LLM-generated Feedback in Real Classes and Beyond: Perspectives from Students and Instructors

Qinjin Jia, Jialin Cui, Haoze Du, Parvez Rashid, Ruijie Xi, Ruochi Li, Edward Gehringer
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
qjia3, efg@ncsu.edu

ABSTRACT

Feedback plays a crucial role in education, offering students explicit guidance on how to enhance their academic performance. In pursuit of providing feedback promptly and efficiently, researchers are actively exploring the use of large language models (LLMs) to automatically generate feedback on student work. However, the deployment of such automated feedback systems in actual classrooms is nascent, and they have yet to survey student and instructor perspectives, thus leaving their limitations in real educational settings unclear. In this paper, we deploy a system that generates feedback for student project reports in a graduate-level computer science course and collect perspectives from authentic users. We solicited student opinions on the generated feedback through questionnaires and engaged the course instructor to delve into their perceptions regarding the alignment of the feedback with their pedagogical objectives. Our work sheds light on the potential impact and limitations of system-generated feedback in real-world educational settings and contributes insights for future research on automated feedback systems.

Keywords

Automated feedback generation, automated review generation, large-language models, education, survey

1. INTRODUCTION

In the realm of education, feedback refers to information provided by individuals (e.g., instructors or peers) regarding students' performance in learning activities [1, 8, 21]. Feedback is integral in guiding students through their learning process, offering insights that enable them to strengthen or correct their understanding of knowledge and content [8, 14, 21]. However, providing quality feedback, especially for assignments that lack straightforward answers, requires considerable effort and educational resources and often faces challenges in being delivered promptly. The demand for immediate and cost-efficient feedback solutions has driven the development of automated feedback systems [9, 12]. With recent advancements in large language models (LLMs), researchers have designed LLM-based systems and shown that generated feedback closely resembles human feedback [3, 10].

However, most LLM-based automated feedback systems have not been deployed in actual classroom settings, thus their applicability in real-world contexts remain unclear [3, 9, 10, 19, 30, 7]. These studies often employ automatic metrics (e.g., ROUGE scores [17], BERTScore [31]) and human judgment across different quality dimensions (e.g., readability, faithfulness) to evaluate the feedback. Although these evaluations may reflect some aspects of the quality of generated feedback, they lack perspectives from students and teachers in real classrooms [2, 5]. Consequently, their evaluation may not fully align with or comprehensively represent the authentic user experience in real educational contexts. This gap raises the question of whether these automated feedback systems can function effectively in real classrooms and provide meaningful assistance to students and instructors.

In this paper, our primary objective is to gain a comprehensive understanding of the potential impact and limitations of automated feedback systems from perspectives of students and instructors in practical educational settings. To this end, we deploy an LLM-based automated feedback system to generate textual feedback on student project reports in a graduate-level computer science course. We then solicit student opinions on the generated feedback through a questionnaire, and engage with the course instructor to delve into their perspectives regarding the alignment of the feedback with their pedagogical objectives and methodologies. Through these, we aspire to contribute the insights derived from this study towards enhancing and optimizing feedback-generation systems to better fulfill the needs of both students and instructors in real-world educational scenarios.

To understand the impact and limitations of LLM-generated feedback in real educational settings, our work seeks to answer the following research questions (RQs):
RQ1: How do students and instructors perceive the system-generated feedback in actual classrooms? Do they find the system-generated feedback helpful for student learning?
RQ2: What limitations do students and instructors identify with the system-generated feedback? What are their specific concerns regarding the system-generated feedback?
RQ3: What insights can be drawn from the perspectives of students and instructors? In what areas can LLM-based automated feedback systems be improved in the future?

Q. Jia, J. Cui, H. Du, P. Rashid, R. Xi, R. Li, and E. Gehringer. Llm-generated feedback in real classes and beyond: Perspectives from students and instructors. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 862–867, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729974

Table 1: Samples of system-generated feedback for student project reports.

