Decoding Actionability: A Computational Analysis of Teacher Observation Feedback

Mayank Sharma
Stanford University
masharma@stanford.edu

Jason Zhang
Stanford University
jasonzyj@stanford.edu

Abstract

This study presents a computational analysis to classify actionability in teacher feedback. We fine-tuned a RoBERTa model on 662 manually annotated feedback examples from West African classrooms, achieving strong classification performance (accuracy = 0.94, precision = 0.90, recall = 0.96, f1 = 0.93). This enabled classification of over 12,000 feedback instances. A comparison of linguistic features indicated that actionable feedback was associated with lower word count but higher readability, greater lexical diversity, and more modifier usage. These findings suggest that concise, accessible language with precise descriptive terms may be more actionable for teachers. Our results support focusing on clarity in teacher observation protocols while demonstrating the potential of computational approaches in analyzing educational feedback at scale.

1 Introduction

Classroom observation plays a crucial role in evaluating and enhancing instructional quality (Adelman and Walker, 1975; Wragg, 2011). By offering a direct perspective on teaching in authentic settings, it provides insights into how educators engage with students and structure their instruction (Millman and Darling-Hammond, 1990; Putnam and Borko, 2000). It also serves as a vital link between teaching practices and student learning outcomes, thus creating the foundation for teacher professional development (Kane and Staiger, 2012).

Given this significance, the quality of feedback derived from classroom observations is essential (Lazarev and Newman, 2015). While various characteristics contribute to effective feedback, including constructive tone and clarity, research emphasizes that specificity and actionability are particularly crucial for enhancing teacher performance (Archer et al., 2016). Truly actionable feedback provides specific recommendations and clear direction, establishing concrete performance expectations and supporting professional growth (Cannon and Witherspoon, 2005). By focusing on observable teaching behaviors rather than personal attributes, such feedback enables meaningful instructional improvements (Archer et al., 2016).

Although research on actionable feedback originated largely outside education, its principles have proven directly applicable to classroom contexts. In organizational psychology, Cannon and Witherspoon (2005) identified key elements of actionable feedback: specificity, balanced positive and constructive components, and clear connections between observed behaviors and suggested improvements. This aligned with Kluger and DeNisi (1996)'s comprehensive meta-analysis of over 3,000 feedback interventions, which found that feedback effectiveness varies dramatically based on specificity and delivery characteristics. Within education-specific research, multiple studies have confirmed and extended these general principles. For example, Allen et al. (2011) demonstrated that structured feedback systems yield measurable improvements in teaching quality. Similarly, Thurlings et al. (2013) found that effective teacher feedback typically contains explicit behavioral descriptions, rationales for suggested changes, and concrete examples of alternative approaches. Quantitative evidence from Steinberg and Sartain (2015)'s analysis of over 12,000 teacher observation records showed that feedback incorporating concrete examples and precise language led to measurable gains in subsequent evaluations. In a similar way, Hill et al. (2012) established that feedback quality directly correlates with improvements in instructional practice, particularly when including specific action steps. In fact, Darling-Hammond et al. (2017)'s work on professional development systems reinforces the critical role of actionable feedback as a bridge between observation and implementation.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 898–907

Despite the importance of actionable feedback, classroom observers often struggle to provide guidance that teachers can readily implement (Kraft et al., 2018). This implementation gap stems from inconsistent understanding of what constitutes actionable feedback and the absence of systematic approaches to analyze feedback quality at scale. Computational approaches offer promising avenues for analyzing observation feedback and identifying patterns in actionable feedback. However, applying these methods to classroom observation requires addressing how actionability can be computationally defined and recognized. Our research bridges educational theory and computational methods to develop methods that can meaningfully evaluate the actionability of teacher feedback.

2 Prior Work

While we were not able to identify any existing studies specifically focused on using NLP approaches to identify actionable teacher feedback, adjacent educational research provides relevant context for our work. In the domain of classroom observation, Demszky et al. (2021) analyzed linguistic features in teacher speech to evaluate instructional effectiveness. Similarly, Suresh et al. (2019) examined different dimensions of teacher feedback, though their work did not address actionability specifically. Beyond teacher-focused research, computational analyses of student-centered feedback have shown promising results. Leeman-Munk et al. (2014) developed methods to evaluate student writing and identify improvement areas, while Madnani et al. (2017) created models for standardized writing assessments that demonstrated reliability comparable to human raters.

