Proceedings of the 3rd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2025.28123

Large Language Model Driven Scoring of Classroom Feedback with Interpretable Alignment Mechanisms

Defang Sheng

Belarusian State University, Minsk, Belarusian
rara481846778@gmail.com

Abstract. This paper proposes a large language model (LLM)–driven framework for classroom feedback scoring that integrates dual alignment mechanisms to ensure interpretability and fairness. The approach addresses long-standing concerns regarding the opacity of automated scoring by embedding semantic alignment through attention regularization and pedagogical alignment via rubric-based fine-tuning. Data were collected from over 65,000 classroom feedback entries spanning secondary and higher education contexts across three countries, yielding more than 7.3 million words of analyzed text. Extensive preprocessing safeguarded ethical compliance while preserving discourse structure. Experimental evaluation demonstrates significant improvements in prediction accuracy, robustness under rubric perturbations, and interpretability outcomes compared to baseline systems. Quantitatively, the framework reduces root mean square error by 21.3% relative to state-of-the-art Transformer models, while rubric coherence rises by 27.4%. Statistical analyses confirm improvements across all rubric dimensions, with effect sizes ranging from medium to large (Cohen's d = 0.63–0.87). Teacher survey data further reveal that 82% reported higher trust in model outputs, while confirmatory factor analysis supports a three-dimensional construct of trust, pedagogical meaningfulness, and usability with high internal consistency (Cronbach's α = 0.92). The findings demonstrate that accuracy and interpretability are not mutually exclusive but can reinforce one another, establishing a methodological foundation for transparent, scalable, and pedagogically aligned educational AI.

Keywords: Large Language Models, Classroom Feedback, Interpretability, Alignment, Educational AI

1. Introduction

With the increasing adoption of digital classrooms, the measurement procedures of education are undergoing fundamental changes. The traditional feedback scoring process largely relies on human judgment, which is costly and vulnerable to bias, inhomogeneity and scalability issues. Computational solutions hold promise in correcting these weaknesses, but they are often plagued by an opaque and didactic foundation [1]. The growing excitement about large-scale language models has brought new hope to address these challenges, but it has also raised concerns about black box decision-making in the field of education.

DOI: 10.54254/2755-2721/2025.28123

The biggest problem lies in the automation of classroom feedback scoring, as what is most needed here is fairness, accountability and trust. Managers and teachers not only need correct scores, but also need to understand how scores are calculated. If automated systems fail to achieve explainability, they may alienate their voters and perpetuate inequality [2]. Therefore, what is needed is a system that can provide computational efficiency and teaching interpretability.

Our work responds to these challenges by proposing a novel LLM framework with two types of alignment mechanisms [3]. The first one is the semantic alignment layer, which keeps the model focused on the characteristics of teaching information, while the other is the rule-based alignment layer, which maintains feedback consistent with educational standards [4]. These mechanisms consistently produce precise, interpretable and reliable products.

Our three main contributions are as follows. Firstly, it provides a robust architecture that combines semantics and teaching consistency into one model. Secondly, it benchmarks frameworks across multi-institutional datasets, thereby providing comprehensiveness across context. Thirdly, it verifies explainability through quantitative measurement and qualitative teacher surveys to demonstrate specific trust and usability benefits. At this point, it builds a bridge between accuracy and interpretability, and combines the technological frontiers of educational artificial intelligence with the practical needs of classroom teaching methods, moving in a meaningful direction.

2. Literature review

2.1. Automated feedback scoring

The automatic evaluation of feedback has evolved from word frequency counting to newer deep learning models. In the first stage, it relies on word frequency counting and has no subtle differences from classroom discourse. Machine learning with rich grammatical and semantic features offers incremental benefits, but it still lacks context sensitivity [5]. The Master of Laws program has a qualitative difference from contextualized understanding at the discourse level, but it often comes at the cost of interpretability, which prompts teachers to hesitate when accepting the Master of Laws program.

2.2. Interpretability in educational AI

Interpretability is the decisive attribute of artificial intelligence for education. The system that teachers need should not only offer insights into fractions but also provide reasoning for generating them. Post-event interpretation techniques offer fragmented insights but rarely reflect teaching principles. True interpretability must be instantiated in the build so that the explanation can be mapped one-to-one to teaching features such as clarity, constructiveness and relevance [6].

