Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Peer Feedback Feature Analysis with Large Language Models: An Exploratory Study

Qianru LYU[a], Zirou LIN[b] & Wenli CHEN[a*]
[a] *National Institute of Education, Nanyang Technological University, Singapore*
[b] *Department of Education Information Technology, East China Normal University, China*
*wenli.chen@nie.edu.sg

Abstract: Peer feedback is a pedagogical strategy for peer learning. Despite recent indications of Large Language Models (LLMs) ' potential for content analysis, there is limited empirical exploration of their application in supporting the peer feedback process. This study enhances the analytical approach to peer feedback activities by employing state-of-the-art LLMs for automated peer feedback feature detection. This research critically compares three models—GPT-3.5 Turbo, Gemini 1.0 Pro, and Claude 3 Sonnet—to evaluate their effectiveness in automated peer feedback feature detection. The study involved 69 engineering students from a Singapore university participating in peer feedback activities on the online platform Miro. A total of 535 peer feedback instances were collected and human-coded for eleven features, resulting in a dataset of 5,885 labeled samples. These features included various cognitive and affective dimensions, elaboration, and specificity. The results indicate that GPT-3.5 Turbo is the most effective model, offering the best combination of performance and cost-effectiveness. Gemini 1.0 Pro also presents a viable option with its higher throughput and larger context window, making it particularly suitable for educational contexts with smaller sample sizes. Conversely, Claude 3 Sonnet, despite its larger context window, is less competitive due to higher costs and lower performance, and its lack of support for training and fine-tuning with researchers' data weakens its learning capabilities. This research contributes to the fields of AI in education and peer feedback by exploring the use of LLMs for automated analysis. It highlights the feasibility of employing and fine-tuning existing LLMs to support pedagogical design and evaluations from a process-oriented perspective.

Keywords: Peer feedback, Large Language Models (LLMs), GPT, Gemini, Claude, peer feedback features

1. Introduction

Peer feedback, as a widely applied pedagogical design, is usually defined as written comments from classmates that identify the strengths and weaknesses of a document and suggest constructive improvements (Wu & Schunn, 2020a). By engaging in these activities, students develop critical thinking and evaluative skills as they assess their peers' work and reflect on their own (Nicol et al., 2014; Tai et al., 2016). This process encourages deeper understanding and knowledge construction, as students are required to articulate their thoughts clearly and constructively (Boud & Molloy, 2013). Additionally, peer feedback fosters a collaborative learning environment, promoting essential communication and teamwork skills (Carless & Boud, 2018). It also increases students' learning motivation and engagement by making learning more interactive and giving students a sense of ownership over their learning (van Gennip et al., 2010). Furthermore, the practice of giving and receiving feedback helps students develop feedback literacy, enabling them to use feedback effectively to improve their work (Mutch et al., 2018).

However, engaging in peer feedback activities does not lead to learning improvement. Accumulating research has pointed out the significant roles of the different peer feedback features on students' learning improvement (Chen et al., 2022; 2024; Gielen & De Wever, 2015; Lyu, et al., 2023; Schunn & Wu, 2020a; Strijbos et al., 2010). For example, more elaborated warm toned peer feedback was found to enhance learning (De Sixte et al., 2019). The clarification and suggestion for improvement were also beneficial for knowledge improvement (Chen et al., 2022; Lyu et al., 2023). Most peer feedback features analysis remains a post-learning analysis, which may not provide instant valuable analytics for students and allow them to make immediate revisions. When there are limited instructor-student ratios in authentic classrooms, it could be difficult to conduct real-time analysis for students' peer feedback content, and accordingly improve their peer feedback regarding the various features to enhance students' learning improvement.

Thanks to the rapid development of AI techniques based on large language models (LLMs) such as ChatGPT from OpenAI, Gemini from Google, and Claude from Microsoft, these models enable real-time text analysis, making it possible to analyze peer feedback content in a very short time. This technology advancement opens new opportunities for providing instant evaluation and analysis for students' learning data, such as peer conversations, online logs as well as learning artefacts (e.g., write-ups) (Grassini, 2023). There is a possibility to enhance students' peer feedback efficiency by automated peer feature analysis by leveraging the LLMs (Bauer et al., 2023).

