiments. Within the test split, 29.3% of the sentences are student utterances, 69.9% are teacher utterances, and 0.8% were utterances that could not be assigned to a speaker. The dataset provides one expert annotation per teaching quality dimension, precluding the assessment of interrater agreement, which we treat as ground truth for both scoring and evaluation. The dataset does not include human-annotated sentence-level rationales or evidence.

3.2 Models

Within our framework, we instantiate the scoring model f using two classes of architectures: PLMs and instruction-tuned LLMs. We compare these two families because PLMs represent the standard supervised approach for rubric-based scoring, requiring task-specific fine-tuning, whereas LLMs offer strong zero- or few-shot capabilities without additional training. This contrast allows us to examine differences in scoring performance but also how explanation methods behave across models with fundamentally different training paradigms, capacities, and levels of transparency.

Pretrained Language Models. We fine-tuned *BERT* (Devlin et al., 2019), *ALBERT* (Lan et al., 2020), *RoBERTa* (Liu et al., 2019), and *DeBERTa V3* (He et al., 2023), using both base and large variants. PLMs operate on transcript segments at the sentence level: each sentence was encoded using the model's `[CLS]` representation, and a trainable attention layer computed an attention-

weighted document embedding. A linear regression head predicted a single scalar QoF score. We used a maximum sentence length of 128 tokens and a maximum of 263 sentences per document, which encompasses 98% of the sentence lengths and 90% of the document lengths without truncation. During fine-tuning, models were optimized using mean squared error; additional hyperparameters are listed in App. A. After fine-tuning, we applied SHAP to the document-level regression output, treating each sentence embedding as a separate feature. SHAP returned one Shapley value per sentence, representing its estimated contribution to the predicted score.

Large Language Models. For LLMs, we used instruction-tuned variants of Llama 3.1 (8B, 70B) (Grattafiori et al., 2024), Mixtral (8×7B, 8×22B) (Jiang et al., 2024), Qwen 3 (4B, 30B, 235B) (Yang et al., 2025), and Mistral (Small, Small 24B) (Jiang et al., 2023). Only open-source LLMs were selected, as they can be deployed locally and therefore mitigate the privacy concerns associated with datasets such as classroom transcripts. LLMs performed two tasks: (1) QoF scoring using a few-shot prompt introducing the task and showing examples, and (2) sentence ranking using a zero-shot prompt. For ranking, we provided transcripts segmented and numbered by sentence and requested a list of ten sentence indices corresponding to the most influential sentences. This ensured alignment with the PLM sentence boundaries and prevented models from altering sentence content. Some outputs contained invalid indices or fewer than ten items; in such cases, the system retried up to ten times. Prompts are shown in App. B. All models were evaluated with deterministic decoding, and all local inference used 4-bit `nf4` quantization via BitsAndBytes.

3.3 Interpretability Methods

We view rubric-based teaching quality assessment as a complex text scoring problem. More specifically, given an input transcript x and a model f, the model outputs a scalar score $f(x)$ on a fixed rubric. An explanation method E maps x and f to a ranked list of textual units (here, sentences) that are claimed to be most influential for the score. To evaluate the faithfulness of such explanations, we adopt a deletion-based protocol: we progressively remove the top-k units selected by E (with $k = 10$ in all experiments), recompute $f(x)$, and measure the change in predictions. Larger changes indicate that the explanation has successfully identified text that the model relies on. To study cross-model consistency, we extend this protocol to pairs of models f and g, apply explanations obtained from one model to perturb the other and compare the resulting prediction shifts. In this paper, we instantiate f as fine-tuned PLMs and g as instruction-tuned LLMs, with explanations E provided either by sentence-level SHAP attributions or by LLM-generated sentence rankings.

3.4 Experiments

To evaluate our proposed sentence-level interpretability evaluation framework, we conduct three sets of experiments. The framework combines sentence-level explanation generation (via SHAP or LLM-based ranking) with faithfulness testing through sentence deletion and robustness analysis via cross-model transfer. Each experiment targets one aspect of this framework across model families. **(1) Model performance for scoring (RQ1):** We evaluate PLMs and LLMs on the transcript segment scoring task. **(2) Faithfulness analysis (RQ2):** We compute score differences (Δ) after each single-sentence deletion to assess whether a model's own explanations meaningfully affect its predictions. All models employ the same sentence-splitting procedure, and deletions are made in accordance with the ranking order. When fewer than ten sentences exist or when deletions result in an empty transcript, models are prompted with an empty input. This only happened for 17 (1.3%) out of 1,230 transcript segments. In addition to deletion-based evaluation, we also quantify the alignment between different explanation methods. For each transcript, we compute the Jaccard similarity between the sets of top-k sentences (here $k = 10$) selected by SHAP and by each LLM, and, where applicable, the Spearman rank correlation between their full sentence rankings. These metrics capture how often explanation methods highlight the same textual evidence, independently of their causal impact on predictions. **(3) Cross-model evaluation (RQ3):** We examine whether sentence-level explanations generalize across models by applying sentence deletion based on SHAP and LLM rankings to the opposite model family. We select six representative models for this analysis: BERT large, DeBERTa V3 large, and ALBERT base (PLMs) and Qwen 3 235B, Mistral Small, and Llama 3 8B (LLMs), which reflect the best-, worst-, and mid-performing models

from prior steps, while also maintaining diversity in model size.