The emergence of large language models (LLMs) has also sparked interest in their potential for educational annotation and classification tasks. However, Wang et al. (2023) found that models like GPT struggled to accurately classify nuanced educational distinctions. This aligns with Hardy (2025)'s assertion that classroom settings represent "out-of-distribution" data for LLMs, which are primarily trained on broad internet crawls. Additionally, concerns about data privacy, environmental impact, and the ethics of automated educational assessments complicate their use in education. In contrast, specialized transformer-based models offer more promising results for educational applications. Research indicates that models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), when properly trained on educational data, can outperform larger LLMs in classifying teacher-student interactions (Wang et al., 2023). Zhang and Litman (2021) demonstrated that these models can be trained on modest amounts of annotated educational data while maintaining strong performance, making them more practical for applications where annotated data may be limited.

Our Study

Despite substantial research on the importance of actionable feedback, computational approaches for identifying actionability in teacher observation feedback remain largely unexplored. This gap appears to exist primarily because of: (1) the lack of clear, consensus definitions of "actionability" in educational contexts; and (2) the scarcity of annotated datasets, as creating these typically requires time-consuming and resource-intensive manual annotation by educational experts (Shah and Pabel, 2019; Shaik et al., 2022).

Our study addresses these gaps through a novel approach where we first established a training dataset by annotation of approximately 660 instances of classroom observation feedback as either actionable or vague. Using this annotated corpus, we fine-tuned RoBERTa to extend this classification to a much larger dataset of over 12,000 feedback instances. With this comprehensive dataset, we conducted an examination of the linguistic features associated with actionability. These findings hold potential to inform the training of classroom observers, guide the development of automated feedback assessment tools, and help improve teacher professional development.

3 Data

This study utilized a large-scale dataset collected from classrooms in Sierra Leone, Liberia, and Ghana by Rising Academies during 2023-2025. The dataset includes N = 13,118 classroom observation records, each documenting teacher feedback provided by trained observers. Descriptive statistics on the schools, grades and subjects from which these observations were sourced are presented in Table 1. As shown, the observations come from 273 schools (approx. 48 observations/school) and were recorded by 76 observers (approx. 173 observations/observer).

Each observation was recorded using a struc-

Category	Value (Percentage)
Observers	
Number of Observers	76
Avg. Observations/Observer	172.6
Schools	
Number of Schools Observed	273
Avg. Observations/School	48.1
School Categories	
Top performing	481 (3.7%)
High Impact	2327 (17.7%)
Middle performing	580 (4.4%)
Moderate	4443 (33.9%)
Developing	3198 (24.4%)
Challenging	501 (3.8%)
Critical	507 (3.9%)
N/A	1081 (8.2%)
Grades	
Grade 1	2393 (18.2%)
Grade 2	2621 (20.0%)
Grade 3	2562 (19.5%)
Grade 4	2949 (22.5%)
Grade 5	1470 (11.2%)
Grade 6	1123 (8.6%)
Subjects	
Math	5201 (39.6%)
Faster Math	1978 (15.1%)
Reading	2806 (21.4%)
Faster Reading	3133 (23.9%)

Table 1: Descriptive statistics on observations. *Note.* For the purposes of this study, data from Grades 1-6 and 4 subjects: Reading, Math, Faster Reading, and Faster Math, were selected. Faster Reading and Faster Math are accelerated learning programs designed to supplement regular school curricula (N = 13118)

tured two-column format that included: (1) *What Went Well* (WWW) statements, which highlighted teacher strengths or effective strategies, and (2) *Even Better If* (EBI) suggestions, aimed at guiding improvements in teaching practices. As shown in the distribution in Figure 1, the average feedback length was 16.95 words (SD = 10.83). While the feedback length was relatively short, there was significant variation in its detail and clarity, ranging from broad praise/criticism to more specific recommendations. Due to the short length, no textual preprocessing was applied.

