LLM Use to Evaluate Student Weekly Reflections in First Year Design Class

Anthony Gwun Hynn Chin, University of California, San Diego

Anthony Chin is a master's student in Mechanical and Aerospace Engineering at the University of California San Diego. Their Research focuses on utilizing education technology and data analysis to improve student learning outcomes in spatial visualization and engineering classes. They have worked on projects involving large scale educational feedback and analysis and are interested in applying AI tools for instructional support and student success.

Dr. Lelli Van Den Einde, eGrove Education

Van Den Einde is a Teaching Professor in Structural Engineering at UC San Diego and the President of eGrove Education, Inc. She has decades of experience teaching hands-on, project-based curricula, spanning high school camps, K-12 outreach, and undergraduate design courses. Dedicated to fostering diversity, she creates supportive environments for students of all backgrounds. Her teaching approach emphasizes scaffolding multidisciplinary skills to boost student self-efficacy and foster meaningful learning outcomes. Dr. Van Den Einde's research focuses on student engagement in large classrooms, developing adaptable design-build-test projects, and creating software to enhance spatial visualization skills, a key factor in improving STEM retention and success.

Dr. Nathan Delson, University of California, San Diego

Nathan Delson, Ph.D. is a Senior Teaching Professor at the University of California at San Diego. He received a PhD in Mechanical Engineering from MIT and his interests include robotics, biomedical devices, product design, engineering education, and maker spaces. In 1999 he co-founded Coactive Drive Corporation (currently General Vibration), a company that provides force feedback solutions. In 2016 Nate co-founded eGrove Education an educational software company focused on teaching sketching and spatial visualization skills.

Jishan Kharbanda, University of California, Los Angeles
Hollis Voinov, University of California, San Diego

Hollis Voinov is an undergraduate student in Structural Engineering at the University of California San Diego. He joined the research team with an interest in analyzing how software can train spatial visualization and how best to support students in large classes.

Andrea Tueanh Huynh, University of California, San Diego

Undergraduate Structural Engineering student at University of California, San Diego.

LLM Use to Evaluate Student Weekly Reflections in a First-Year Engineering Design Class

Abstract

This empirical paper investigates whether Large Language Models (LLMs) are effective in assisting professors in summarizing and scoring students' weekly reflections. Reflections support improved retention and enable instructors to identify those who may need personalized assistance. Driven by the need for scalable, timely, and actionable feedback analysis in large classes, this study examines the reliability of LLM summarizations. The LLM-generated qualitative data is then validated by correlating it with student self-reported quantitative improvements in a first-year engineering course.

Weekly student feedback collected from this course was pre-processed and analyzed using the LLM ChatGPT-4o-mini. The analysis included LLM-based summarization and rubric-based Likert scoring on a 1 to 5 scale. The summarization rating categories included roadblocks, tone/mood, perceived learning, team sentiment, persistence, and engagement quality. These categories were based on Likert responses and open-ended questions from the feedback form, in which students were asked to discuss breakthroughs and roadblocks on their design project, teamwork, and general feedback about the class. Manual analysis of past student reflections served as a benchmark to evaluate the accuracy and consistency of LLM-generated summaries and sentiment scores. The comparison emphasized how effectively both methods identified students who may require additional support or intervention. After the LLM demonstrated performance comparable to manual grading, Spearman's correlation tests were used to examine how LLM-derived rubric scores align with students' end-of-quarter Likert responses and to illustrate the potential for using these validated scores to establish concurrent validity.

Results from this study indicate that LLM-generated scoring based on rubrics is comparable to human scoring. These results suggest that LLMs can serve as effective, scalable graders of short reflection forms in large engineering classes when paired with clear rubrics and prompts. This work highlights how integrating LLM-generated reflection scores establishes a foundation for supporting timely interventions for at-risk students in the future.

Introduction/Background

Weekly reflections are a common, well-established practice in engineering education, particularly in self-regulated learning frameworks [1][2]. Structured reflection encourages students to articulate their thoughts and identify knowledge gaps, essential skills in engineering [3]. In addition to their learning value, reflections generate substantial qualitative data on students' mental and motivational states that quizzes and tests cannot capture on their own. However, in large classes, manually reviewing hundreds of student reflections to understand students' mental and motivational states or issues within a class presents a substantial time barrier to leveraging reflections [4][5]. As a result, many instructors forgo systematic reflection of feedback, which limits the pedagogical impact to the reflective practice itself [6].

Recent advances in large language models (LLMs) offer new avenues for quicker large-scale reflection analysis [8][9]. LLMs, neural networks trained on large amounts of data, can be

prompted with scoring rubrics and guidelines to generate assessments of open-ended questions on feedback forms with reliability comparable to that of human scorers [6][8]. Additionally, LLMs can be used to extract sentiment, thematic patterns, and engagement indicators from reflective text, which can inform instructional decisions and the analysis of learning trends and trajectories within the class [8][9]. Learning analytic systems that combine reflective data with behavioral and engagement metrics provide a more complete picture of the student experience than tests alone [4][10][11]. Early-alert systems based on such learning analytics have shown promising results for flagging at-risk students and enabling timely targeted interventions [11][12][13]. However, most of these learning analytics rely on quantifiable data [4][12]. Therefore, extending these systems to incorporate LLM-analyzed reflection data could enhance the timeliness and accuracy of the analysis of at-risk students and improve overall pedagogical outcomes [6][8][12].