We report mean absolute error (MAE) and mean squared error (MSE). For constant baselines, we use the median prediction for MAE and the mean prediction for MSE, corresponding to the respective risk minimizers. All experiments use identical sentence segmentation, deletion rules, and deterministic LLM decoding.

4 Results And Discussion

4.1 Model Performance for Scoring

Table 1 reports Mean Absolute Error (MAE) and Mean Squared Error (MSE) for all models. For PLMs, results are shown before and after fine-tuning, whereas for LLMs, a single score is reported based on the few-shot scoring prompt (see App. B).

Model	MAE	MSE	MAE	MSE
Constant Baseline	0.96	1.35	–	–
	Non-Fine-Tuned		**Fine-Tuned**	
ALBERT base	4.34	20.12	0.98	1.34
ALBERT large	4.22	19.10	0.99	1.45
BERT base	4.58	22.29	0.98	1.36
BERT large	4.45	21.07	0.97	1.34
DeBERTaV3 base	4.08	17.97	1.00	1.56
DeBERTaV3 large	3.66	14.65	**0.96**	**1.31**
RoBERTa base	4.33	20.08	0.97	1.34
RoBERTa large	3.27	12.01	0.97	1.37
Llama 3.1 8B Instruct	1.63	3.98	–	–
Llama 3.1 70B Instruct	1.98	5.32	–	–
Mistral Small Instruct	**1.02**	**1.78**	–	–
Mistral Small 24B Instruct	1.75	4.28	–	–
Mixtral 8x7B Instruct	1.39	2.85	–	–
Mixtral 8x22B Instruct	1.21	2.41	–	–
Qwen3 4B Instruct	1.18	2.29	–	–
Qwen3 30B A3B Instruct	1.56	3.59	–	–
Qwen3 235B A22B Instruct	1.67	4.16	–	–

Table 1: Mean Absolute Error (MAE) and Mean Squared Error (MSE) for PLMs and LLMs. PLM results are reported separately for non-fine-tuned and fine-tuned models, while LLM results correspond to prompted inference without task-specific training.

For PLMs, fine-tuning yields a substantial and expected performance improvement. While non-fine-tuned models show MAEs above 4.0, all fine-tuned PLMs achieve uniformly low errors (MAE 0.96–1.00; MSE 1.31–1.56), comparable to the constant baseline. Performance differences among fine-tuned PLMs are minimal ($\Delta = 0.04$ MAE, $\Delta = 0.25$ MSE), indicating comparable behavior across architectures. The best-performing model, DeBERTaV3 large, achieves an MAE of 0.96 and an MSE of 1.31.

Despite their strong numerical performance, fine-tuned PLMs exhibit limited label coverage. For DeBERTaV3 large, the mean predicted score is 4.14 ($\sigma = 0.16$), closely matching the dataset average of 4.22, but predictions never fall below 2.03 or exceed 5.89, meaning that extreme labels (1 and 7) are never predicted (see App. C). This behavior is consistent across all fine-tuned PLMs, with some models (e.g., ALBERT and RoBERTa variants) effectively collapsing to a narrow mid-range of labels (3–5). This pattern is likely driven by strong label imbalance at the extremes of the QoF scale, indicating high sensitivity to the underlying data distribution.

LLMs show weaker overall scoring performance than fine-tuned PLMs. The best-performing LLM, Mistral Small Instruct, achieves an MAE of 1.02 and an MSE of 1.78, which remains higher than both the best PLM and the constant baseline. In contrast to PLMs, performance variability across LLMs is substantially larger ($\Delta = 0.96$ MAE, $\Delta = 3.54$ MSE), reflecting notable differences across models. However, most LLMs produce predictions spanning the full 1–7 score range. For Mistral Small Instruct, the mean predicted score is 4.37 ($\sigma = 0.77$), indicating a much broader dispersion than observed for PLMs.

Overall, these results highlight a trade-off between accuracy and flexibility. Fine-tuned PLMs achieve substantially higher scoring accuracy but are very sensitive to training data distribution, whereas LLMs provide wider score distributions at the cost of reduced accuracy. In educational settings, LLMs offer practical advantages due to their out-of-the-box usability and lack of task-specific training requirements, but incur higher computational costs and lower predictive reliability. In response to RQ1, fine-tuned PLMs outperform prompted LLMs in accuracy, while LLMs better preserve score variability.