4 Methods

We organized our study into five sequential phases (visualized in Figure 2):

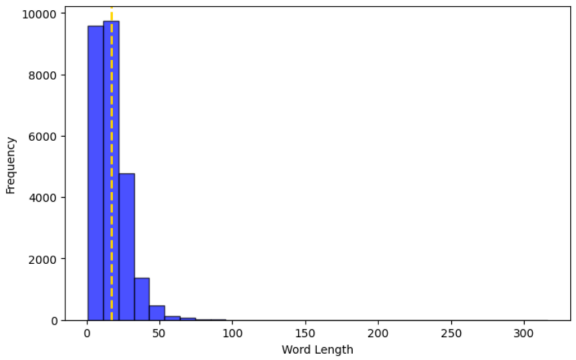

Figure 1: Probability density distribution of feedback length (in words; N = 13118)

Phase 1: Annotation and rubric development

In this phase, we developed a training dataset for fine-tuning RoBERTa through a rigorous annotation process. A stratified subsample of 750 comments was selected across multiple dimensions (school categories, grade levels, and subject areas) to ensure representativeness. Two independent researchers annotated each observation feedback according to a standardized rubric derived from established literature (see Appendix A for details), classifying comments as either "actionable" or "vague." This dual-annotation approach facilitated the calculation of inter-rater reliability using Cohen's Kappa (κ), yielding a coefficient of 0.60, which indicated moderate agreement.

Discrepancies were methodically resolved through iterative analytical discussions, which simultaneously informed the refinement of our annotation protocol, culminating in the revised rubric presented in Appendix A. The operationalization of "actionable" feedback centered on the presence of concrete, specific suggestions with explicit guidance on both implementation targets and mechanisms. For instance, the comment "*The teacher did great grouping learners and made them pick one word on a flash card where the group later leads in learning the new word. It would be better if the teacher completed the lesson in one hour to allow time for other lessons*" exemplified actionable feedback due to its specific temporal recommendation and clear rationale. Conversely, feedback was classified as "vague" when it lacked implementation specificity, regardless of the presence of ostensibly directive phrases such as "*even better if*" or "*could have*." The comment "*Giving more energy to make the class exciting was absolutely missing*" illustrates this classification, as it presents an evaluative

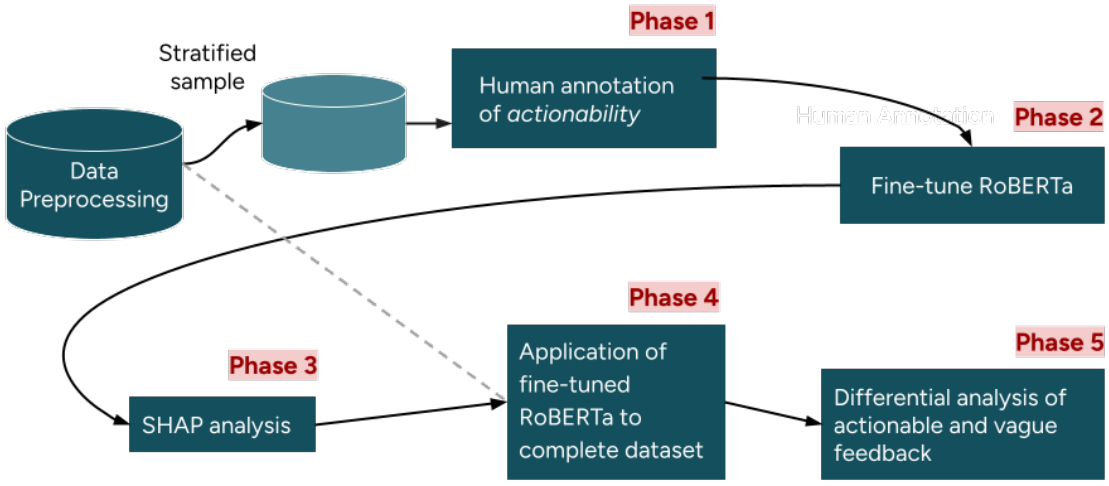

Figure 2: Flow chart for the study

statement without concrete behavioral specifications for improvement. We also excluded instances lacking sufficient evaluative clarity, resulting in a final annotated corpus of 662 comments (reduced from the initial 750). This dataset served as the training corpus for the next phase.