This study aims to investigate whether LLMs can analyze student reflections from a large first-year engineering class, validate LLM analyses against human scores, and lay the foundation for identifying meaningful correlations and indicators of at-risk students. This work bridges learning analytics and LLM-based analysis of feedback surveys to assess the practical value of automated reflection analysis in supporting student success and instructional effectiveness.

Methods

Approach

This study builds on data from a first-year engineering graphics course, where students learned CAD, spatial visualization skills, and worked on a hands-on team design project. Data from weekly reflections were analyzed using ChatGPT 4o mini. To perform this analysis, a rubric and instructions were provided to the LLM, ensuring clear grading criteria and a structured output. After obtaining initial scores from the LLM, the results were compared with those of three human researchers to demonstrate reliability and validity. Lastly, once human researchers and LLM scores were in relative alignment, a Spearman correlation analysis was conducted to examine how LLM reflection scoring directly relates to students' Likert responses. Spearman's correlation was chosen because the Likert-scale data were ordinal and not normally distributed. Through this process, the goal was to develop a tool to effectively grade student reflection responses for further correlation analysis or to identify at-risk students.

This study was classified as human subjects research by the UCSD Institutional Review Board (#812291). All study participants gave informed consent before their data were utilized in this analysis.

Context and Participants

In the course evaluated, a total of 286 students were enrolled, though not all students completed every weekly feedback form. Of these, 276 consented to have their data used for research, and each completed at least one feedback form that was included in the sample. During each weekly homework deadline, students were tasked to complete a form after finishing all other homework tasks to reflect on their overall learning and sentiment in the class. To avoid overwhelming students, they were required to complete only 8 of the 10 weekly reflections throughout the

10-week quarter; failure to do so would result in a loss of up to 5% of their total grade, depending on the number of reflections the student missed. Each reflection form consisted of two components: Likert questions and open-ended free-response questions. The Likert questions used a 5-point scale and required students to assess multiple aspects of their learning experience, including the following: Perceived understanding of course content (CAD skills, Spatial Visualization ability, and making tools), confidence in applying design skills (engineering theory application), perceptions of course climate (feeling included and supported), workload and time spent on coursework, and team dynamics. Additionally, students were also recommended to answer at least one of these short answer prompts: describe any creative breakthroughs, describe any creative blocks, or describe any utilization of math or physics. The short answer portion of the feedback forms was not required. All of this data was then provided to the LLM for analysis and scoring.

A rubric was created for LLM and human researchers to use to score students. The grading rubric helped determine whether the LLM could generate reliable, human-like evaluations of student reflections, which could then be compared.

Analysis Methods

To generate LLM scores for students, a Python script was written to organize student data and prompt the ChatGPT-4o-mini model to score student reflections using the rubric (see appendix). The script read in student reflections from an Excel spreadsheet containing all nine weekly reflection forms and one end-of-quarter reflection. Additionally, the rubric and a prompt outlining requirements and specifications for grading the students were provided. The rubric consisted of five broad categories encompassing 12 subcategories to help identify the traits of successful students and those in need of assistance. The five categories that measured learning consisted of struggle, motivation, teamwork, engagement quality, and the need for instructor support. These categories were selected because prior research in education has shown that students' motivation and responses to struggle, the quality of their teamwork experiences, and their perceptions of instructor support are closely linked to engagement and academic performance [13][14][15][16][17]. Listed below are the rubric categories and the rationale for each selection.

Learning and Struggle

- Perceived learning: Self-reported learning is a core component of self-regulated learning and is an important metric in monitoring student progress toward learning goals.
- Encountered roadblocks: Struggle is an important aspect of learning, and measuring how often students encounter obstacles helps us understand if roadblocks are barriers or learning opportunities
- Overcoming roadblocks: Persistence in problem-solving is central to learning. As such, students who can persist and overcome obstacles will be more successful in learning

Affect and Motivation

- Mood tone: Students' emotional states influence learning outcomes. Negative effects might signal that the student has decreased motivation and is less likely to put in effort in the classroom.

- Overall enthusiasm: A course's overall enthusiasm is a good predictor of engagement, and low enthusiasm may signal disengagement and a need for assistance.
- Persistence: Defines a student's ability to maintain effort despite difficulty. We hypothesize that students who can persist will have higher academic success and retention. Thus, we want to analyze and score this trait.
- General trends: Used to track overall trajectory and see if the student improved or declined, allowing the ability to distinguish between struggling and well-performing students

Team Experience

- Team sentiment: Positive team experiences are associated with higher satisfaction with the course and greater learning retention.
- Team roadblocks: Negative team experience could impact an individual's performance and well-being, and would be a likely identifier for students in need of assistance

Engagement and Quality

- Response rate: Regular responses mean the student is engaged with the class and will likely perform better.
- Reflection quality: Detailed and thoughtful responses likely indicate self-awareness in the course and a deeper engagement with the feedback form.