4.2 Faithfulness Analysis

The second experiment evaluates the faithfulness of SHAP- and LLM-based sentence importance rankings. For SHAP, the number of selected sentences is deterministically fixed as the minimum of ten and the number of sentences in each transcript segment. In contrast, LLMs show limited controllability: even when prompted to return exactly ten sentences, they produce more or fewer often. Models such as Mixtral 8×7B and Qwen 3 235B deviate most strongly from the expected average of 9.931

Group	Model	$\overline{\Delta}$
PLMs	ALBERT base	0.0219
	ALBERT large	0.0172
	BERT base	0.0256
	BERT large	**0.0329**
	DeBERTaV3 base	0.0242
	DeBERTaV3 large	0.0049
	RoBERTa base	0.0053
	RoBERTa large	0.0119
LLMs	Llama 3.1 8B Instruct	0.0174
	Llama 3.1 70B Instruct	0.0090
	Mistral Small Instruct	0.0033
	Mistral Small 24B Instruct	0.0123
	Mixtral 8x7B Instruct	0.0121
	Mixtral 8x22B Instruct	0.0199
	Qwen3 4B Instruct	0.0211
	Qwen3 30B A3B Instruct	0.0174
	Qwen3 235B A22B Instruct	**0.0388**

Table 2: Average consecutive performance change $\overline{\Delta}$ across sentence removals for PLMs and LLMs.

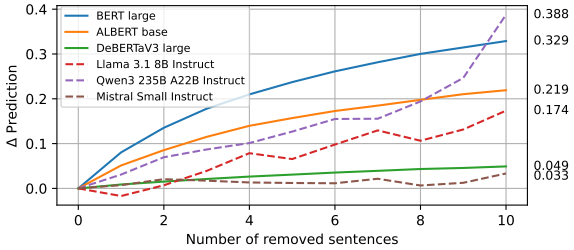

Figure 2: Prediction change Δ under progressive sentence removal for selected PLMs and LLMs. Sentences are chosen using SHAP (solid lines) or LLM-based rankings (dashed lines), and inputs are re-scored by the same model; final changes are shown on the right.

sentences, typically returning fewer sentences despite up to ten retry attempts for malformed outputs. Llama 3.1 8B is the only model that, on average, returns more sentences than requested and comes closest to the target value (see App. D).

These results highlight the inherent variability and limited controllability of LLM outputs, even under structured prompting and multiple retries. This unpredictability represents a practical limitation when deploying LLMs in educational systems, where reliability and strict adherence to output formats are critical. Systems that incorporate LLMs must therefore be explicitly designed to handle such inconsistencies if these models are to be used effectively in real-world educational settings.

We then remove one sentence at a time from the top ten ranked sentences and re-score each transcript segment using the same model that produced the ranking. Table 2 reports the average prediction change after each consecutive removal, denoted as $\overline{\Delta}$. Among PLMs, BERT models are most sensitive to sentence removal, with BERT large exhibiting the highest average change ($\overline{\Delta} = 0.0329$). For LLMs, the largest change is observed for Qwen 3 235B ($\overline{\Delta} = 0.0388$), while all other LLMs show $\overline{\Delta}$ values below 0.02. Overall, this indicates that both model families can identify influential sentences, though PLMs do so more consistently, while LLM behavior is more variable (see Section 4.3).

Beyond faithfulness, we observe largely similar patterns in the types of sentences selected by PLMs and LLMs. Both predominantly prioritize teacher-authored utterances, consistent with the QoF rating dimension: LLMs select teacher utterances in 79.5% of cases and student utterances in 19.5%, while PLMs show a comparable distribution (74.0% teacher, 24.7% student). Despite this similarity, alignment between SHAP- and LLM-based explanations remains consistently low. Across nine LLMs, the mean Jaccard similarity between the top-10 sentence sets is 0.085, corresponding to an overlap of roughly one to two sentences per transcript, while the mean Spearman rank correlation is 0.062, indicating weak agreement in sentence importance ordering. These trends are consistent across model families and sizes, suggesting that explanation alignment is driven primarily by the explanation method rather than model scale.

Fig. 2 illustrates prediction changes under progressive sentence removal for selected PLMs and LLMs. For completeness, App. F reports results for all models, App. G provides representative sentence examples, and App. E contains the full alignment statistics.

Together, these results answer RQ2: SHAP-based sentence rankings are more faithful for PLM scorers, whereas LLM-generated rationales induce smaller and often unstable prediction changes, even for the models that produce them.

4.3 Cross-Model Evaluation

For cross-model evaluation, we select three PLMs and three LLMs representing high, medium, and low faithfulness in the single-model deletion analysis. Specifically, we use **BERT large** (most affected), **ALBERT base** (moderately affected), and **DeBERTaV3 large** (least affected) among PLMs, and **Qwen3 235B**, **Llama 3.1 8B**, and **Mistral Small** as corresponding LLMs. Their sentence-removal trajectories under self-generated explana-