Phase 2: Fine-tuning RoBERTa on the annotated dataset

Model Specification

We fine-tuned RoBERTa (Liu et al., 2019) (using Hugging Face's `roberta-base`[1]) on our annotated dataset. An input token length of 512 was selected for the textual embeddings, with truncation and padding applied as needed. This vector was thresholded using `threshold=0.6` to produce the output vector. We chose this value to prioritize precision over recall, as our context requires high-confidence predictions of actionability rather than maximizing the identification of all potentially actionable feedback.

Training

The model was trained on a T4 GPU via Google Colab[2] using `adam` optimizer with `learning_rate=2e-5` with `linear_decay` of 0.01. For training, a `batch_size=16` was chosen, while `batch_size=32` was chosen for evaluation, keeping in mind compute bandwidths. We trained the model for 5 epochs and reported results from epoch 3 as the final epoch because model performance degraded afterward. The standard cross-entropy loss function was chosen (default for `roberta-base`).

Evaluation

To evaluate the model's performance, we used a held-out test set comprising 20% of the total dataset. The assessment was based on standard classification metrics: `accuracy`, `precision`, `recall`, and `f1-score`.

Phase 3: SHAP analysis

To gain deeper insights into the textual features driving our model's decisions, we employed SHapley Additive exPlanations (SHAP) analysis (Lundberg and Lee, 2017). This model-agnostic technique provides interpretability by attributing prediction outcomes to specific input features; in this study, words and phrases within the observation comments.

Recent work by Benslimane et al. (2024) validated SHAP's effectiveness for analyzing short, informal texts, demonstrating its reliability in identifying semantic patterns including emotional tone, gender references, and political language. Building on this empirical evidence, we applied SHAP to analyze 500 teacher feedback instances (250 ac-

[1]Available at https://huggingface.co/FacebookAI/roberta-base

[2]Available at https://colab.research.google.com/

tionable, 250 vague) and quantified each token's influence on model classification decisions. To identify consistent patterns, we aggregated SHAP values by unique tokens, calculated mean importance scores across all samples, and ranked terms according to their average contribution to classification outcomes.

Phase 4: Application of fine-tuned RoBERTa to the complete dataset

We utilized our fine-tuned RoBERTa model to categorize the rest of the observations as "vague" or "actionable." The model from the best-performing cross-validation fold was selected and used to make these predictions. A custom PyTorch `Dataset` class was implemented, which tokenized input text using the RoBERTa tokenizer with a maximum sequence length of 512. Tokenized inputs were converted into tensors with appropriate attention masks. To ensure computational efficiency, batch predictions (with `batch_size=32`) were performed using PyTorch's `DataLoader`. For each batch, input tensors were used to extract logits. From these logits, class predictions were obtained using the `argmax` function, and class probabilities using the `softmax` function. Instances with `softmax` probabilities less than 0.90 were classified as "low probability" instances and removed from the dataset.

Phase 5: Differential Analysis of Actionable and Vague Feedback

In this step, we extracted several linguistic features known to be associated with text clarity, specificity, and directiveness. These features were selected to potentially distinguish actionable observations from vague observations classified in the last step:

1. **Word Count**: We calculated the total number of words in each observation using `NLTK`'s word tokenization. Previous research suggests that actionable feedback tends to be more detailed, which could potentially result in higher word counts that provide implementable information (Winstone et al., 2016).

2. **Reading Ease**: We calculated Flesch reading ease using `textstat`. In this metric, the readability of the observation was scored on a 100-point scale, with higher scores indicating easier reading (Flesch, 1948). More accessible language may correlate with feedback that can be readily understood and implemented. Previous work has demonstrated that Flesch Reading Ease can be effectively used with short-form textual data such as tweets, and can enable robust analysis of readability even in brief, informal text (Davenport and DeLine, 2014).