Support

- Intervention need: Identifies students who would benefit most from intervention. Students in this category are likely disengaged, struggling, or have unresolved roadblocks and could use the professor's assistance.

Each rubric category was scored on a 1 to 5 scale, with the general definitions being as follows: 5 = Strong/Consistently Positive, 4 = Above average, 3 = Average, 2 = Below average, and 1 = Poor/strong concern. After the rubric was drafted, it was iteratively refined through small pilot runs on a subset of student data and by comparing it with manual grading to ensure that each category accurately captured the intended aspect of students' experiences.

In addition to the rubric, a short prompt was sent to the LLM to improve its grading accuracy [7]. Prior research has shown that context, role prompting, chain-of-thought prompting, examples, and explicit constraints help the LLM provide high-quality feedback and consistent responses. Guided by this, the prompt asked the LLM to serve as an educational evaluator for the engineering course and to analyze each reflection form methodically. The prompt included context and graded examples, along with clear, explicit instructions for grading each reflection. After initial testing, additional context and restrictions were given to the final prompt. These included reflection context, emphasizing that students only needed to complete 8 of the 10 weekly reflections (to prevent the model from penalizing missing weeks), and providing clear instructions for the model to use Likert evidence as the primary evidence when short-answer responses were unavailable (to prevent "insufficient evidence" responses). Figure 1 shows the final prompt provided to the LLM for scoring the student reflection forms.

Context
You are a careful educational evaluator for an undergraduate Computer Graphics Course in Engineering. Your task is to evaluate each student based on their feedback. Use ONLY evidence in the provided student data. You will be provided with the reflection context, scoring rubric, graded examples, and instructions, which are detailed below.

Reflection Context: """ Students are given reflections once a week and are only required to complete 8/10 weekly reflections to receive the full 5% of their grade. Additionally, students are advised to complete at least one short answer response in addition to the required Likert responses."""

Scoring Rubric """ {Rubric} """

Graded Examples: """ {examples} """

Instructions: Let's think step by step:

1. Analyze the reflections with the provided rubric and graded examples
2. Based on the rubric, assign a score between 1 - 5 to each of the rubric categories. IMPORTANT: Likert-scale responses are numeric fields in the data; use them to score even if short answers are missing. Only return scoring_status='insufficient_evidence' if the student's data is essentially empty.
3. Return data in a single Json object

Figure 1. Final Prompt Provided to LLM

Manual Verification and Rubric Refinement
To verify and validate the LLM, the rubric and prompt were provided to the three researchers to independently score the reflections. The researchers consisted of two undergraduate and one graduate student, under the supervision of two professors specializing in engineering pedagogy. The sample provided to the researchers consisted of 20 randomly selected students, that were superficially reviewed before scoring to ensure the student responses varied in length. Each researcher then independently scored the sample reflections using the same 1-5 scale and written rubrics that were provided to the LLM model, and scores between human graders and the LLM were compared to identify differences.

A common difference was that the distribution of LLM scores often did not utilize the full rubric 1 through 5 range, typically staying in the middle 2 through 4 values, whereas human scorers more frequently utilized the full range. Histograms comparing Human and LLM scores highlighted a concentration of scores in the LLM and revealed that LLMs were particularly clustered in categories where the rubric was unclear about distinguishing between grading levels. These rubric categories were then revised to improve clarity. For example, in the reflection-quality category, the original rubric scores "Substantive and insightful" (4) and "Highly substantive and insightful" (5) lacked sufficient specificity, so the model frequently defaulted to 4. After initial testing, these categories were revised so that level 4 represented

“Solid and complete; answers the question directly with some explanation or example, but depth, specificity, or connections are limited” and level 5 to be “Highly substantive and insightful; explains reasoning clearly, uses specific examples or details, makes connections (e.g., to prior weeks, concepts, or strategies), and shows clear metacognition about learning”. After these revisions, a follow-up test of the LLM showed its scores aligned much more closely with human ratings, indicating that clarifying the boundaries between adjacent score levels helped the model apply the full-scale rating more consistently.

Results and Discussion

After calibrating the rubric, a baseline analysis to compare the consistency by which the three human graders scored the reflection data. The results for human scores on the sample set of 20 students showed that researchers assigned the same score to a student 26% of the time, were within 1 point 66% of the time, and were within 2 points 92.1% of the time (see Table 1). These values indicate non-trivial variability among human scorers applying a shared rubric to the reflection data, highlighting meaningful spread in human judgments.