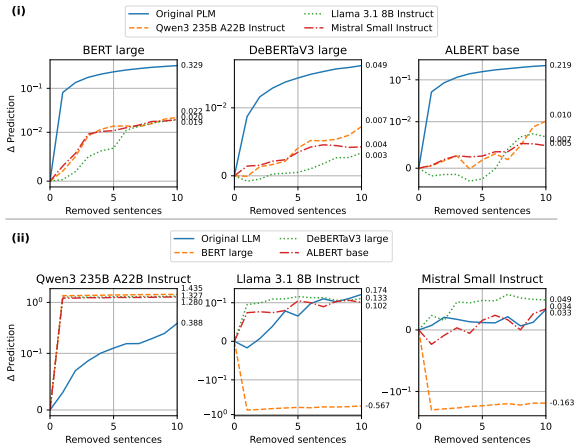

Figure 3: Prediction change Δ under progressive sentence removal for selected PLMs and LLMs. Panel (i) shows PLM re-scoring after removing sentences selected by SHAP (solid) or LLM-based rankings (dashed/dotted); panel (ii) shows LLM re-scoring after removing sentences selected by the LLM itself (solid) or by SHAP from fine-tuned PLMs (dashed/dotted).

tions are shown in Fig. 2, with full results in App. F.

To assess transfer from LLMs to PLMs, we apply LLM-generated sentence rankings to PLMs and re-score the perturbed inputs. Fig. 3 (i) compares these results with the baseline condition where PLMs are perturbed using SHAP-selected sentences. Across all models, removing LLM-ranked sentences produces substantially smaller prediction shifts than removing SHAP-ranked sentences, even for PLMs that are weakly sensitive to SHAP. Moreover, deletion trajectories induced by LLM rationales are often non-monotonic, with predictions fluctuating as additional sentences are removed, indicating limited alignment with the features PLMs rely on. Among LLMs, **Qwen3 235B** shows the strongest cross-model transfer, yielding the largest and most stable perturbations, though still markedly weaker than those induced by SHAP.

We then consider the reverse direction, applying PLM-derived SHAP explanations to LLM scoring. As shown in Fig. 3 (ii), removing the single most influential SHAP-ranked sentence frequently causes a large immediate shift in LLM predictions, followed by stabilization in subsequent deletions. This effect is most pronounced for PLM–LLM pairs that are highly sensitive to sentence removal. The consistent "first-step jump" suggests that SHAP identifies sentence-level features that are relevant not only to PLMs but also to LLMs, despite architectural and training differences.

Overall, these results indicate that PLMs and LLMs rely on different sentence-level evidence. While LLM rationales often surface intuitively relevant content, they do not reliably capture the features driving PLM predictions, regardless of model sensitivity. In contrast, SHAP explanations consistently induce larger, more stable, and more coherent prediction shifts both within PLMs and when transferred to LLMs, demonstrating superior faithfulness for sentence-level interpretability.

This difference is likely influenced by architectural factors: PLMs are trained with sentence-level representations, whereas LLMs operate primarily at the token level. Nevertheless, sentence-ranking attributions remain a useful tool for improving the interpretability of rubric-based scoring, particularly when LLMs are used as classifiers. Addressing RQ3, the cross-model evaluation shows that SHAP explanations generalize across architectures, while LLM rationales transfer poorly and appear unreliable as general-purpose explanations.

5 Conclusion

We propose a general framework for evaluating the truthfulness of sentence-level explanations in rubric-based scoring, systematically contrasting Shapley-value attributions with LLM-generated rationales based on their causal impact on model predictions. Applied to the QoF dimension, the framework shows that fine-tuned PLMs outperform prompted LLMs in scoring accuracy (RQ1), although PLMs exhibit label compression, whereas LLMs provide broader but less precise predictions. Deletion-based tests demonstrate that SHAP explanations are substantially more faithful than LLM rationales (RQ2), and cross-model evaluations reveal that SHAP-selected sentences transfer more robustly across architectures, whereas LLM rationales exert limited influence on PLM predictions (RQ3). Overall, the results suggest that current LLM rationales are unreliable as faithful justifications for rubric-based scores, whereas Shapley-based attributions provide a more stable foundation for transparent and actionable automated assessment. More broadly, our findings suggest that the proposed framework offers a robust and extensible way to evaluate scoring models and their explanations in high-stakes settings such as education and can be readily applied to other rubric-based language assessment tasks (e.g., essay scoring, peer feedback quality, or clinical note evaluation).

Limitations

A primary limitation of this work is the dataset's size and label distribution. Although we selected the least skewed CLASS dimension, only 19% of the labels fall outside the 3–5 range across 6k transcript segments. This imbalance likely contributes to the observed label compression in fine-tuned PLMs, limiting model performance at the extremes of the scale. Future work could explore data augmentation or synthetic data generation to improve coverage of underrepresented classes. In addition, our experiments focus solely on the Quality of Feedback dimension, and it remains unclear how well the findings generalize to other CLASS dimensions, particularly those that are less discourse-driven, such as Productivity within the Classroom Management domain.

Our analysis is further restricted to text-only transcripts. CLASS scoring in practice relies on rich, multimodal cues, including prosody, timing, and visual interactional signals, which are currently ignored. Moreover, each transcript segment is annotated by a single expert, preventing assessment of inter-rater reliability and leaving open the possibility of annotation noise and subjective bias. Whereas Quality of Feedback is primarily teacher-centered, a substantial portion of the transcripts consists of student utterances, which were also frequently selected by the sentence-ranking methods. Future work should investigate how explicitly filtering or modeling student contributions affects both scoring and the interpretability of results.