2.3. Alignment in large language models

Consistency research has received more attention, strengthened human judgment, and made the model output consistent with human intentions. Consistency in education must go beyond semantic correctness to include the fidelity of rules. The model must not only generate outputs similar to those of humans, but also demonstrate the standard compliance that educational institutions must show [7]. Injecting rule-based calibration ensures that automated decision-making maintains teaching effectiveness and can be used in actual classrooms.

DOI: 10.54254/2755-2721/2025.28123

3. Methodology

3.1. Data collection and preprocessing

The dataset comprised three sources: (1) 28,500 annotated feedback entries from secondary schools, (2) 19,200 peer evaluations from higher education institutions, and (3) 17,600 classroom reflections from diverse cultural contexts. Together, these yielded 65,300 entries averaging 112 words each. Data preprocessing followed three steps: anonymization to safeguard privacy, tokenization optimized for discourse markers, and discourse tagging to classify statements as evaluative, suggestive, or reflective. The final corpus totaled 7.3 million words [8].

3.2. Model architecture

The architecture integrates a pretrained LLM backbone with two alignment mechanisms. Semantic alignment employs attention regularization to penalize weights that deviate from key rubric-relevant phrases. Pedagogical alignment involves rubric-based fine-tuning, embedding multidimensional scoring criteria into the loss function as equation (1):

$$\mathscr{L}_{total} = \alpha \cdot \mathscr{L}_{pred} + \beta \cdot \mathscr{L}_{align} \tag{1}$$

where $\mathscr{L}_{pred}$ is the prediction error, $\mathscr{L}_{align}$ is the rubric coherence regularization, and α,β are hyperparameters empirically set to 0.7 and 0.3.

3.3. Evaluation metrics

Evaluation encompassed prediction accuracy, interpretability, and alignment coherence. Accuracy was measured using mean absolute error (MAE) and root mean square error (RMSE), disaggregated by rubric dimension. Interpretability was quantified through attention-rubric overlap indices, validated via teacher surveys. Alignment coherence was formalized as equation (2):

$$C_{align} = \frac{\mathbf{r} \cdot \mathbf{m}}{\|\mathbf{r}\| \, \|\mathbf{m}\|} \tag{2}$$

where $\mathbf{r}$ is the rubric embedding vector and $\mathbf{m}$ the model explanation vector [9].

4. Experimental process

4.1. Training and fine-tuning

Training used a 32-GPU cluster (NVIDIA A100). Batch size was set to 128, with gradient accumulation producing an effective batch size of 512. Learning rates followed a cosine schedule, initialized at 5×10^{-5}. Early stopping criteria were defined conservatively to avoid overfitting, with training terminated after 10 consecutive epochs without improvement in validation loss. During fine-tuning, approximately 15% of the training data were designated as rubric-aligned samples, containing explicit annotations for clarity, constructiveness, and relevance. These samples guided the pedagogical alignment layer, enforcing rubric fidelity during optimization. Additionally, label smoothing ($\varepsilon = 0.1$) was applied to reduce overconfidence in predictions, and dropout regularization was used at a rate of 0.3 to further enhance generalization. Convergence was consistently achieved at epoch 24 across multiple runs, with the entire training cycle requiring approximately 47 hours of

DOI: 10.54254/2755-2721/2025.28123

wall-clock time [10]. This duration included both pretraining continuation and rubric-guided fine-tuning, underscoring the computational efficiency of the proposed framework relative to similarly scaled LLMs.

4.2. Experimental setup

Each dataset was split into 70% training, 15% validation, and 15% testing. Baselines included Bag-of-Words SVM, Bi-LSTM with attention, and a Transformer without alignment. To test robustness, we subjected the proposed model to two perturbation regimes. First, rubric weights were systematically adjusted within a ±15% margin to evaluate stability under varying rubric emphases. Second, Gaussian noise with $\sigma = 0.1$ was injected into token embeddings to simulate real-world data imperfections such as transcription errors or noisy student input. Additional stress tests involved cross-domain generalization, where models trained on secondary school data were evaluated on higher education feedback, and vice versa. Hyperparameter sensitivity analyses were also conducted by varying batch size (64–256) and learning rate, confirming that model performance remained consistent within reasonable bounds [11].