Table 1: Validation of Human Researchers

Percent of Time All Human Researchers Agree	Percent of Time All Human Researchers Score Within 1 of Each Other	Percent of Time All Human Researchers Score Within 2 of Each Other
25.9%	65.5%	92.1%

Although the human scorers' agreement was modest, the high percentages within one and within two suggest a reasonable level of consistency despite differences in individual interpretations of the reflection data. This variability likely reflects the subjective nature of several rubric categories, in which differences between adjacent scores (e.g., 4 versus 5 or 1 versus 2) require nuanced judgment, particularly when students provide minimal written feedback.

Subsequent analysis compared LLM scores with human ratings for the same set of 20 students (see Table 2). The results compared the relative agreement among all the researchers and the LLM, and how often the LLM scored within 1 or 2 points of the average human score. When the LLM was treated as a fourth scorer, all human researchers and the LLM were within one point of each other 56.8% of the time, and were within two points 87.8% of the time. Additionally, compared with the *average* human score, the LLM scores were within one point of it 82.7% of the time and within two points 99.2%. The pattern suggests that the LLM typically scores within the band of disagreement among human scorers and does not deviate from it, demonstrating that the LLM can effectively replicate a human scorer’s judgment when analyzing a reflection form. This follows similar studies on LLM grading of short-answer questions, which suggest that LLMs can achieve substantial agreement with humans when graded with solid rubrics and prompts, while still requiring human oversight, particularly for borderline or low-quality responses [7][8].

Table 2: Validation of LLM results

Percent of Time All Human Researchers and LLM Scores Within 1 of Each Other	Percent of Time All Human Researchers and LLM Scores Within 2 of Each Other	Percent of Time LLM Scores Within 1 of the Average of the Human Researchers	Percent of Time LLM Scores Within 2 of the Average of the Human Researchers
56.8%	87.8%	82.7%	99.2%

Additionally, to assess whether the LLM scores aligned with students' course experience, a Spearman correlation was computed between the LLM rubric scores and the End-of-Quarter reflection responses for all 276 students in the class to establish concurrent validity between the LLM scores and student self-reports. A full correlation matrix was generated by combining all rubric categories with the Likert-scale items from the survey (31 variables in total), yielding a 31×31 matrix of pairwise correlations. Due to the large sample size (n=276 students), many correlations were statistically significant. Therefore, the items shown in Table 3 were selected to highlight the strongest and clearest correlations. The strongest correlation was in the area of Team Sentiment, which showed a strong positive correlation between students' end-of-quarter ratings of teamwork effectiveness with an $r = 0.67$. Reflection Quality correlated moderately with feeling supported in the class ($r = 0.42$), and Perceived learning correlated with self-rated CAD ability ($r = 0.39$).

Table 3: Statistically Significant Correlations Spearman Correlation Table For All Students comparing Likert Scoring by ChatGPT and End of Quarter Likert Responses

Likert Scored by the LLM	End of Quarter Likert Responses	p-values	r-values
Team Sentiment	How effective was the teamwork in your group for the past week?	< 0.001	0.67
Reflection Quality	Did you feel well supported in the class?	< 0.001	0.42
Perceived Learning	Rate your ability with Computer-Aided Design (CAD).	< 0.001	0.39

The strong positive correlations across these categories were in line with expectations and supported the idea that the LLM effectively scored students. Students with high team effectiveness should have positive team sentiment. Similarly, students who feel well supported in the class should feel more confident and comfortable providing more reflective feedback, resulting in higher-quality reflections. And students who rate their CAD ability high should have a high perceived learning, as it is a large focus of this class. Histogram data further supported the strong correlations between these groups, as shown in Figures 2 and 3. These figures demonstrate how closely the overall LLM scoring aligns with the Likert scale responses on the End of Quarter report, closely mirroring the student self-reported trends. Notably, the LLM data sample is larger, as not all students who submitted weekly reflections completed the end-of-quarter reflection.