Finally, the deletion-based faithfulness protocol perturbs the natural discourse structure of classroom interaction and may alter pragmatic meaning and speaker intent, limiting its ability to reflect true causal influence. This disruption may disproportionately affect LLM-based scoring and explanations, as LLMs are trained on coherent sentence generation and strongly rely on intact discourse structure and pragmatic flow. Finally, we also observed notable reliability issues in LLM sentence-ranking outputs, which required strict prompt engineering and external sentence segmentation to enforce structured predictions, thereby constraining the natural expressive capacity of LLM-based explanations. Nevertheless, we observe a clear asymmetry: removing SHAP-selected sentences leads to substantially larger changes in predicted scores than removing LLM-selected sentences. This suggests that the observed differences in faithfulness cannot be attributed solely to discourse disruption, but rather reflect differences in how well each method identifies sentences that are truly influential for the model's predictions.

Ethical considerations

A central ethical concern in automated rubric-based scoring is the risk of bias amplification from training labels. Models trained on human annotations may inherit subjective judgments or structural biases and propagate them at scale (Barocas and Selbst, 2016; Mehrabi et al., 2021). In educational contexts, such effects are particularly concerning, as algorithmic assessment systems can reproduce or exacerbate existing inequalities (Nguyen et al., 2023).

Accordingly, automated scoring systems should not replace human raters in high-stakes settings, where unchecked deployment may create self-reinforcing feedback loops of biased predictions. Instead, these systems should be positioned as self-assessment or decision-support tools that augment, rather than replace, professional judgment, consistent with established principles for ethical and human-centered AI in education (Holmes et al., 2022; Nguyen et al., 2023).

Relatedly, there is a risk that model predictions and explanations may be interpreted as objective ground truth. In educational settings, this is particularly problematic, as model-generated feedback can influence teaching practices, evaluations, and institutional decision-making. This underscores the critical need for transparency: educators and stakeholders must be able to understand how and why a model arrives at a given score to appropriately contextualize and challenge its outputs. To mitigate misuse, it is essential to communicate that both scores and explanations are probabilistic and imperfect, and that interpretability methods are intended to support teacher reflection and informed decision-making rather than to serve as authoritative or definitive assessments.

Privacy and data protection are also key ethical considerations. Although the NCTE transcripts used in this work are anonymized, classroom interactions inherently involve sensitive information about teachers and minors. Any real-world deployment of similar systems must ensure strict safeguards for data security, informed consent, and compliance with regulations governing the processing of educational data.

Furthermore, the use of LLM-generated rationales introduces additional risks. LLMs are known to hallucinate, produce inconsistent outputs, and lack transparent decision processes, which makes their explanations particularly problematic in high-stakes settings such as education. Limited reproducibility further complicates auditing and accountability. These factors reinforce the need for caution when using LLM rationales as justifications for automated judgments, and motivate the emphasis of this work on faithfulness-based evaluation. In particular, explanation methods should be evaluated not only for plausibility but also for their causal impact on model predictions, and systems should avoid presenting unvalidated LLM rationales as authoritative interpretations of teaching practice.

Use of AI Assistance We used AI assistance tools (ChatGPT, and GitHub Copilot) to aid in rewriting code, and paraphrasing. All AI-generated content was thoroughly reviewed and verified by the authors. AI was not used to generate new research ideas or original findings; rather, it served as a support tool to improve clarity, efficiency, and organization. In accordance with ACL guidelines, our use of AI aligns with permitted assistance categories, and we have transparently reported all relevant usage in this paper. While AI contributed to enhancing the quality of the work, no direct research outputs are the result of AI assistance.

References

Orlando Ayala and Patrice Bechard. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238, Mexico City, Mexico. Association for Computational Linguistics.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review*.

Dorottya Demszky and Heather Hill. 2023. The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. 2024. Can automated feedback improve teachers' uptake of student ideas? evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, 46(3):483–505.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

European Parliament and Council of the European Union. 2024. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act). Official Journal of the European Union, OJ L, 2024/1689, 12.7.2024. Entered into force on 1 August 2024.

Tim Fütterer, Ruikun Hou, Babette Bühler, Efe Bozkir, Courtney Bell, Enkelejda Kasneci, Peter Gerjets, and Ulrich Trautwein. 2026. Validating automated assessments of teaching effectiveness using multimodal data. *Learning and Instruction*, 101:102264.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C. Santos, Mercedes T. Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, and Kenneth R. Koedinger. 2022. Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32(3):504–526.

Ruikun Hou, Babette Bühler, Tim Fütterer, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2025a. Multimodal assessment of classroom discourse quality: A text-centered attention-based multi-task learning approach. *arXiv preprint arXiv:2505.07902*.

Ruikun Hou, Tim Fütterer, Babette Bühler, Efe Bozkir, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2024. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and chatgpt. In *Artificial Intelligence in Education*, pages 60–74, Cham. Springer Nature Switzerland.