Though some recent studies indicated the possibility to introduce the LLMs for learning content analysis (Mayer et al., 2023), there remains limited empirical explorations of applying LLMs to support peer feedback process. Some very recent studies such as Hutt et al. (2024) employed LLMs to evaluate the quality of peer feedback content. However, there is a lack of investigation on a fine-grained analysis of peer feedback features with LLMs, which could provide valuable insights for educators as well as actionable suggestions for students. Another underexplored question is how to fine-tune the LLMs for peer feedback feature analysis with the limited data sample size (e.g., student number, lesson number) in authentic education context. For the peer feedback feature analysis task with limited learning data, what are the actual performance of the various LLMs? These questions are difficult ones for educators to employ the trending models to support the everyday peer feedback activities. Therefore, this study aims to explore the feasibility and effectiveness of applying LLMs for peer feedback feature analysis. The guiding research question is: To what extent do the LLMs support peer feedback features analysis in university classroom context?

2. Method

2.1 Participants and learning context

The participants were 69 engineering students from a Singapore university. The study was approved by university IRB committee (IRB-2020-04-031) and all participants indicated their consent by signing consent forms. In a mechanical engineering course, they participated in in-class peer feedback activities on the online platform Miro, commenting on each other's design solutions for an automated vehicle.

2.2 Data collection

All participants' peer feedback on the Miro platform and their design solutions were collected for content analysis. A total of 535 peer feedback were collected.

2.3 Data analysis

2.3.1 Training data preparation

Qualitative content analysis was conducted to identify the peer feedback features, with every slice of peer feedback as one unit of analysis. The coding scheme of peer feedback features is based on the peer feedback literature (Chen et al., 2022; Gielen & De Wever, 2015; Lyu, et al., 2023; Schunn & Wu, 2020a; Strijbos et al., 2010). A total of eleven peer feedback features were human coded in a binary way (code 0/code 1): the cognitive dimension (positive evaluation (Cohen's kappa: 0.96); neutral evaluation (Cohen's kappa: 0.82); negative evaluation (Cohen's kappa: 0.79); problem identification (Cohen's kappa: 0.81); suggestion (Cohen's kappa: 0.94); solution (Cohen's kappa: 0.86) and clarification (Cohen's kappa: 0.77)), the affective dimension (hedge (Cohen's kappa: 0.71); mitigation (Cohen's kappa: 0.82)), elaboration (Cohen's kappa: 0.79) and specificity (Cohen's kappa: 0.90). Eventually, human coders produced eleven features encoding results for each peer feedback sample, and finally produced 5885 data samples with labels.

2.3.2 Models selection

In this study, we selected three comparable AI models—GPT-3.5 Turbo, Gemini 1.0 Pro, and Claude 3 Sonnet—to analyze peer feedback, considering factors such as data size, API costs, and potential performance. Both GPT-3.5 Turbo and Gemini 1.0 Pro were fine-tuned using our own training data to better detect peer feedback features, whereas Claude 3 Sonnet does not support fine-tuning. In terms of context window size, Claude 3 Sonnet offers a significant advantage with a 200K token limit, compared to 33K for Gemini 1.0 Pro and 16K for GPT-3.5 Turbo, making it more suitable for processing extensive texts. However, this comes at a higher cost, with Claude 3 Sonnet charging $3.00 per million input tokens and $15.00 per million output tokens, while GPT-3.5 Turbo and Gemini 1.0 Pro are more economical, both priced at $0.50 per million input tokens and $1.5 per million output tokens. In terms of throughput, which measures output tokens per second, Gemini 1.0 Pro leads with 82 tokens per second, followed by Claude 3 Sonnet at 61 tokens per second, and GPT-3.5 Turbo at 52 tokens per second. These metrics highlight the trade-offs between processing capacity, cost efficiency, and performance across the three models.

2.3.3 Experiment setup

To explore the performance of above LLMs for peer feedback features analysis, this study utilized peer feedback data that had been previously human-coded. This dataset was divided into three parts in the ratio of 7:2:1, with 70% for training, 20% for validation, and 10% for testing. The 7:2:1 ratio ensures ample training data while providing sufficient data for validation and testing. This ratio strikes a relative balance, comprehensively considering various factors in the current

research context, such as dataset size, task complexity, and computational resources. The fine-tuning process involved training each model on the peer feedback dataset to enhance higher quality to detect specific feedback features. For GPT-3.5 Turbo and Gemini 1.0 Pro, different number of examples are included in the fine-tune process. Since Claude 3 Sonnet does not support fine-tuning with human-coded data up to now, we directly used all data samples for prompting. For all models, we construct the same prompts by Python script with API setting according to users guide. This study analysed the models' performance using key metrics such as accuracy, precision, recall, and F1 score.

