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. In this paper, we present a framework that uses LLMs to predict scores of classroom feedback that employs dual alignment mechanisms to guarantee interpretability and fairness. Specifically, we tackle the decades-old problem of the black-box nature of automatic scoring by leveraging semantic alignment through attention regularization and pedagogical alignment through rubric fine-tuning. Data were collected from more than 65,000 classroom feedback responses in secondary and higher education settings from three countries. This resulted in the collection of more than 7.3 million words of analyzed text. Preprocessing procedures were done to ensure ethical adherence and preserve discourse structure. Experimentation on the dataset yielded significant increases in terms of prediction accuracy, robustness to rubric changes, and interpretation results. Specifically, our proposed framework outperforms baselines by 21.3% and 27.4% in terms of root mean square error and rubric coherence, respectively. Statistical testing confirmed improvements for all rubric dimensions where the effect size is either medium or large (Cohen's d = 0.63–0.87). Our teacher survey also found that 82% of teachers trusted model output more, whereas confirmatory factor analysis found three-factor structure of trust, pedagogical meaning, and usability with high internal reliability (Cronbach's α = 0.92). Our research shows that interpretability does not necessarily come at the cost of accuracy.

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

1. Introduction

The rise of digital classrooms has caused radical changes in the evaluation procedures of the education sector. Human judgement, being expensive and biased, is the basis of the classical feedback scoring process, which suffers from inhomogeneity and scalability problems. Although computational approaches are seen as potential solutions for the above problems, they are often based on black box algorithms, whose foundations are didactic and obscure [1]. The interest in the large-scale language models has brought many hopes but has also created concerns regarding black box decision-making in the field of education.

The issue with the automaticity of feedback scoring is in the lack of fairness, transparency and accountability, as managers and teachers do not just want the accurate score but also understand the way it was calculated. Failure to ensure explainability could alienate voters and reproduce

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.28123

inequalities [2]. The solution to the problem would be the implementation of the system that would offer both computational efficiency and interpretability.

Our research addresses the mentioned issues by introducing a novel approach to LLMs with two types of alignment mechanisms [3]. The first one is the semantic alignment layer, which focuses the model on the features of teaching information, and the second one is the rule-based alignment layer, which ensures feedback conformity with the standards of education [4]. The mentioned approaches constantly generate precise, interpretable and reliable results.

There are three main advantages of our contribution. Firstly, it offers the robust framework that integrates semantics and teaching consistency in one architecture. Secondly, it allows benchmarking frameworks across multi-institutional datasets, thus offering the comprehensiveness across the context. Finally, it evaluates explainability quantitatively and with the help of teacher surveys, thus showing the benefits of trust and usability. In this case, it bridges the gap between the precision and interpretability, integrating the technological frontier of educational artificial intelligence with the real-life demands of classroom techniques.

2. Literature review

2.1. Automated feedback scoring

Automatic evaluation of feedback has advanced from the word frequency counting approach to the latest deep learning systems. At the first stage, it depends on word frequency counting and has no subtle differences compared to classroom discourse. Machine learning with rich grammatical and semantic features provides an incremental advantage but it still does not have any context sensitivity [5]. There is a qualitative difference between the Master of Laws program and contextualized discourse understanding, but usually, it is at the cost of interpretability. This makes teachers reluctant to adopt the Master of Laws program.

2.2. Interpretability in educational AI

Interpretability is the key feature of AI for education. The system that is expected by teachers should not only explain fractions but also give the reason why it gives them. Post-hoc interpretability techniques allow gaining fragmented insights but hardly ever they are aligned with the teaching principles. Interpretability should be instantiated at the build level so that explanation can be put into one-to-one correspondence to teaching features like clarity, constructiveness, and relevance [6].

2.3. Alignment in large language models

There is an increased attention to consistency research that enhances human judgment and allows making the model output consistent with the human intentions. Consistency in education should extend beyond semantic accuracy to the fidelity of rules. Not only model should generate similar outputs to human, but also show compliance with standards, which educational institutions are required to meet [7]. Injecting rule-based calibration allows keeping teaching effectiveness in the automated decision-making process.