Ruikun Hou, Tim Fütterer, Babette Bühler, Patrick Schreyer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2025b. LLM-human alignment in evaluating teacher questioning practices: Beyond ratings to explanation. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 239–249, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Abderrazak Khediri, Hamda Slimi, Ayoub Yahiaoui, Makhlouf Derdour, Hakim Bendjenna, and Charaf Eddine Ghenai. 2024. Enhancing machine learning model interpretability in intrusion detection systems through shap explanations and llm-generated descriptions. In *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pages 1–6.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *Preprint*, arXiv:1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Behnam Mohammadi. 2024. Explaining Large Language Models Decisions Using Shapley Values. *arXiv preprint*. ArXiv:2404.01332 [cs].

Andy Nguyen, Ha Ngan Ngo, Yvonne Hong, Belle Dang, and Bich-Phuong Thi Nguyen. 2023. Ethical principles for artificial intelligence in education. *Education and Information Technologies*, 28(4):4221–4241.

Robert C. Pianta, Karen M. La Paro, and Bridget K. Hamre. 2008. *Classroom Assessment Scoring System™: Manual K-3.* Brookes Publishing.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. *Preprint*, arXiv:1602.04938.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, and Lindsay Clare Matsumura. 2023. Utilizing natural language processing for automated assessment of classroom discussion. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 490–496, Cham. Springer Nature Switzerland.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

A Training Hyperparameters

The model was fine-tuned using the Hugging Face `Trainer` API with the following key hyperparameters:

- **Learning rate:** 1×10^{-5}
- **Batch size:** 1 per device (with gradient accumulation of 8 steps)
- **Number of epochs:** 10
- **Weight decay:** 0.01
- **Maximum gradient norm:** 1.0
- **Evaluation strategy:** per epoch
- **Saving strategy:** per epoch (keeping best model based on validation MAE)
- **Learning rate scheduler:** cosine schedule
- **Warmup ratio:** 0.1
- **Mixed precision:** FP16 when available
- **Early stopping:** patience of 3 validation checks

All other hyperparameters were kept at their default values. One single H200 GPU was used and computations took a total of ∼322 hours (∼20 hours for PLMs and ∼302 hours for LLMs). All experiments could have been done in a single A100 GPU, since we used the quantized versions of the models, except for the LLMs Qwen3 235B A22B Instruct, and Mixtral 8×22B Instruct, which require more than 80GB VRAM, even in their 4-bit `nf4` quantized versions.

B Prompts

Fig. 4 shows an example of the few-shot prompt transcript Quality of Feedback scoring, and Fig. 5 show the zero-shot prompt for sentence ranking.

C Prediction Statistics

Table 3 presents the full prediction statistics for PLMs and LLMs. It shows the mean values for prediction, standard deviation, and if any label was not present in the predictions of a specific model.

Model	Mean	Std. Dev.	Labels Missing
Data (Training and Test)	4.22	1.14	–
ALBERT base	4.10	0.28	1, 2, 6, 7
ALBERT large	4.33	0.41	1, 2, 7
RoBERTa base	4.27	0.22	1, 2, 6, 7
RoBERTa large	4.01	0.11	1, 2, 6, 7
BERT base	4.15	0.38	1, 6, 7
BERT large	4.14	0.41	1, 2, 7
DeBERTa V3 base	4.49	0.47	1, 7
DeBERTa V3 large	4.14	0.16	1, 2, 6, 7
Llama 3.1 8B Instruct	4.91	1.60	–
Llama 3.1 70B Instruct	2.95	1.65	–
Mistral Small Instruct	4.37	0.77	1, 7
Mistral Small 24B Instruct	2.55	0.69	6, 7
Mixtral 8x7B Instruct	3.02	0.52	6, 7
Mixtral 8x22B Instruct	4.30	1.19	–
Qwen3 4B Instruct	3.96	1.12	–
Qwen3 30B A3B Instruct	2.92	1.00	6, 7
Qwen3 235B A22B Instruct	2.80	1.13	7

Table 3: Prediction statistics for fine-tuned PLMs and LLMs. We report the mean and standard deviation of predicted QoF scores on the test set, as well as the CLASS labels (1–7) that were never predicted by each model.

You are a rater for classroom transcripts. You are tasked with evaluating the transcripts to rate the dimension 'Quality of Feedback' as one of the dimensions of the domain 'Instructional Support' according to the Classroom Assessment Scoring System (CLASS).
Your goal is to rate the provided transcript by focusing on the teacher's interactions labeled as "Teacher" and the students' responses labeled as "Student".

Your Task:
Please read the following classroom transcript and assign a rating based on the Quality of Feedback dimension.
Quality of Feedback assesses the degree to which feedback expands and extends learning and understanding and encourages student participation. In upper elementary classrooms, significant feedback may also be provided by peers. Regardless of the source, the focus here should be on the nature of the feedback provided and the extent to which it "pushes" learning.