3. Results

The performance of the models was evaluated and compared in terms of accuracy, precision, recall, and F1 score. The results are summarized in the table 1 below:

Table 1. *Summary Table of Models Performance*

Models	Accuracy	Precision	Recall	F1 Score
GPT-3.5 Turbo	0.88	0.88	0.88	0.88
Gemini 1.0 Pro	0.83	0.82	0.83	0.83
Claude 3 Sonnet	0.58	0.41	0.77	0.53

GPT-3.5 Turbo outperformed the other two models, achieving the highest scores across all metrics, indicating superior performance in detecting peer feedback features. Gemini 1.0 Pro also performed well but with slightly lower metrics, while Claude 3 Sonnet had the lowest performance among the three models. This study also compared the models based on their operational efficiency, focusing on the context window, cost, and throughput, as previously provided in the table 2. Gemini 1.0 Pro demonstrated the highest throughput with 82 output tokens per second and a larger context window of 33K tokens, making it cost-effective despite its slightly lower performance metrics compared to GPT-3.5 Turbo. It is worth to mention that the maximum limit on data examples is 500 up to now as required by Gemini 1.0 Pro. Therefore, it only requires a very small amount of data input to achieve a relatively ideal accuracy rate. GPT-3.5 Turbo, with a context window of 16K tokens and a cost of $0.50 per input and $1.5 per output per million tokens, showed balanced performance and operational efficiency. Claude 3 Sonnet, despite having the largest context window of 200K tokens, had the highest cost and lower throughput, making it less favourable for this application.

4. Discussion

To enhance the current analytical approach of peer feedback activities, this study employs the state-of-the-art LLMs to realize automated peer feedback features detection. A critical comparison was conducted among the three models GPT-3.5 Turbo, Gemini 1.0 Pro, and Claude 3 Sonnet. The results indicate that GPT-3.5 Turbo is the most effective model for automated detection of peer feedback features, offering the best combination of performance and cost-effectiveness. However, Gemini 1.0 Pro also presents a viable option with its higher throughput and larger context window, making it particularly suited to the education domain, where research contexts typically have a smaller sample size. Claude 3 Sonnet, while offering a significantly larger context

window, is less competitive due to its higher costs and lower performance. In addition, the lack of support for training and fine-tuning with the researcher's data weakens its learning capabilities to some extent. These findings suggest that LLMs can indeed support the automated detection of peer feedback features, with GPT-3.5 Turbo and Gemini 1.0 Pro being the most promising candidates for practical implementation.

This study contributes to AI in education and peer feedback research fields by exploring the use of LLMs for automated peer feedback analysis. As for the AI in education field, this study highlights the feasibility of employing and fine-tuning the existing LLMs to support pedagogical design and evaluations from a process-oriented perspective. This study demonstrates that educators, even without deep knowledge and expertise in natural language processing, can access and employ the current LLMs for their daily classroom practices and evaluations.

With regard to the contribution to the peer feedback research, these LLMs can be integrated with existing peer feedback implementation/learning improvement models (e.g., Wu & Schunn, 2020b; Wu & Schunn, 2021; Fong & Schallert, 2023). This integration will make it possible to evaluate peer feedback in real-time, providing immediate insights and predictions about learning improvements and knowledge gains (Lin et al., 2024), which is more timely compared to traditional post-class assessments. In summary, using LLMs for real-time peer feedback feature evaluation offers a practical advancement in peer feedback research and AI for education practices. It provides insights for the adoption of the state-of-the-art LLMs to support students' learning efficiency during peer feedback activities.

While this study provides valuable insights into the use of LLMs for automated peer feedback analysis, the limitations regarding sample size, contextual applicability, the range of models tested, and ethical considerations highlight the need for further research. Firstly, the sample size is relatively small, consisting of only 535 slices for each peer feedback features. This limited sample may not fully capture the diversity of peer feedback practices across different educational contexts. Future research could include a larger and more varied dataset to validate and generalize the findings. Secondly, the study was conducted in a specific context of engineering classrooms. The unique characteristics and dynamics of this educational setting may not be directly transferable to other disciplines or educational levels. Different subject areas and classroom environments may present distinct challenges and nuances in peer feedback processes that were not addressed in this study. Thirdly, this study only included three LLMs: GPT-3.5 Turbo, Gemini 1.0 Pro, and Claude 3 Sonnet. While these models provide valuable insights, a more comprehensive analysis involving a greater number and variety of models is necessary to fully understand the capabilities and limitations of LLM-based AI for peer feedback analysis. Future studies may compare a wider range of models to determine the best fit for different educational applications.