Generated Feedback 1: "The writeup is quite readable, and it explains the rationale for the design of the project in a readable way. The design doc is divided into sections, with sections on the design, implementation, and testing. It would have been better to describe the functionality in more detail, e.g., how it works, and what changes are made to the code. In general, the design document is quite good."
Generated Feedback 2: "Good job of describing the changes to be made, and explaining why they were made. Would have been helpful to have more detail on the code changes, e.g., what was changed, and why. It would be better to link to the code in Github, where it would be easier to see what was changed. Also, it would have been good to have a summary of what was done, rather than just a list of changes."
Generated Feedback 3: "The writeup is quite readable, and the changes are described in detail. The code changes are clearly described, and they are described with sufficient detail. It would have been better to link to the Github diff view to show the code that has been changed. Also, I would have liked to see more explanation of the tests, e.g., what are the tests for late_policy_spec.rb, and why are they needed?"

The main conclusions are as follows: Both students and instructors find the system-generated feedback to be valuable, yet highlight some areas requiring enhancement. First, a major concern is the presence of unfaithful content in the feedback (i.e., prose that is erroneous, misleading, or entirely irrelevant to the student work). Second, students indicate a preference for more specific feedback tailored to their individual projects, rather than generic comments. Third, there is a consensus that automated feedback, while helpful, cannot fully replace the nuanced insights offered by instructors. Lastly, students express a desire for an interactive feedback system that can address follow-up questions they may have.

Our primary contributions are: 1) we conduct an empirical study that deploys an LLM-based automated feedback system in a graduate-level computer science course, collecting authentic user experiences that support the effectiveness of generated feedback in real classroom settings; 2) perspectives from students and instructors highlight the strengths of the generated feedback in actual classes while also revealing limitations such as the presence of unfaithful content and lack of project specificity; 3) our analysis contributes insights for future research on automated feedback systems, steering them towards better fulfilling the educational needs and expectations of students and instructors in real classes.

The paper is organized as follows: Section 2 first outlines the methodology and procedures employed in this study, including research setup, how feedback is generated, questionnaire design, and interview protocol. Subsequently, Section 3 presents the survey results and interview summary. After that, Section 4 analyzes perspectives from students and instructors, and discusses four insights drawn from their opinions. Finally, Section 5 concludes the paper, highlights the limitations of our work, and discusses future directions.

2. STUDY DESIGN

2.1 Study Setup

This study was approved by the IRB (institutional review board) and conducted in an object-oriented graduate-level course at a public university in the United States. A total of 82 students participated in the research, divided into 28 groups to work on course projects. The projects covered in this study involved students engaging in activities such as refactoring code, integrating new features and functionalities, or implementing automated unit tests for software modules. For the course-project deliverables, each group was required to submit a group report documenting the work completed, the methodologies they used, and other project-related materials, such as how they tested their programs.

The overall process was as follows: students first uploaded near-finalized projects, upon which their peers provided feedback based on rubric criteria devised by the instructor. Following the peer feedback, students revised their reports and submitted the final versions. Subsequently, an automated feedback-generation system produced feedback for the final reports and disseminated it to the students and the instructor via email. The questionnaire was simultaneously sent to each group, who were asked to complete it within one week. At the same time, the instructor examined the project reports and provided feedback. After the instructor reviewed all reports, we conducted an interview with the instructor.

2.2 Feedback Generation

Two cutting-edge methods for implementing LLM-based feedback generation systems are data-driven and prompt-driven approaches, which correspond to two distinct strategies for customizing LLMs for feedback generation tasks: *pre-training and then fine-tuning* [22, 29, 15] and *prompt engineering* [23, 24, 4, 3]. The data-driven systems involve further training LLMs on student work and feedback data to learn underlying patterns for producing feedback [10, 6]. The prompt-driven systems rely on human-designed prompts to guide LLMs in generating feedback [3, 20]. This study used a data-driven system because our preliminary experiment found it could better mimic the tone and style of instructor feedback.