Here are examples for the different score values in the Quality of Feedback dimension:
An Example for score value 1 is {*Transcript segment from training set with score 1*}
An Example for score value 2 is {*Transcript segment from training set with score 2*}
An Example for score value 3 is {*Transcript segment from training set with score 3*}
An Example for score value 4 is {*Transcript segment from training set with score 4*}
An Example for score value 5 is {*Transcript segment from training set with score 5*}
An Example for score value 6 is {*Transcript segment from training set with score 6*}
An Example for score value 7 is {*Transcript segment from training set with score 7*}

You must begin your answer with
'### Rating: <the score on a 1-7 integer scale here>'

Transcript:
{*Transcript segment to score*}

Rating:

Figure 4: Prompts used for few-shot prompting of LLMs for the task of Quality of Feedback scoring of a classroom transcript segment. {*Transcript segment from training set with score 1-7*} represents an actual example from the training set, where the specific score was given to the transcript by an expert annotator, and {*Transcript segment to score*} represents the transcript segment to be scored.

You are an expert in classroom discourse analysis and the CLASS (Classroom Assessment Scoring System) framework. You will receive:

1. A transcript of a classroom interaction.
2. The same transcript divided by numbered sentences, formatted as:
   (1) - <SENTENCE #1>
   (2) - <SENTENCE #2>
   ...
3. A score (1-7) for the "Quality of Feedback" domain according to the CLASS framework, annotated by an expert. 1 indicates very low quality feedback, while 7 indicates very high quality feedback.

Your task:
* Identify the sentences (by their numbers) that most strongly influenced this score.
* Focus specifically on aspects relevant to the "Quality of Feedback" domain (e.g., scaffolding, prompting, encouragement of thought processes, questioning strategies, or absence of these features).
* Return exactly 10 sentence numbers, ordered by estimated importance (most influential first).
* If fewer than 10 sentences exist, return only those available.

Format your response **strictly** as a single line of comma-separated numbers (no spaces, no brackets, no explanations). For example: 9,15,6,37,2,7,8,1,64,66

Do not include any commentary, reasoning, or additional text.

You must begin your answer with
'### Output: <your comma-separated list>'

1. Transcript:
   {Full transcript segment}

2. Numbered Sentences:
   (1) - {First sentence}
   (2) - {Second sentence}
   (3) - {Third sentence}
   ...

3. Score:
   {Previously predicted score}

Output:

Figure 5: Zero-shot prompt for LLM sentence ranking. {Full transcript segment} represents the full transcript segment, with the same formatting as it appears in the dataset, {First/Second/Third/... sentence} represents the transcript split into sentences and numbered by order of appearance, and {Previously predicted score} represents the Quality of Feedback score predicted for the transcript segment either by the LLMs or by the PLMs.

D Average Number of Sentences Predicted

Table 4 reports the average number of sentences output by each LLM, together with the absolute difference (Δ) from the expected average of 9.931 sentences.

Model	# of sentences	Δ
Llama 3.1 8B Instruct	9.939	**-0.008**
Llama 3.1 70B Instruct	9.886	0.045
Mistral Small Instruct	9.654	0.277
Mistral Small 24B Instruct	9.871	0.060
Mixtral 8x7B Instruct	9.291	0.640
Mixtral 8x22B Instruct	9.836	0.095
Qwen3 4B Instruct	9.711	0.220
Qwen3 30B A3B Instruct	9.828	0.102
Qwen3 235B A22B Instruct	9.516	0.415

Table 4: Average number of sentences each model outputs for sentence ranking, and the difference between the expected and observed number of sentences.

E Additional Details on Explanation Alignment

Table 5 reports the alignment between SHAP-based and LLM-based explanations across 1,230 transcripts. Jaccard similarity is computed between the sets of the top-10 sentences identified by each method for every transcript. Spearman rank correlation is computed between sentence importance rankings derived from absolute SHAP values and LLM deletion order. Spearman correlations are reported only for transcripts where the correlation is well-defined, resulting in between 1,226 and 1,228 transcripts per model. Multiple SHAP runs are aggregated by averaging absolute SHAP values per sentence prior to ranking.

F Results of Sentence Removal

Table 6, shows the cumulative delta values after each step of the top 10 sentences removal. Fig. 6, represents it in a graphical form.

G Removed Sentence Examples

Table 7 lists randomly selected sentences removed during the sentence-deletion experiments. Sentences are grouped by the explanation method(s) that selected them.

Model	Jaccard@10	Spearman ρ
Llama 3.1 8B Instruct	0.081 ± 0.135	0.048 ± 0.134
Llama 3.1 70B Instruct	0.089 ± 0.136	0.067 ± 0.129
Mistral Small Instruct	0.081 ± 0.128	0.054 ± 0.129
Mistral Small 24B Instruct	0.090 ± 0.141	**0.082** ± 0.138
Mixtral 8x7B Instruct	0.072 ± 0.131	0.029 ± 0.125
Mixtral 8x22B Instruct	0.088 ± 0.142	0.070 ± 0.134
Qwen3 4B Instruct	0.084 ± 0.139	0.057 ± 0.138
Qwen3 30B A3B Instruct	0.089 ± 0.154	0.071 ± 0.134
Qwen3 235B A22B Instruct	**0.095** ± 0.159	0.081 ± 0.142
Average	**0.085**	**0.062**

Table 5: Alignment between SHAP and LLM explanations measured by Jaccard similarity over top-10 sentences and Spearman rank correlation. Values are reported as mean $\pm$ standard deviation.