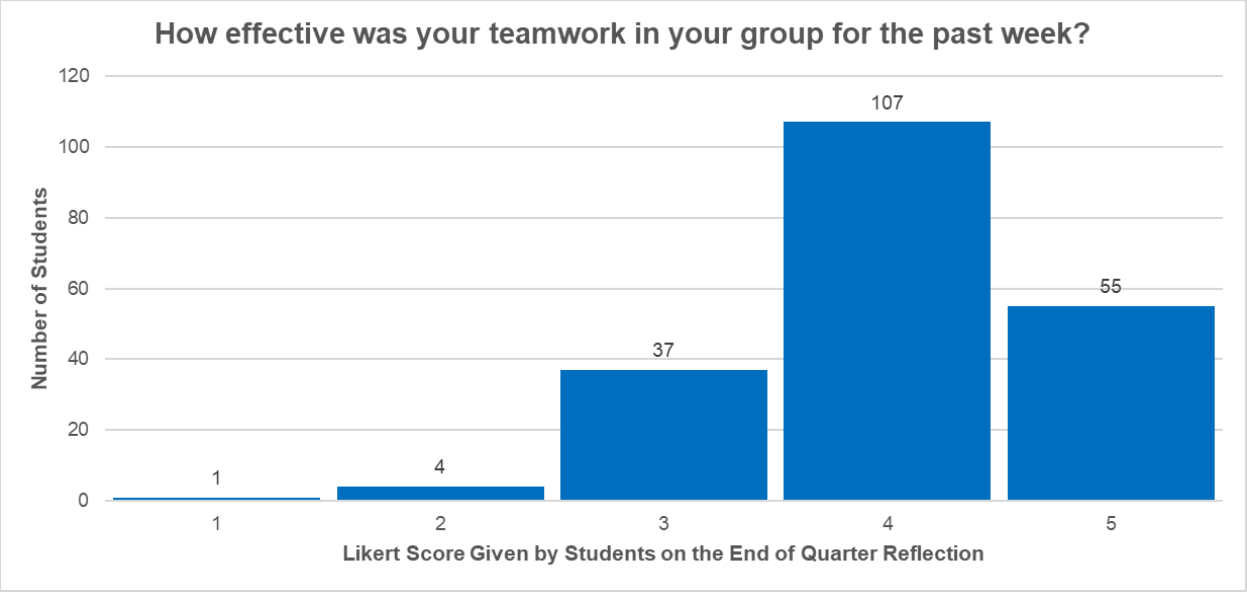

Figure 2. Distribution of student-reported teamwork effectiveness scores on the End-of-Quarter Reflection (1 = not effective, 5 = highly effective).

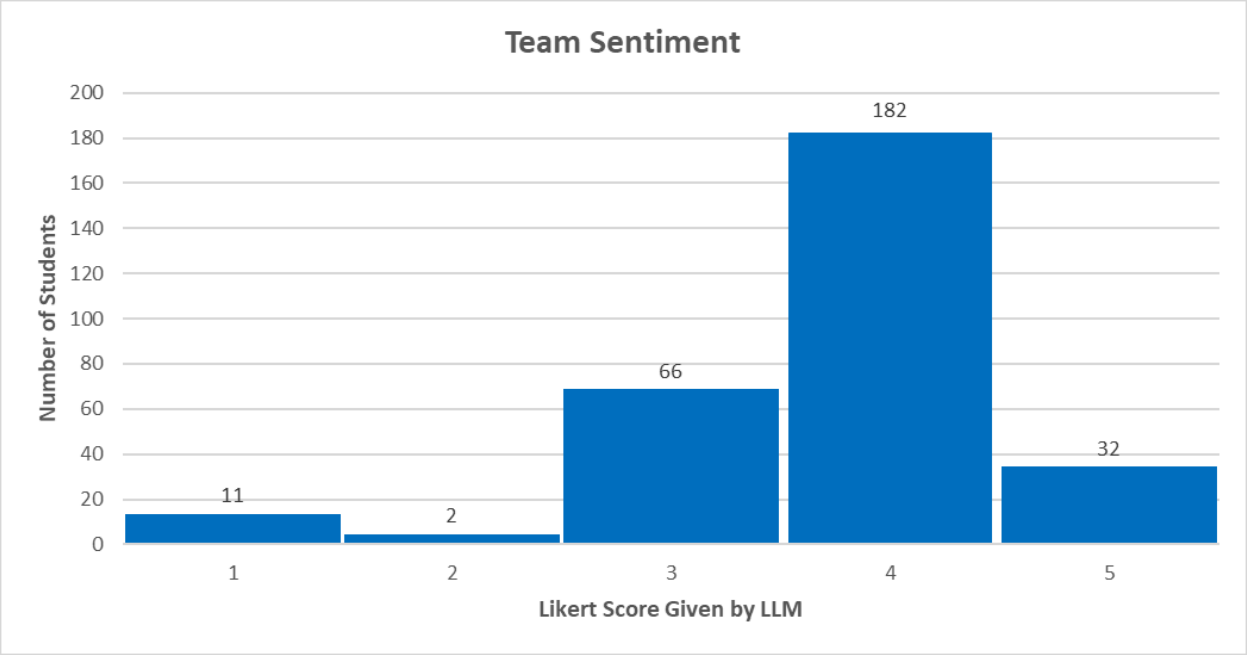

Figure 3. Distribution of LLM-assigned Team Sentiment scores across all students (1 = consistently negative, 5 = consistently positive).

Overall, the evidence indicates that LLM-rubric-based scoring of student reflections can provide human-comparable scoring in large engineering classes. This is consistent with the literature, which shows that when LLMs are given clear, explicit rubrics and carefully designed prompts, they can effectively grade students [7]. The effectiveness of the LLM scoring and strong correlations to the end-of-quarter reflection provide a strong basis for future work in which correlations of the LLM can be compared to class grades on tests and assignments to determine

characteristics of student success and help professors in large classes identify students in need of improvement.