3. **Lexical Diversity**: This was calculated using `NLTK`'s word tokenization as the ratio of unique words to total words in the observation text. Higher lexical diversity may indicate more specific feedback, potentially offering clearer guidance for action. Conversely, excessive diversity might introduce complexity that reduces actionability.

4. **Modifier Count**: This was calculated by counting modifiers (adjectives and adverbs) in the observations using `spaCy`'s POS tagger. Higher modifier counts might indicate more descriptive or qualifying language, which could potentially correlate with either actionability.

We ran a logistic regression model that included the linguistic features as predictors and feedback category as the outcome to examine the odds ratio for the categories.

The code used in the study is available on a publicly accessible GitHub repository.

5 Results and Discussion

In this section, we present results from Phases 2-5, as Phase 1 has already been described completely in the Methods section.

Phase 2: Fine-tuning RoBERTa on the annotated dataset

Fine-tuned RoBERTa demonstrated strong and stable performance in distinguishing actionable from vague teacher feedback. Using stratified 5-fold cross-validation on 662 annotated examples, the model achieved a mean `accuracy` of 0.94, `precision` of 0.90, `recall` of 0.96, and `f1` of 0.93 across folds, with an F1 standard deviation of 0.03. These metrics reflect performance on held-out validation sets and suggest the model generalizes well despite the modest dataset size. This aligns with findings by Zhang and Litman (2021) that well-curated educational data, even in small quantities, can yield high-performing models when paired with appropriate architectures.

Overfitting was monitored via 5-fold cross-validation and early stopping. The model showed

consistently high validation performance (mean F1 = 0.93, SD = 0.03), with no signs of overfitting.

Phase 3: SHAP analysis

Table 2 presents the top 20 most influential words for actionable and vague feedback based on SHAP analysis (positive values indicate features associated with "actionable" class, while negative values indicate association with "vague" class). The results provide mixed evidence without clear-cut patterns. While some action verbs and specific instructional behaviors (e.g., "struggled," "checks," "encourages," "provide") appear in the actionable feedback category, and certain comparative and conditional terms ("whether," "enough," "instead") appear in the vague feedback category, the overall linguistic distinctions lack sufficient consistency to draw definitive conclusions. The absence of strong patterns suggests that actionability may be determined by relationship between words rather than individual word choices alone.

Actionable Feedback		**Vague Feedback**	
Word	**SHAP Value**	**Word**	**SHAP Value**
struggled	0.055	sa	-0.051
checks	0.052	equal	-0.025
genders	0.044	whether	-0.025
sentences	0.033	needed	-0.025
tried	0.030	called	-0.023
encourages	0.029	easier	-0.019
improv	0.029	25	-0.018
introduced	0.027	avoid	-0.017
achers	0.026	kick	-0.017
provide	0.026	8	-0.016
helpful	0.026	enough	-0.016
stage	0.026	q	-0.015
minutes	0.026	arus	-0.015
excellent	0.024	had	-0.015
teaches	0.023	name	-0.015
helped	0.023	creative	-0.014
pared	0.023	instead	-0.014
rew	0.023	enable	-0.013
creat	0.022	supposed	-0.013
days	0.022	note	-0.012

Table 2: Top 20 most important words for feedback classification with their SHAP values.

Phase 4: Application of fine-tuned RoBERTa to the complete dataset

"Low probability" predictions constituted about 329 observations (2.5%) of the total data. After their removal, distribution in the complete dataset was as follows: 52.7% (or $n = 6741$) classified as "vague", and 47.3% (or $n = 6048$) as "actionable."

Phase 5: Differential Analysis of Actionable and Vague Feedback

Logistic regression analysis (Table 3 and Figure 3) revealed several significant associations between linguistic features and feedback actionability. Word count showed a strong negative relationship with actionability (-16.637, $p < .001$), indicating shorter feedback was more likely classified as actionable, contrary to our proposed hypothesis.

Flesch Reading Ease demonstrated a strong positive association with actionability (11.751, $p < .001$), suggesting more readable feedback was more likely to be actionable, aligning with our hypothesis about language complexity.