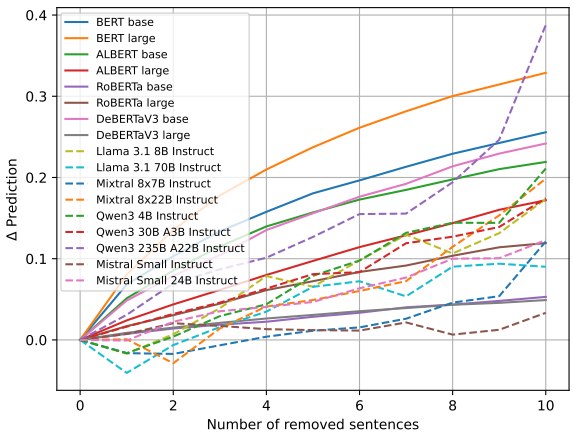

Figure 6: Average prediction changes after removing most important sentences and re-scoring within model family.

Model	Number of Sentences Removed									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
PLMs										
ALBERT base	0.051	0.085	0.115	0.140	0.157	0.173	0.185	0.198	0.210	0.219
ALBERT large	0.024	0.044	0.062	0.080	0.097	0.114	0.129	0.144	0.160	0.172
RoBERTa base	0.007	0.014	0.019	0.022	0.028	0.033	0.040	0.044	0.048	0.053
BERT base	0.068	0.103	0.134	0.158	0.180	0.196	0.213	0.229	0.242	0.256
BERT large	0.080	0.135	0.178	0.210	0.237	0.261	0.282	0.300	0.315	**0.329**
DeBERTaV3 base	0.048	0.079	0.104	0.135	0.156	0.176	0.192	0.214	0.229	0.242
DeBERTaV3 large	0.009	0.015	0.021	0.026	0.031	0.035	0.039	0.043	0.046	0.049
LLMs										
Llama 3.1 8B Instruct	-0.017	0.007	0.039	0.079	0.065	0.098	0.129	0.106	0.131	0.174
Llama 3.1 70B Instruct	-0.041	-0.006	0.016	0.035	0.065	0.072	0.054	0.090	0.094	0.090
Mistral Small Instruct	0.007	0.021	0.017	0.013	0.012	0.011	0.022	0.007	0.012	0.033
Mistral Small 24B Instruct	-0.001	0.022	0.035	0.041	0.047	0.064	0.077	0.100	0.101	0.123
Mixtral 8x7B Instruct	-0.016	-0.017	-0.007	0.004	0.011	0.015	0.026	0.046	0.054	0.121
Mixtral 8x22B Instruct	0.001	-0.029	0.014	0.041	0.049	0.060	0.072	0.115	0.153	0.199
Qwen3 4B Instruct	-0.016	0.004	0.029	0.044	0.079	0.098	0.132	0.144	0.144	0.211
Qwen3 30B A3B Instruct	0.016	0.032	0.046	0.063	0.081	0.083	0.119	0.127	0.139	0.174
Qwen3 235B A22B Instruct	0.031	0.070	0.087	0.101	0.127	0.155	0.156	0.194	0.247	**0.388**

Table 6: Performance deltas after removing consecutive numbers of sentences. Top: PLMs. Bottom: LLMs.

Model	Sentence
LLM	- You think it's not a polygon because it has a curved side. - Teacher: Okay, what you do is, yeah, you cut it in half again so you have their equivalents. - Remember in math I spent a lot of time going over with you every time you answer a question you have to a? - Is this area right here inside the O? - What's ten times 32? - What does she need to put right next to the inches? - Student: Um - Teacher: Try it - try a number for X and see how well it works for you. - Teacher: Add my two sums of my two grids. - Technically if you're finding area, you're finding the number of squares inside the shape, correct? - So are they different?
PLM	- 1, 2, 3. - Teacher: I could divide it by 2, yeah. - Why don't we count on the 9th? - Tell somebody in your group what you learned in math today. - What happens to 6? - 635. - Student: Can we write an answer sentence? - Multiple Students: Label. - Student: ... - Teacher: Plus 19 - Student D, what's 2 plus 9?
Both	- And now I can find my total which will be what, Student H? - Student: It's going to get bigger. - Teacher: So what is the denominator there? - As you're doing this and you're answering the questions I want to label it. - So what is 12 divided by 2? - Okay so 85 plus 15 would bring me to my next whole. - What happens to 6? - Teacher: You have to go ahead .... - Multiple Students: Less. - Student: It's three times bigger.

Table 7: Example sentences removed during sentence-deletion experiments, grouped by whether they were selected by LLM-based explanations, PLM-based explanations, or by both.