Specifically, we generate feedback for student project reports by utilizing a BART-based data-driven automated feedback system, as described in [10]. The BART model is an encoder-decoder LLM [15], which is adept at capturing relationships from one sequence of text (e.g., project reports) to another (e.g., feedback). We fine-tuned the model with 484 pairs of project reports and feedback. In addition, recognizing the absence of effective evaluation metrics for text generation [2], we generated three sets of feedback by leveraging different combinations of decoding strategies and hyperparameters, and manually selected the best feedback from them. Table 1 above exhibits three samples of system-generated feedback.

Table 2: All questions in the questionnaire, their average scores (Avg.) and standard deviations (SD).

Question (Scale)	Question Description	Avg.	SD
Q1. Overall Score (1–10)	Please provide an overall score for the feedback.	8.14	2.28
Q2. Helpfulness (1–5)	How helpful do you think the feedback is?	4.04	1.21
Q3. Faithfulness (1–5)	How accurate do you think the feedback is?	3.89	1.14
Q4. Comprehensiveness (1–5)	How comprehensive do you think the feedback is?	4.25	1.02
Q5. Replaceability (1–5)	Do you think the generated feedback can replace the instructor feedback?	3.25	1.45
Q6. Open Question (text)	Please give your detailed comments on the feedback.	–	–

2.3 Survey Design

To gather student opinions on the system-generated feedback for their project reports, we designed a questionnaire incorporating a series of Likert scale questions and an open-ended response. In constructing the questionnaire and formulating the questions, we followed the guidance outlined in [13]. Additionally, we opted for an anonymous survey to increase response rates, encourage honest responses, and mitigate privacy concerns [26, 27]. Students were informed, both verbally and in writing, that participation in the survey was entirely voluntary. Their responses were collected electronically using Google Forms, and the results were securely stored on a cloud service administered by our institution.

Besides the IRB information and consent form at the beginning of the questionnaire, each survey consists of six questions. The specific survey questions are listed in Table 2. The first question (Q1) asks students to provide an overall rating of the feedback. Subsequently, Q2–Q4 engage students in assessing the feedback from the perspectives of helpfulness, faithfulness, and comprehensiveness, respectively. Q5 probes whether students perceive the system-generated feedback as an effective substitute for instructor feedback. Lastly, Q6 is an open-ended question that allows students to offer textual feedback, providing explanations for their ratings or suggestions for future improvements to the system.

2.4 Interview Protocol

The insights of the course instructor are of paramount significance in comprehending the practical implications of the LLM-generated feedback within an educational framework. We followed the guidance outlined in [11], and conducted a qualitative semi-structured interview with the course instructor to delve into his nuanced perspectives. The semi-structured interview offers a balanced approach that combines structure with flexibility, allowing for thorough exploration of topics while accommodating the unique perspectives and experiences of the instructor. This interview was performed at the end of the course, after the instructor had reviewed all project reports and LLM-generated feedback.

In the interview, we first elicited the instructor's overall impressions of the LLM-generated feedback. Then, we delved into several specific aspects of the feedback, including its helpfulness, faithfulness, specificity, and comprehensiveness. Finally, we discussed with the instructor regarding his perspectives on how to integrate automated feedback systems into existing pedagogical frameworks and methodologies, as well as explored potential synergies or challenges that may arise in real-world applications. Throughout the interview, we also engaged in impromptu discussions to probe further into viewpoints and insights mentioned by the instructor.

3. RESULTS

3.1 Survey results

For our survey, we received a total of 28 responses, with 23 of them providing detailed textual comments. The average score and standard deviation for each Likert scale question are presented in Table 2. The distribution of scores for each Likert scale question can be found in Appendix A. Representative snippets of textual comments are shown interspersed throughout the following analysis to illustrate the results.