Lexical diversity showed a moderate positive association (0.418, $p < .001$, odds ratio = 1.52), with more varied vocabulary correlating with actionability. Similarly, modifier count had a significant positive relationship (0.187, $p < .001$, odds ratio = 1.21), suggesting adjectives and adverbs may help describe teaching behaviors with needed precision.

Overall, the model showed a pseudo R^2 of 0.159, accuracy of 0.68, precision of 0.70 (actionable), recall of 0.58, F1-score of 0.63, and an AUC-ROC of 0.76.

6 Conclusion

Our study provides empirical support for computational approaches to analyzing actionable teacher feedback. The high performance of our fine-tuned RoBERTa model (`accuracy` = 0.94, `precision` = 0.90, `recall` = 0.96, `f1 = 0.93`) demonstrates that RoBERTa can effectively distinguish between actionable and vague feedback, even with a relatively modest training dataset of 662 annotated examples.

The SHAP analysis revealed several interesting patterns in the linguistic features associated with actionable feedback. Action verbs (e.g., "struggled," "checks," "encourages") and specific instructional behaviors appeared more frequently in actionable feedback, while comparative and conditional language (e.g., "whether," "enough," "instead") was more characteristic of vague feedback. However, these patterns were not uniformly consistent, suggesting that actionability may be determined more by the relationships between words and phrases rather than by individual word choices alone.

An analysis of linguistic features suggested that contrary to our initial expectations, word count showed a significant negative relationship with actionability, indicating that shorter feedback was

Feature	Coefficient	Std Error	Odds Ratio
Word Count	-16.637***	1.510	5.95×10^{-8}
Flesch Reading Ease	11.751***	1.013	1.3×10^{5}
Lexical Diversity	0.418***	0.029	1.52
Modifier Count	0.187***	0.033	1.21

Table 3: Results of logistic regression predicting feedback actionability (*** $p < .001$)

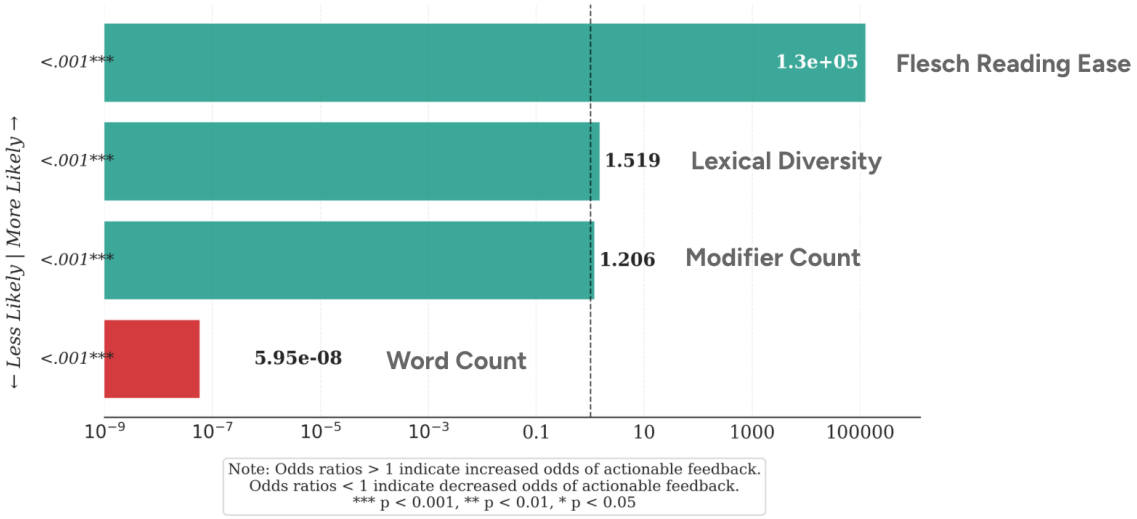

Figure 3: Odds ratios for predictors of actionable feedback

more likely to be classified as actionable. This finding challenges the common assumption that more detailed feedback is necessarily more actionable. It suggests that concision may actually enhance clarity and implementability—verbose feedback might obscure key action points with extraneous information.

The positive association between Flesch Reading Ease and actionability aligns with our hypothesis that more readable feedback is more actionable. This finding indicates that accessible language is crucial for feedback that can be readily understood and implemented.