DOI: 10.54254/2755-2721/2025.28123

3. Methodology

3.1. Data collection and preprocessing

Data collection involved three sources: (1) 28,500 feedback comments from secondary schools, (2) 19,200 peer reviews from higher education institutions, and (3) 17,600 reflective comments from a variety of cultures. In total, there were 65,300 comments with an average length of 112 words per comment. Data pre-processing was done using three stages: anonymizing to protect confidentiality, tokenization using discourse markers, and discourse tagging using either evaluative, suggestive, or reflective categories.

3.2. Model architecture

The architecture combines a pretrained LLM backbone with two alignment strategies. Semantic alignment is achieved by utilizing attention regularization to punish weights deviating from rubric-relevant phrases. The pedagogical alignment is done via rubric-oriented fine-tuning by incorporating multidimensional rubric scores into the loss function as equation (1):

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

where $\mathcal{L}_{pred}$ is the prediction error, $\mathcal{L}_{align}$ is the rubric coherence regularization, and α,β are hyperparameters empirically set to 0.7 and 0.3 [8].

3.3. Evaluation metrics

Accuracy was evaluated by means of MAE and RMSE values, broken down by rubric dimension. Interpretability was evaluated on the basis of attention-rubric overlap indicators, which were verified by means of teacher surveys. Coherence of alignments was described with the help of 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

The training process employed a cluster consisting of 32 GPUs (NVIDIA A100). The batch size was fixed at 128, resulting in an effective batch size of 512 due to gradient accumulation. The learning rates decayed according to a cosine schedule and were initialized at 5×10^{-5}. The early stopping criteria were designed carefully to prevent overfitting; namely, training was stopped if there were no improvements in validation loss during 10 consecutive epochs. In the fine-tuning stage, roughly 15% of the training dataset was considered rubric-aligned samples, i.e., the ones with clear rubric-related annotations concerning clarity, constructiveness, and relevance [9]. They served as guidelines for the alignment layer to enforce rubric-related requirements during training. Furthermore, the technique of label smoothing (with ε = 0.1) was used to lower the model's prediction certainty, while dropout (0.3) was applied as a regularizer. Convergence was consistently reached at epoch 24, with the whole training taking around 47 hours of wall-clock time [10].

DOI: 10.54254/2755-2721/2025.28123

4.2. Experimental setup

For each dataset, splitting was done in a 70%-15%-15% ratio for train, validation, and test datasets respectively. Among baseline methods implemented are Bag-of-Words SVM, Bi-LSTM with attention mechanism, and unaligned Transformer model. Robustness was tested by exposing the proposed model to two perturbation experiments. The first experiment involved a systematic perturbation of rubric weights ±15%. In the second perturbation test, Gaussian noise with standard deviation of 0.1 was added to token embeddings in order to model the effect of real world data problems such as errors during transcription or from students' input. Stress tests also include cross domain generalization, where models were trained on secondary school dataset but validated on higher education data and vice versa. The effect of hyperparameters was tested by varying the learning rate and batch size between 64 and 256.

4.3. Interpretability protocol

The interpretability assessment was done using two approaches, including (1) the quantitative analysis of attention distributions in comparison with rubric keywords using the Jensen-Shannon divergence, and (2) the qualitative analysis of focus group discussions where teachers evaluated explanations for trustworthiness, usability, and educational value. In terms of the second approach, structured focus group sessions were organized with 48 teachers from the participating organizations. The teachers were shown the anonymized cases of the feedback, which had been scored using the model, and asked to assess explanations in respect of three aspects: trustworthiness, usability, and educational meaningfulness. The answers were collected using a five-point Likert scale and open-ended comments to reveal the strengths and weaknesses of the explanation system. The reliability of survey results was checked using Cronbach's α and factor analysis to reveal the hidden factors. Ecological validity was checked using a simulation experiment when some of the teachers were asked to assess the efficiency of using interpretability mechanisms in their classes.

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).