Overall, responses to Q1 indicate a prevailing satisfaction with the LLM-generated feedback. The majority of students (75.0%) rated the feedback 8 or above out of 10. Students commended the feedback for its fluency, coherence, and effectiveness in guiding them to enhance their project reports.

- *"The feedback is detailed and very helpful as an initial guidance to improve our report for the final project."*
- *"It is very impressive. It is almost like a real person's feedback. I found the feedback accurate to some extent. However, some part of the feedback was not clear."*

The responses to Q2 (Helpfulness) suggest that students perceive the LLM-generated feedback as helpful, with most students (78.6%) rating it as 4 or higher out of 5. However, student comments suggest that the primary factor diminishing the helpfulness of the feedback is lack of project specificity.

- *"It suggested me some changes for the code comparison addition which seemed helpful but I felt like the feedback seemed to be generic rather than document specific.*"

The answers to Q3 (Faithfulness) indicate that students have a concern about the faithfulness of the feedback. The average score for this question is 3.89 out of 5, and 67.9% of students scored it as 4 or above. Students perceive that such unfaithful content undermine the reliability of the feedback.

- *"The feedback said 'Include screenshots' and 'More explanation of the tests.' But they were already in place."*

The responses to Q4 (Comprehensiveness) demonstrate satisfaction with the comprehensiveness of the feedback. Nearly all students (92.9%) rated it as 4 or higher out of 5. However, the results may be biased as students may not be aware of all the aspects that comprehensive feedback could encompass.

- *"The tool highlights some areas of improvements but was not able to accurately provide all errors and suggestions. The feedback obtained by mentors and the professor during the demo was much more insightful."*

The ratings for Q5 (Replaceability) show that students have varied perspectives on whether the generated feedback can replace the instructor feedback. Some students (50% of them rated 4 or above out of 5) believe it can, while others feel that instructors can provide more helpful and nuanced insights.

- *"The feedback puts light on the tests not being described in detail, I think it was a good point but the particulars that we get from peers, and professor are better where there is someone actually monitoring your work."*

3.2 Interview Summary

The following points were distilled from the semi-structured interview with the instructor. First, the instructor expressed cautious optimism regarding the LLM-generated feedback. While acknowledging its potential to assist students in improving project reports, the instructor highlighted the need for feedback to be faithful and more specific. Second, the instructor articulated an expectation for the flexibility to adjust various aspects of the generated feedback, such as its tone and areas of focus, to better align with instructional objectives. Lastly, regarding the integration of automated feedback systems into existing pedagogical frameworks, the instructor considered that directly sending the generated feedback to students still poses risks. Thus, the instructor deemed the most pragmatic approach to be using the LLM-generated feedback as the initial feedback draft.

4. ANALYSIS AND DISCUSSION

Insight 1: Faithfulness is a primary concern. The first insight from our analysis is the concern regarding faithfulness. Despite the LLM-generated feedback closely resembling instructor feedback in fluency and coherence, it may contain hallucinated content that is erroneous, misleading, or entirely irrelevant to the original project reports. Both the students and the instructor in this study expressed apprehension that such hallucinated content could confuse or mislead students, and consider it to significantly impact the reliability of the automated feedback system. To address the concern of hallucination, future research could attempt to understand the underlying causes of hallucination [28], and explore various hallucination detection and post-editing techniques [18]. These efforts can pave the way for the deployment of automated feedback systems in real classrooms.

Insight 2: Project-specific feedback is expected. The second insight underscores that the LLM-generated feedback tends to be generic. The students expressed a strong preference for feedback that is tailored to their project reports, as they felt that specific and detailed feedback would be more helpful for their learning. This sentiment was echoed by the instructor during the interview, further emphasizing the significance of project specificity. We speculate that the lack of specificity is primarily due to the feedback-generation system itself lacking relevant knowledge to provide detailed feedback. We also observed from preliminary experiments that forcing an increase in specificity appears to decrease its faithfulness. To this end, future research could explore techniques such as retrieval-augmented generation (RAG) [16] to incorporate knowledge for providing more specific feedback.