Conclusion

The results of this study demonstrate that LLMs offer a reliable, scalable solution for analyzing student reflections in large engineering courses, where identifying students who require additional support can be difficult for instructors. Our validation phase indicated that the LLM performed similarly to human grading in terms of agreement on scores, with model ratings falling within one point of the average human score 82.7% of the time across rubric categories. This suggests that when LLMs are given prompts with clear, specific rubrics, they can provide rubric-based numeric assessments that fall within the spread observed among our human researchers and score near the average of human scorers nearly all of the time.

This work yielded strong initial results; however, it had some limitations. These limitations included inconsistent student responses, as students were not required to answer any of the short-answer questions. This may have led the LLM to rely too heavily on students' Likert scores rather than using both Likert and short-answer questions as intended. The lack of short-answer responses also made it difficult at times for human scorers to evaluate students effectively, as there was little feedback to guide their evaluations. Additionally, human verification was limited to just three researchers with limited training; future work should include additional graded examples from more experienced graders. Lastly, because the reflection was lengthy, students may have answered questions quickly rather than respond to the best of their abilities.

Future work for this research should include a correlation analysis of actual student grades and LLM scores to identify trends that predict student success. Additional work should also implement week-by-week reflection analysis to find trends in an engineering course. The weekly scoring would allow instructors to quickly identify at-risk students during the course using AI to detect downward trends in sentiment and engagement. This could allow instructors to transition from delayed feedback to proactive intervention, a crucial practice in the classroom. Longer-term studies could also examine how instructors integrate these LLM scorers into their existing practices and whether such implementation provides measurable student improvement.

Disclosure

Nathan Delson and Lelli Van Den Einde have an equity interest in eGrove Education, Inc, a company that may potentially benefit from the research results. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

References

[1] Y. Wang and J. J. Harris, “Self-Regulated Learning Instructions in Engineering Education: A Systematic Scoping Review,” *2022 IEEE Frontiers in Education Conference (FIE)*, vol. 109, no. 2, Oct. 2022, doi: https://doi.org/10.1109/fie56618.2022.9962475.

[2] A. Hadwin, S. Järvelä, and M. Miller, “Self-Regulation, Co-Regulation, and Shared Regulation in Collaborative Learning Environments,” *Handbook of Self-Regulation of Learning and Performance*, pp. 83–106, Sep. 2017, doi: https://doi.org/10.4324/9781315697048-6.

[3] M. E. Kihal, C. Wallwey, J. David, and J. N. Newcomer, “Self-Regulated Learning in First Year Engineering: Opportunities for Practical Implementation,” *Asee.org*, Jul. 28, 2024.http://peer.asee.org/self-regulated-learning-in-first-year-engineering-opportunities-for-practical-implementation (accessed Jan. 22, 2026).

[4] H. A. Diefes‐Dux and L. M. Cruz Castro, “Reflection Types and students’ Viewing of Feedback in a First‐year Engineering Course Using Standards‐based Grading,” *Journal of Engineering Education*, vol. 111, no. 2, Jan. 2022, doi: https://doi.org/10.1002/jee.20452

[5] Muhammad Afzaal and J. Nouri, “A Systematic Review of Software for Learning Analytics in Higher Education,” *International Journal of Emerging Technologies in Learning (iJET)*, vol. 19, no. 07, pp. 17–43, Sep. 2024, doi: https://doi.org/10.3991/ijet.v19i07.50313.

[6] C. Grévisse, “LLM-based automatic short answer grading in undergraduate medical education,” *BMC Medical Education*, vol. 24, no. 1, Sep. 2024, doi: https://doi.org/10.1186/s12909-024-06026-5.

[7] E. M. AlGhamdi, Y. Li, Dragan Gašević, and G. Chen, “Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education,” *Computers & Education*, vol. 243, pp. 105511–105511, Nov. 2025, doi: https://doi.org/10.1016/j.compedu.2025.105511.

[8] O. Bolgova, P. Ganguly, M. F. Ikram, and V. Mavrych, “Evaluating large language models as graders of medical short answer questions: a comparative analysis with expert human graders,” *Medical Education Online*, vol. 30, no. 1, Aug. 2025, doi: https://doi.org/10.1080/10872981.2025.2550751.

[9] I. Daqiqil ID, H. Saputra, S. Syamsudhuha, R. Kurniawan, and Y. Andriyani, “Sentiment analysis of student evaluation feedback using transformer-based language models,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 2, p. 1127, Nov. 2024, doi: https://doi.org/10.11591/ijeecs.v36.i2.pp1127-1139.

[10] E. Foster and R. Siddle, “The effectiveness of learning analytics for identifying at-risk students in higher education,” *Assessment & Evaluation in Higher Education*, pp. 1–13, Nov. 2019, doi: https://doi.org/10.1080/02602938.2019.1682118.

[11] Ryan S.J.d. Baker and Kalina Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” *JEDM | Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009, doi: https://doi.org/10.5281/zenodo.3554657.