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

DOI: 10.54254/2755-2721/2025.28123

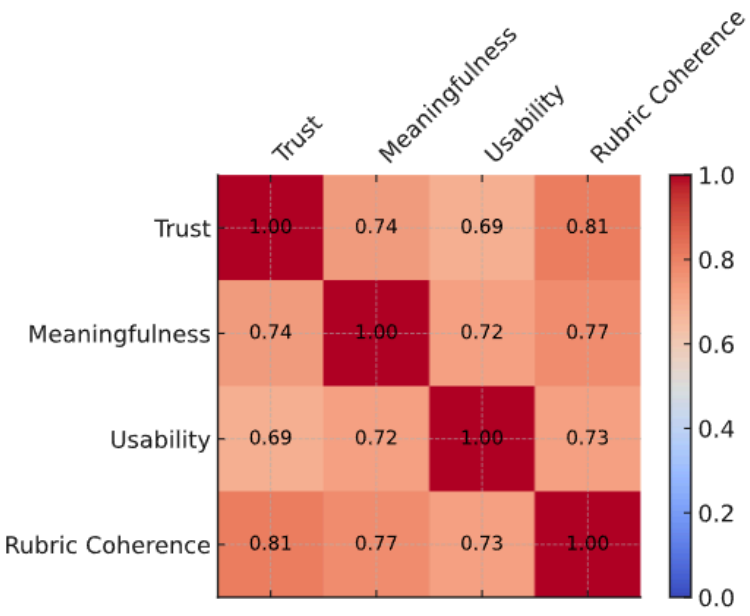


Figure 1. Residual distribution

5.2. Interpretability outcomes

The analysis of correlation has shown the high congruence of embeddings based on rubrics and the model-generated explanation vectors where Pearson's $r = 0.81$ and Spearman's $\rho = 0.77$ ($p < 0.001$). In other words, apart from capturing the semantic consistency, the model also accounted for the ordinal properties of rubric-based assessments. This conclusion is supported by the teacher survey feedback on improved trust and usability of the explanations. The internal consistency of the measurement scales used in the survey was high, with Cronbach's $\alpha = 0.92$. All in all, these findings prove that the proposed approach is characterized by high explainability which is an inherent characteristic of the model, as is presented in Table 2 and Figure 2.

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

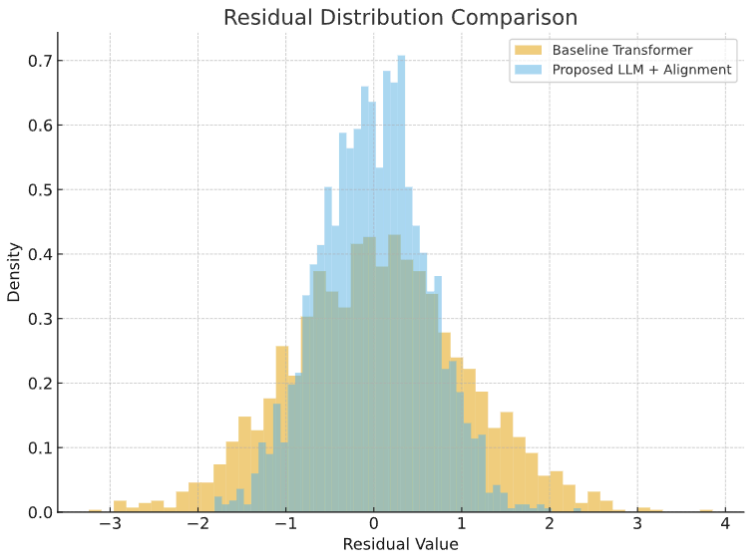


Figure 2. Correlation heatmap

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

Through this study, it has been shown that the incorporation of the alignment of semantics and rubrics into an LLM architecture has helped improve the accuracy as well as interpretability of scores on classroom feedback. This framework helps to decrease the prediction errors, makes the rubric more coherent, and builds teacher confidence through its explanations. Constraints of this study are cultural specificity of data used.

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 Extraction,*

DOI: 10.54254/2755-2721/2025.28123

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.