Insight 3: Human feedback remains irreplaceable. The third insight emphasizes the irreplaceable value of human feedback. While LLM-generated feedback offers valuable assistance, the students believe it cannot replace the nuanced understanding and personalized insights provided by human instructors. The instructor also felt that the feedback did not always align with the pedagogical objectives. Therefore, while automated feedback systems can supplement and enhance the learning process, they currently cannot entirely replace the role of human instructors. Instead, a synergistic approach may be adopted, where automated feedback systems support and complement human feedback, maximizing the benefits of both approaches. For example, the instructor found that using automated feedback as an anchor when personally crafting feedback reduced the time and effort needed.

Insight 4: Interactive feedback could be the way forward. The last insight is that interactive feedback can be a promising direction for future improvements. In the survey, some students expressed a desire for the feedback system to allow them to inquire about specific aspects of the feedback that they find confusing or in need of further explanation. Existing static feedback, although valuable, lacks the flexibility to address individual concerns or queries. In contrast, interactive feedback systems offer the potential to engage students more actively in the feedback process by providing them with opportunities to seek clarification and additional information as needed, which may enhance the personalization and effectiveness of feedback. Thus, the focus of future research may shift from static feedback systems to interactive feedback systems that enable dynamic student engagement.

5. CONCLUSION

LLM-based automated feedback systems hold immense potential for transforming the educational landscape [25]. However, the absence of perspectives from students and instructors leaves uncertainties regarding their effectiveness and limitations in practical educational settings. To bridge this gap and gain practical insights, we deployed an LLM-based automated feedback system in a graduate-level computer science course. We then solicited student opinions on the generated feedback through a questionnaire and conducted a semi-structured interview with the course instructor. The results demonstrate the value of the system-generated feedback, but also reveal areas requiring enhancement, such as faithfulness and specificity. This study contributes insights towards improving feedback-generation systems to better fulfill the needs of students and instructors in real classes.

Limitations and future work: There are two main limitations to this study. Firstly, the study was conducted in a single graduate-level computer science course at one university. This limits the generalizability of the findings to other educational contexts, particularly those with different subject matters, student populations, or instructional approaches. Therefore, future research can be conducted in different environments. Second, this study relies on anonymous questionnaire surveys and semi-structured interviews for collecting opinions. While these methods provide valuable insights, they may not capture the full spectrum of perspectives or experiences, and the findings may be subject to bias or limitations inherent in self-reporting. Thus, future studies could consider complementing these methods with additional quantitative measures or extrinsic evaluations.

6. REFERENCES

[1] W. Alharbi. E-feedback as a scaffolding teaching strategy in the online language classroom. *Journal of Educational Technology Systems*, 46(2):239–251, 2017.

[2] A. Celikyilmaz, E. Clark, and J. Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.

[3] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, and G. Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE, 2023.

[4] T. Gao, A. Fisch, and D. Chen. Making pre-trained language models better few-shot learners. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, pages 3816–3830. Association for Computational Linguistics (ACL), 2021.

[5] C. Garbacea and Q. Mei. Neural language generation: Formulation, methods, and evaluation. *arXiv preprint arXiv:2007.15780*, 2020.

[6] S. Gombert, A. Fink, T. Giorgashvili, I. Jivet, D. Di Mitri, J. Yau, A. Frey, and H. Drachsler. From the automated assessment of student essay content to highly informative feedback: a case study. *International Journal of Artificial Intelligence in Education*, pages 1–39, 2024.

[7] J. Han, H. Yoo, J. Myung, M. Kim, H. Lim, Y. Kim, T. Y. Lee, H. Hong, J. Kim, S.-Y. Ahn, et al. Fabric: Automated scoring and feedback generation for essays. *arXiv preprint arXiv:2310.05191*, 2023.