Acknowledgements

This study was funded by the EdeX Teaching and Learning (T&L) Grant (#021799-00001) supported by Nanyang Technological University, Singapore.

References

Bauer, E., Greisel, M., Kuznetsov, I., Berndt, M., Kollar, I., Dresel, M., Fischer, M. & Fischer, F. (2023). Using natural language processing to support peer-feedback in the age of artificial intelligence: a cross-disciplinary framework and a research agenda. *British Journal of Educational Technology*, 54(5), 1222-1245.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. Educational Psychology Review, 38(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chen, W., Lyu, Q., Su, J., Aileen, C. S. C., Zhang, W., Su, G., & Li, X. (2023). How Does Feedback Formulation Pattern Differ between More-Improvement and No-Improvement Student Groups? An Exploratory Study. *In Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023,* pp. 577-584. International Society of the Learning Sciences. https://doi.org/10.22318/icls2023.747522

Chen, W., Lyu, Q., & Su, J. (2024). How more-improvement and less-improvement groups differ in peer feedback giving and receiving practice-an exploratory study. *Instructional Science*, 1-21. https://doi.org/10.1007/s11251-024-09667-7

De Sixte, R., Mañá, A., Ávila, V., & Sánchez, E. (2020). Warm elaborated feedback. Exploring its benefits on post-feedback behaviour. *Educational Psychology*, 40(9), 1094-1112.

Fong, C. J., & Schallert, D. L. (2023). "Feedback to the future": Advancing motivational and emotional perspectives in feedback research. *Educational Psychologist*, 58(3), 146-161.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. Learning and instruction, 20(4), 304-315.

Grassini, S. (2023). Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, 13(7), 692.

Hutt, S., DePiro, A., Wang, J., Rhodes, S., Baker, R.S., Hieb, G., Sethuraman, S., Ocumpaugh, J. and Mills, C. 2024. Feedback on Feedback: Comparing Classic Natural Language Processing and Generative AI to Evaluate Peer Feedback. *Proceedings of the 14th Learning Analytics and Knowledge Conference* (Kyoto Japan, 2024), 55–65.

Lin, Z., Yan, H., & Zhao, L. (2024). Exploring an effective automated grading model with reliability detection for large-scale online peer assessment. *Journal of Computer Assisted Learning.* https://doi.org/10.1111/jcal.12970

Lyu, Q., Chen, W., Su, J., & Heng, K. H. J. G. (2023). Steps to implementation: the role of peer feedback inner structure on feedback implementation. *Assessment & Evaluation in Higher Education*, 1–14. https://doi.org/10.1080/02602938.2023.2291340

Mayer, C. W. F., Ludwig, S., & Brandt, S. (2023). Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. *Journal of Research on Technology in Education*, 55(1), 125–141. https://doi.org/10.1080/15391523.2022.2142872

Mutch, A., Young, C., Davey, T., & Fitzgerald, L. (2018). A journey towards sustainable feedback. *Assessment & Evaluation in Higher Education*, 43(2), 248–259. https://doi.org/10.1080/02602938.2017.1332154

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment & Evaluation in Higher Education*, 39(1), 102–122. https://doi.org/10.1080/02602938.2013.795518

Tai, J. H.-M., Canny, B. J., Haines, T. P., & Molloy, E. K. (2016). The role of peer-assisted learning in building evaluative judgement: Opportunities in clinical medical education. *Advances in Health Sciences Education*, 21(3), 659–676. https://doi.org/10.1007/s10459-015-9659-0

Strijbos, J. W., & Sluijsmans, D. M. A. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction*, 20, 265-269.

van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction*, 20(4), 280–290. https://doi.org/10.1016/j.learninstruc.2009.08.010

Wu, Y., & Schunn, C.D. (2020) a. When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. C*ontemporary Educational Psychology*, 62, 101897.

Wu, Y., & Schunn, C. D. (2020) b. From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, 60, 101826.
Wu, Y., & Schunn, C. D. (2021). From plans to actions: A process model for why feedback features influence feedback implementation. *Instructional Science*, 49(3), 365-394.