[12] D. R. Tampke, “Developing, implementing, and assessing an early alert system,” Journal of College Student Retention: Research, Theory & Practice, vol. 14, no. 4, pp. 523–532, 2013, doi: https://doi.org/10.2190/CS.14.4.e.

[13] S. Anwar, M. Menekse, and A. Kardgar, “Engineering Students’ Self-Reflections, Teamwork Behaviors, and Academic Performance,” Jun. 15, 2019. https://www.google.com/url?q=https://peer.asee.org/engineering-students-self-reflections-teamwork-behaviors-and-academic-performance&sa=D&source=docs&ust=1771425600033273&usg=AOvVaw0WtMPIIytIBsVyzeE6FPL4 (accessed Feb. 19, 2026).

[14] M. Sousa and E. Fontão, “Team-based learning—An approach to enhance collaboration and academic success in engineering education: A comprehensive study,” Forum for Education Studies, vol. 3, no. 1, p. 2239, Mar. 2025, doi: https://doi.org/10.59400/fes2239.

[15] S. Outerbridge, M. Taub, M. Nader, S. Pal, R. Zaurin, and H. Jin Cho, “Investigating Motivation and Self-Regulated Learning for Students in a Fundamental Engineering Course,” 2024. https://peer.asee.org/investigating-motivation-and-self-regulated-learning-for-students-in-a-fundamental-engineering-course.pdf (accessed Jan 26, 2026).

[16] Y. Zhang et al., “The impact of perceived teacher support on students’ learning approach: the chain mediating role of academic engagement and achievement goal orientation,” Frontiers in Psychology, vol. 16, Jun. 2025, doi: https://doi.org/10.3389/fpsyg.2025.1513538.

[17] P. H. Carnell, N. J. Hunsu, D. F. Ray, and N. W. Sochacka, “Exploring the Relationships Between Resilience and Student Performance in an Engineering Statics Class: A Work in Progress,” 2018. https://peer.asee.org/exploring-the-relationships-between-resilience-and-student-performance-in-an-engineering-statics-class-a-work-in-progress.pdf

Appendix

Rubric

Rate the student on these 11 dimensions using integers 1–5, following EXACTLY the definitions below:

CONTEXT:

- Students were only required to complete 80% of feedback forms to receive 100% credit
- Students were only recommended to answer ONE open-ended question per feedback form
- Use all available data

Perceived Learning:

5: The student shows mastery beyond expectations, deep understanding demonstrated
4: The student shows strong, clear, substantial learning and skill development with specific examples
3: The student shows adequate meets learning objectives, shows moderate progress
2: The student is below expectation and struggles with concepts, shows minimal progress
1: The student provides insufficient information in feedback forms with no evidence of learning or engagement

Encountered Roadblocks frequency and severity of obstacles

1: The student encounters frequent major obstacles (serious, repeated challenges).
2: The student encounters some major obstacles and occasional minor obstacles.
3: The student frequently faces minor obstacles (many small challenges).
4: The student faces occasional minor obstacles (few challenges overall).
5: The student reports very few obstacles; progress is smooth throughout.

Overcoming Roadblocks ability to resolve challenges

5: The student consistently develops effective strategies and independently overcomes obstacles with confidence.
4: The student usually overcomes challenges on their own, showing resilience and resourcefulness. Sometimes overcomes obstacles with help from others.
3: The student overcomes most obstacles, typically with help from peers, instructor, or other resources.
2: The student rarely overcomes challenges and often needs intervention or support.
1: The student struggles to overcome obstacles and remains stuck or frustrated despite attempts or support.

Mood Tone (overall satisfaction and affect)

5: The student is highly satisfied and positive about the class, instructor, and peers.
4: The student is generally satisfied, with only minor distractions or negatives.
3: The student regularly expresses frustration or dissatisfaction.
2: The student is often or consistently dissatisfied and frustrated with the course structure/content.
1: The student reports being overwhelmed and extremely dissatisfied with the course.

Overall Enthusiasm (motivation and engagement)

5: The student demonstrates excitement, motivation, and energy in participation and learning throughout the course.
4: The student is generally motivated and participates with interest, though with occasional lapses.
3: The student's enthusiasm is mixed, with periods of engagement and periods of disengagement.
2: The student is generally disengaged, passive, or uninterested in the course.
1: The student demonstrates little or no motivation, enthusiasm, or engagement.

Team Sentiment

5: The student consistently enjoys collaborating with their team and reports a positive, productive experience throughout the course.
4: The student generally enjoys working with their team and maintains a positive attitude despite minor challenges.
3: The student works effectively with their team and occasionally experiences frustration, but maintains a neutral to positive outlook overall.
2: The student experiences ongoing tension or interpersonal challenges with teammates that impact their overall sentiment but make efforts to stay engaged.
1: The student consistently struggles to enjoy teamwork and expresses clear dissatisfaction or disengagement with the team dynamic.