[8] J. Hattie and H. Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[9] Q. Jia, M. Young, Y. Xiao, J. Cui, C. Liu, P. Rashid, and E. Gehringer. Insta-reviewer: A data-driven approach for generating instant feedback on students' project reports. *International Educational Data Mining Society*, 2022.

[10] Q. Jia, M. Young, Y. Xiao, J. Cui, C. Liu, P. Rashid, E. Gehringer, et al. Automated feedback generation for student project reports: A data-driven approach. *Journal of Educational Data Mining*, 14(3):132–161, 2022.

[11] H. Kallio, A.-M. Pietilä, M. Johnson, and M. Kangasniemi. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of advanced nursing*, 72(12):2954–2965, 2016.

[12] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[13] B. A. Kitchenham and S. L. Pfleeger. Personal opinion surveys. In *Guide to advanced empirical software engineering*, pages 63–92. Springer, 2008.

[14] S. Kusairi. A web-based formative feedback system development by utilizing isomorphic multiple choice items to support physics teaching and learning. *Journal of Technology and Science Education*, 10(1):117–126, 2020.

[15] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[17] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In M.-F. Moens and S. Szpakowicz, editors, *Text summarization branches out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

[18] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, 2020.

[19] H. McNichols, W. Feng, J. Lee, A. Scarlatos, D. Smith, S. Woodhead, and A. Lan. Exploring automated distractor and feedback generation for math multiple-choice questions via in-context learning. *arXiv preprint arXiv:2308.03234*, 2023.

[20] T. Phung, J. Cambronero, S. Gulwani, T. Kohn, R. Majumdar, A. Singla, and G. Soares. Generating high-precision feedback for programming syntax errors using large language models. *arXiv preprint arXiv:2302.04662*, 2023.

[21] P. Race. Using feedback to help students to learn. *The Higher Education Academy*, 2001.

[22] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[25] J. Roschelle, J. Lester, and J. Fusco. Ai and the future of learning: Expert panel report., 2020. https://circls.org/reports/ai-report.

[26] T. Scherer, J. Straub, D. Schnyder, and N. Schaffner. The effects of anonymity on student ratings of teaching and course quality in a bachelor degree programme. *GMS Zeitschrift für Medizinische Ausbildung*, 30(3), 2013.

[27] P. K. Tyagi. The effects of appeals, anonymity, and feedback on mail survey response patterns from salespeople. *Journal of the Academy of Marketing Science*, 17:235–241, 1989.

[28] Z. Xu, S. Jain, and M. Kankanhalli. Hallucination is inevitable: An innate limitation of large language

models. *arXiv preprint arXiv:2401.11817*, 2024.

[29] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[30] S.-Y. Yoon, E. Miszoglad, and L. R. Pierce. Evaluation of chatgpt feedback on ell writers' coherence and cohesion. *arXiv preprint arXiv:2310.06505*, 2023.

[31] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

APPENDIX

A. THE DISTRIBUTION OF SCORES

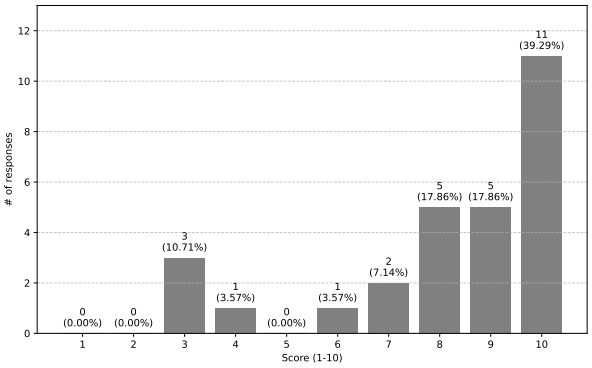

Figure 1: The distribution of scores for "Q1 – Please provide an overall score for the feedback." Scale 1 (bad) – 10 (good), mean=8.14, SD=2.28, n=28 groups and 82 students.