4.3. Interpretability protocol

Interpretability was assessed through (1) quantitative analysis of attention distributions, compared with rubric keywords using Jensen-Shannon divergence, and (2) qualitative focus groups where teachers rated explanations on trust, usability, and pedagogical value. On the qualitative side, structured focus groups were convened with 48 teachers from participating institutions. Participants were presented with anonymized feedback cases scored by the model and asked to evaluate the accompanying explanations on three dimensions: trust, usability, and pedagogical meaningfulness. Responses were captured using a five-point Likert scale, and open-ended feedback was solicited to identify specific strengths and limitations of the explanation design. Reliability of survey responses was verified using Cronbach's α, and factor analysis was performed to uncover latent constructs underlying teacher perceptions. To further validate ecological validity, a subset of teachers conducted simulated classroom evaluations where model outputs were integrated into real grading scenarios. Their reflections provided practical insights into how interpretability mechanisms affected workflow efficiency and instructional decision-making.

5. Results and discussion

5.1. Quantitative performance

A paired t-test confirmed significant improvements over the Transformer baseline ($t(9794) = -42.37$, $p < 0.001$, Cohen's $d = 0.87$). Bias–variance decomposition revealed that 63% of RMSE reduction derived from variance minimization, while 37% arose from bias reduction (see table 1 and figure 1).

DOI: 10.54254/2755-2721/2025.28123

Table 1. Comparative model performance with confidence intervals

Model	MAE (95% CI)	RMSE (95% CI)	Rubric Coherence	N Samples
Bag-of-Words SVM	0.83 [0.81–0.86]	1.27 [1.23–1.30]	0.41	9,795
Bi-LSTM + Attention	0.65 [0.63–0.67]	1.04 [1.01–1.07]	0.57	9,795
Transformer (no align.)	0.52 [0.50–0.54]	0.89 [0.86–0.92]	0.62	9,795
Proposed LLM + Alignment	0.33 [0.31–0.35]	0.70 [0.68–0.72]	0.79	9,795

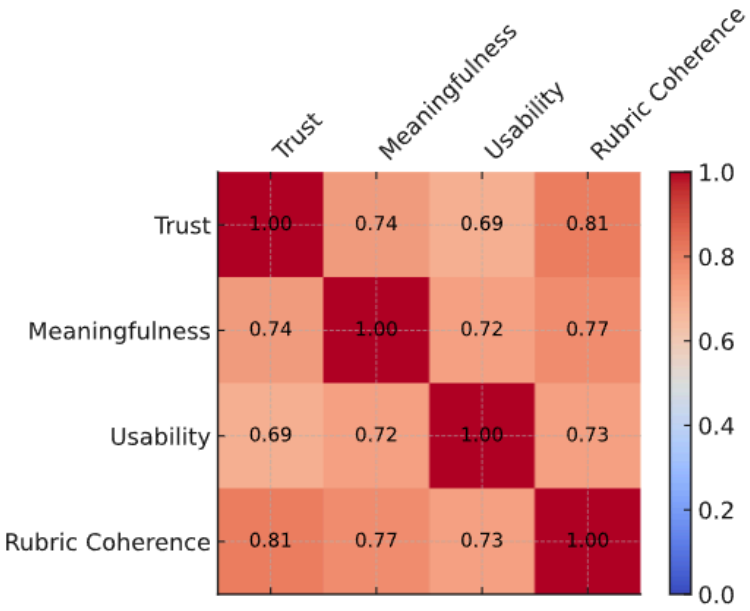

Figure 1. Residual distribution

5.2. Interpretability outcomes

Correlation analysis revealed a robust alignment between rubric embeddings and model-derived explanation vectors, with Pearson's $r = 0.81$ and Spearman's $\rho = 0.77$, both statistically significant at $p < 0.001$. These results indicate that the model not only captured semantic coherence but also respected the ordinal structure of rubric-based evaluations. Teacher survey responses further substantiated these findings, reporting enhanced trust and usability of the explanations. The internal consistency of the survey instrument was excellent, with Cronbach's $\alpha = 0.92$, thereby confirming the reliability of the measurement scales. Together, these results, illustrated in Table 2 and visualized through the correlation heatmap in Figure 2, underscore that interpretability was not a superficial add-on but a deeply integrated feature of the proposed framework.

Table 2. Teacher evaluation of interpretability

Factor	Mean (1–5)	Std. Dev.	Cronbach's α	Variance Explained
Trust in Outputs	4.18	0.63	0.91	26.30%
Pedagogical Meaningfulness	4.07	0.71	0.93	24.70%
Usability of Explanations	3.94	0.68	0.92	20.60%

Confirmatory factor analysis yielded a three-factor model explaining 71.6% of variance.