Our findings suggest three practical implications for teacher professional development. First, our study suggests that concise, readable feedback with precise descriptive language could enhance actionability of feedback given by observers. Second, training programs for classroom observers could benefit from incorporating linguistic guidelines that emphasize readability, appropriate lexical diversity, and effective use of modifiers to enhance feedback actionability. Third, computational approaches like our RoBERTa model could serve as supportive tools for observers to assess and potentially improve the actionability of their feedback before sharing it with teachers, though such applications should complement rather than replace human judgment.

7 Limitations and Future Work

This study has several limitations that point to directions for future research. While our RoBERTa model performed strongly even with 662 annotated examples, the relatively small training set still poses challenges for generalizability. Its success reflects the effectiveness of fine-tuning on well-curated educational data, but broader representation across feedback styles, school contexts, and observer types would strengthen model robustness and reduce the risk of overfitting.

Second, the scope of this study was limited to early primary classrooms (Grades 1–6) and core subjects (English and Math) in a specific cultural setting. Findings may not fully generalize to other grade levels, subjects, or educational systems. Additionally, because the model was trained on English-language feedback, linguistic and cultural differences in how actionability is expressed remain underexplored.

Third, while SHAP analysis revealed some useful patterns, many influential words, especially

in vague feedback, were ambiguous or context-dependent, highlighting the challenge of capturing actionability through isolated word-level features.

Finally, our binary classification approach, while practical, likely oversimplifies the feedback quality spectrum. Actionability may be better understood as a continuum (from highly vague to highly specific). A multi-point ordinal scale (e.g., 5–7 categories) could offer more granular insights, especially for training observers or improving vague feedback. Moving to such a framework would require more complex annotation protocols, higher inter-rater alignment, and substantially larger datasets—but the added nuance may justify this investment by producing models that offer not just detection, but actionable guidance.

Future work should: (1) expand annotations across more diverse educational contexts, (2) test cross-cultural variation in feedback actionability, and (3) explore methods for refining or rewriting vague comments into actionable ones to support professional development more directly.

Acknowledgments

We are deeply grateful to Francisco Carballo Santiago from Rising Academies for providing us with this dataset and assisting in our initial exploration. We also extend our thanks to Sanne Smith and Michael Hardy for their invaluable insights into our research questions, figures, tables, and early drafts of the paper. Finally, we thank our peers for their participation in peer review sessions, which significantly contributed to strengthening our study.

Ethics Statement

Potential Misuse: Our model could be misused in high-stakes evaluations, leading to the automated assessment of observer performance without appropriate human oversight. We explicitly discourage such use and advocate for responsible deployment.
Privacy Considerations: Implementing this technology would require strict privacy protocols to protect teacher identities. Observation data should be de-identified before being fed into the model to ensure confidentiality.

References

C. Adelman and R. Walker. 1975. *A Guide to Classroom Observation (1st ed.)*. Routledge.

Joseph P Allen, Robert C Pianta, Anne Gregory, Amori Yee Mikami, and Janetta Lun. 2011. An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, 333(6045):1034–1037.

J. Archer, S. Cantrell, S. L. Holtzman, J. N. Joe, C. M. Tocci, and J. Wood. 2016. *Better feedback for better teaching: A practical guide to improving classroom observations*. Jossey-Bass, a Wiley Brand.

S. Benslimane, T. Papastergiou, J. Azé, S. Bringay, M. Servajean, and C. Mollevi. 2024. A shap-based controversy analysis through communities on twitter. *World Wide Web*, 27(5).

M. D. Cannon and R. Witherspoon. 2005. Actionable feedback: Unlocking the power of learning and performance improvement. *Academy of Management Perspectives*, 19(2):120–134.

Linda Darling-Hammond, Maria E Hyler, and Madelyn Gardner. 2017. Effective teacher professional development. *Learning policy institute*.

James R. A. Davenport and Robert DeLine. 2014. The readability of tweets and their geographic correlation with education. ArXiv preprint arXiv:1401.6058.