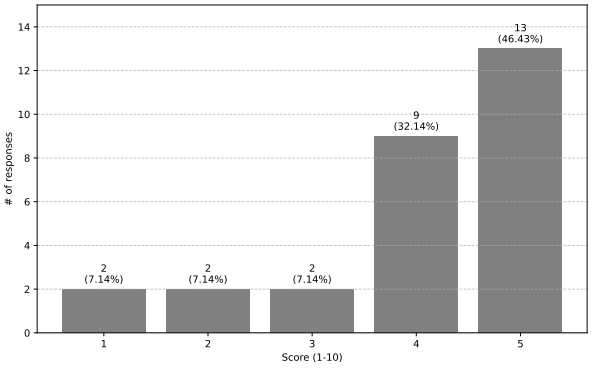

Figure 2: The distribution of scores for "Q2 – How helpful do you think the feedback is?" Scale 1 (not helpful) – 5 (very helpful), mean=4.04, SD=1.21, n=28 groups and 82 students.

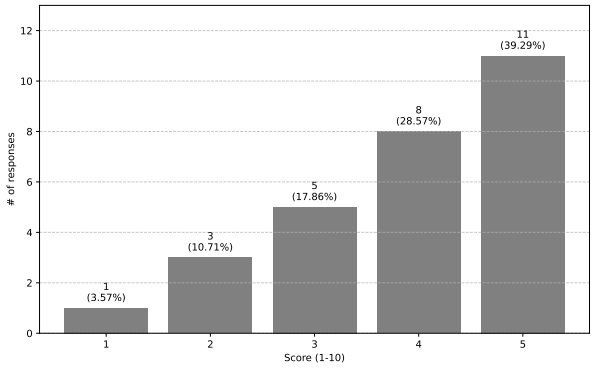

Figure 3: The distribution of scores for "Q3 – How accurate do you think the feedback is?" Scale 1 (not accurate) – 5 (very accurate), mean=3.89, SD=1.14, n=28 groups, 82 students.

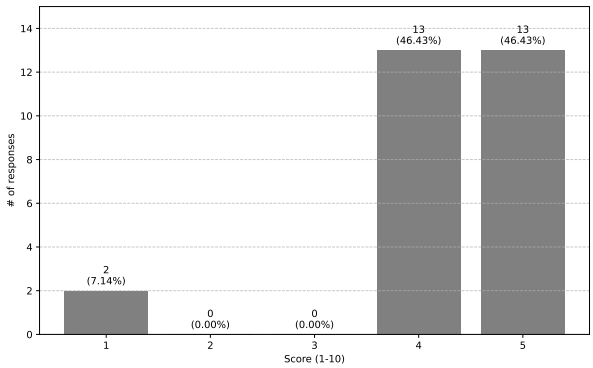

Figure 4: The distribution of scores for "Q4 – How comprehensive do you think the feedback is?" Scale 1 (not comprehensive) – 5 (very comprehensive), mean=4.25, SD=1.02, n=28 groups and 82 students.

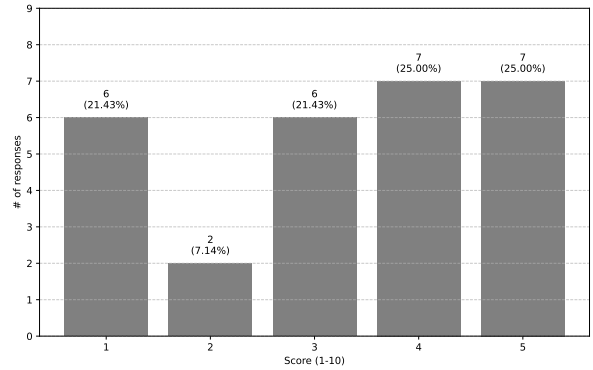

Figure 5: The distribution of scores for "Q5 – Do you think the generated feedback can replace the instructor feedback?" Scale 1 (not likely) – 5 (very likely), mean=3.25, SD=1.45, n=28 groups and 82 students.