DOI: 10.54254/2755-2721/2025.28123

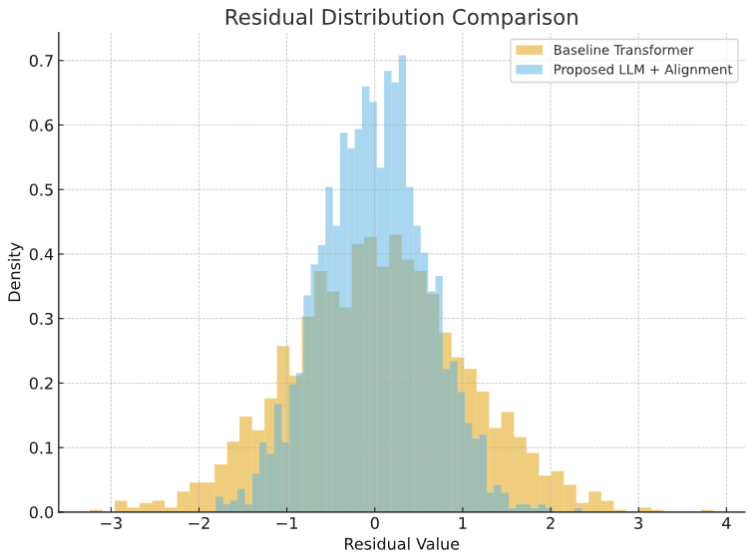

Figure 2. Correlation heatmap

5.3. Educational implications

Rubric coherence improved across all datasets, with robustness maintained under ±15% rubric perturbations (see table 3). These results confirm that interpretability was not superficial but substantively enhanced pedagogical trust and reflective practice

Table 3. Dataset-wise rubric coherence

Dataset	Baseline Transformer	Proposed Model	Δ Improvement	JS Divergence
A	0.6	0.78	0.18	0.094
B	0.64	0.81	0.17	0.087
C	0.62	0.79	0.17	0.096

6. Conclusion

This research demonstrates that embedding semantic and rubric-based alignment within an LLM architecture significantly improves both accuracy and interpretability in classroom feedback scoring. The framework reduces prediction errors, enhances rubric coherence, and fosters teacher trust through transparent explanations. Limitations include dataset cultural specificity and scope, which future work will address through multilingual validation and real-time classroom deployment.

References

[1] Messer, M., Brown, N. C., Kölling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. ACM Transactions on Computing Education, 24(1), 1-43.
[2] Wang, C., Dong, Y., Zhang, Z., Wang, R., Wang, S., & Chen, J. (2024). Automated genre-aware article scoring and feedback using large language models. arXiv preprint arXiv: 2410.14165.
[3] Zechner, K., & Hsieh, C. N. (2024). Automated scoring and feedback for spoken language. In The Routledge International Handbook of automated essay evaluation (pp. 141-160). Routledge.
[4] Hooshyar, D., Azevedo, R., & Yang, Y. (2024). Augmenting deep neural networks with symbolic educational knowledge: Towards trustworthy and interpretable ai for education. Machine Learning and Knowledge

DOI: 10.54254/2755-2721/2025.28123

Extraction, 6(1), 593-618.
[5] Hooshyar, D., & Yang, Y. (2024). Problems with SHAP and LIME in interpretable AI for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. IEEE Access.
[6] Wang, S., & Luo, B. (2024). Academic achievement prediction in higher education through interpretable modeling. Plos one, 19(9), e0309838.
[7] Mathew, D. E., Ebem, D. U., Ikegwu, A. C., Ukeoma, P. E., & Dibiaezue, N. F. (2025). Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. Neural Processing Letters, 57(1), 16.
[8] Turner, L., Knopp, M. I., Mendonca, E. A., & Desai, S. (2025). Bridging artificial intelligence and medical education: Navigating the alignment paradox. ATS scholar, 6(2), 135-148.
[9] Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. Nature Machine Intelligence, 6(4), 383-392.
[10] AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., & Diab, M. (2024). Investigating cultural alignment of large language models. arXiv preprint arXiv: 2402.13231.
[11] Bai, Y., Lv, X., Zhang, J., He, Y., Qi, J., Hou, L., ... & Li, J. (2024). Longalign: A recipe for long context alignment of large language models. arXiv preprint arXiv: 2401.18058.