D. Demszky, J. Liu, Z. Mancenido, J. Cohen, H. Hill, D. Jurafsky, and T. Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

R. Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

M. Hardy. 2025. "all that glitters": Approaches to evaluations with unreliable model and human annotations.

Heather C Hill, Charalambos Y Charalambous, and Matthew A Kraft. 2012. When rater reliability is not enough. *Educ. Res.*, 41(2):56–64.

T. J. Kane and D. O. Staiger. 2012. Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.

Avraham N Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol. Bull.*, 119(2):254–284.

Matthew A Kraft, David Blazar, and Dylan Hogan. 2018. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Rev. Educ. Res.*, 88(4):547–588.

V. Lazarev and D. Newman. 2015. How teacher evaluation is affected by class characteristics: Are observations biased? In *Paper presented at the Annual Meeting of AEFP, Washington, DC*.

S. P. Leeman-Munk, E. N. Wiebe, and J. C. Lester. 2014. Assessing elementary students' science competency with text analytics. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, pages 143–147.

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. *CoRR*, abs/1705.07874.

N. Madnani, A. Loukina, A. von Davier, J. Burstein, and A. Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52.

J. Millman and L. Darling-Hammond. 1990. *The new handbook of teacher evaluation*. SAGE Publications, Ltd.

R. T. Putnam and H. Borko. 2000. What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, 29(1):4–15.

M. Shah and A. Pabel. 2019. Making the student voice count: using qualitative student feedback to enhance the student experience. *Journal of Applied Research in Higher Education*, 12(2):194–209.

T. Shaik, X. Tao, Y. Li, C. Dann, J. McDonald, P. Redmond, and L. Galligan. 2022. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access: Practical Innovations, Open Solutions*, 10:56720–56739.

Matthew P Steinberg and Lauren Sartain. 2015. Does teacher evaluation improve school performance? experimental evidence from chicago's excellence in teaching project. *Educ. Finance Policy*, 10(4):535–572.

Abhijit Suresh, Tamara Sumner, Jennifer Jacobs, Bill Foland, and Wayne Ward. 2019. Automating analysis and feedback to improve mathematics teachers' classroom discourse. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9721–9728.

Marieke Thurlings, Marjan Vermeulen, Theo Bastiaens, and Sjef Stijnen. 2013. Understanding feedback: A learning theory perspective. *Educ. Res. Rev.*, 9:1–15.

Deliang Wang, Dapeng Shan, Yaqian Zheng, Kai Guo, Gaowei Chen, and Yu Lu. 2023. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 515–519, Bengaluru, India. International Educational Data Mining Society.

N. E. Winstone, R. A. Nash, M. Parker, and J. Rowntree. 2016. Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, 52(1):17–37.

T. Wragg. 2011. *An Introduction to Classroom Observation (Classic Edition)*. Routledge.

H. Zhang and D. Litman. 2021. Essay quality signals as weak supervision for source-based essay scoring. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–96. Association for Computational Linguistics.

A Appendix

Feedback Type	Rubric
Actionable	• Provides **clear and specific suggestions** for improvement (Archer et al., 2016; Cannon and Witherspoon, 2005) • Offers explicit guidance on: – *What* the teacher should do next – *How* the suggested change can be implemented • Focuses on observable behaviors rather than personality traits (Archer et al., 2016) • Establishes clear connections between observed behaviors and suggested improvements (Cannon and Witherspoon, 2005) • Provides balanced positive and constructive components (Cannon and Witherspoon, 2005) • May or may not contain indicative phrases (e.g., *"even better if," "could have"*); presence of such phrases is not required • Includes concrete examples of alternative approaches
Vague	• Lacks concrete or specific suggestions for improvement (Archer et al., 2016) • Fails to provide clear guidance on implementation steps (Kraft et al., 2018) • May focus on general impressions rather than specific teaching behaviors (Archer et al., 2016) • Lacks explicit connection between observation and suggested change (Cannon and Witherspoon, 2005) • Provides limited or no concrete examples of alternative approaches • May use evaluative language without actionable direction (Allen et al., 2011) • May include general phrases (e.g., *"even better if," "could have"*), but their presence does not ensure clarity; feedback is considered vague if the intended action or direction remains ambiguous or insufficiently specified