Team Roadblocks

5: The student encounters few or no team-related obstacles and communicates effectively to maintain strong collaboration.
4: The student faces occasional team issues but resolves them collaboratively and promptly.
3: The student experiences some recurring team challenges that require effort to overcome but ultimately manages to do so.
2: The student encounters significant difficulties working with team members, though they take steps to address and resolve them.
1: The student faces serious and persistent conflicts with their team that remain unresolved and impede progress.

Persistence

5: The student puts in lots of time and effort despite obstacles (Especially deserved if they put in many hours of effort despite difficult/frequent obstacles)
4: The student puts in lots of time and effort despite minor obstacles
3: The student puts in a reasonable amount of time and may encounter some obstacles that they overcome
2: The student does not spend a lot of time on coursework and overcomes some obstacles
1: The student does not spend a lot of time on coursework and does not overcome or overcomes very few and minor obstacles

General Trends

5: The student demonstrates growth and improving steadily over time.
4: The student maintains solid consistent performance throughout.

3: The student maintains relatively consistent performance with moderate or limited improvement.
2: The student shows little change in ability or performance throughout most of the course.
1: The student's performance declines over time.

Need for Intervention
5: The student is significantly struggling or in clear distress. Shows frequent frustration, low or declining learning, serious team problems, or disengagement. Immediate, proactive contact and support are necessary.
4: The student is struggling in one or more major areas (course concepts, teamwork, motivation, or workload). Reflections show repeated problems or negative sentiment. A targeted intervention (office hours, email, or meeting) is recommended soon.
3: The student is doing "okay" overall but has recurring or moderate issues (confusion about material, inconsistent effort, uneven teamwork, or stress). A check-in would likely help, but issues are not urgent.
2: The student is generally performing well with only minor or occasional issues. Reflections show mostly positive or neutral experiences with some small challenges. No formal intervention is needed; light encouragement or optional office hours could help.
1: The student is doing very well. Strong or improving performance, positive reflections, and no meaningful indicators of struggle. No intervention is needed or recommended.

Response Rate - Completion percentage of open-ended questions
5: 80-100% of short answer questions answered across most/all weeks (comprehensive engagement)
4: 60-79% of short answer questions answered consistently (regular engagement)
3: 40-59% of short answer questions answered (selective engagement)
2: 20-39% of short answer questions answered (minimal engagement)
1: Under 20% of short answer questions answered (low engagement)

Reflection Quality - Depth and thoughtfulness of answers when provided
5: Highly substantive and insightful; explains reasoning clearly, uses specific examples or details, makes connections (e.g., to prior weeks, concepts, or strategies), and shows clear metacognition about learning.
4: Solid and complete; answers the question directly with some explanation or example, but depth, specificity, or connections are limited compared to level 5.
3: Barely sufficient; addresses the prompt in a literal way but with minimal detail, generic statements, or little explanation of "how" or "why."
2: Weak or superficial; very short, vague, off-topic in parts, or shows minimal effort, with little evidence of actual thinking or learning.
1: Essentially no meaningful content; blank, one- or two-word replies, copied text, or responses that are incoherent or unrelated to the question.

General Sentiment Regarding Emails - ONLY rate if student explicitly mentions receiving/reading emails from Professor
5: Student felt greatly encouraged and motivated by the emails; felt they helped with learning or visualization skills

4: Student felt moderately encouraged or motivated by receiving the emails
3: Student mentioned receiving emails but felt they had no impact or mixed/neutral impact
2: Student felt the emails were somewhat discouraging or not helpful
1: Student felt the emails were discouraging or negative in tone
NA: Student did not receive emails, or did not mention emails in their reflections (no evidence of receiving them)

JSON instructions
Return ONLY a single JSON object with these fields (ALL fields are REQUIRED):
{
"perceived_learning": 1-5 (integer),
"encountered_roadblocks": 1-5 (integer),
"overcoming_roadblocks": 1-5 (integer),
"mood_tone": 1-5 (integer),
"overall_enthusiasm": 1-5 (integer),
"team_sentiment": 1-5 (integer),
"team_roadblocks": 1-5 (integer),
"persistence": 1-5 (integer),
"general_trends": 1-5 (integer),
"response_rate": 1-5 (integer),
"reflection_quality": 1-5 (integer),
"intervention_need": 1-5 (integer),
"email_sentiment": 1-5 (integer),

"summary": "A comprehensive paragraph (6-10 sentences) synthesizing the student's overall performance, trajectory, strengths, and challenges throughout the quarter. Include specific evidence from their reflections, explain your reasoning for key ratings, and provide context for intervention recommendations. Reference specific weeks or patterns when relevant.",
"improvement_level": "Moderate Decline" OR "Slight Decline" OR "No Improvement" OR "Slight Improvement" OR "Great Improvement",
"intervention_type": "No intervention needed" OR "Office Hours intervention" OR "Student Team Check in" OR "Student email check in" OR "Private Meeting with Professor",
"evidence": ["<=3 short quotes or paraphrases with week tags that justify the ratings"]
}

No markdown. No extra text. Ensure ALL ratings are present as integers 1-5.