Adoption of Large Language Model for Improving Grading Timing in an Engineering Economics Course

Dr. Billy Gray, Tarleton State University

Billy Gray is an Associate Professor at Tarleton State University in the Department of Engineering Technology. He holds a PhD in Industrial Engineering from the University of Texas at Arlington, a Master's degree from Texas Tech University in Systems and Engineering Management, and a BS in Manufacturing Engineering Technology from Tarleton State University.

Dr. Gloria Margarita Fragoso-Diaz,

Dr. Fragoso-Diaz is Assistant Dean for Outreach and Recruitment and Associate Professor of Engineering Technology in the Mayfield College of Engineering at Tarleton State University. She received her Ph.D. in Industrial Engineering and Master's degree in Industrial Engineering from New Mexico State University. Dr. Fragoso's research interest is in supply chains and student success.

Adoption of Large Language Model for Improved Grading in an Engineering Economics Course

Abstract

This research expands on the use of Large Language Models (LLM), to evaluate and provide feedback on student papers. This paper uses a custom GPT created through OpenAI's ChatGPT. Previous work by the authors found that there is significant time savings from using an LLM to assist in providing feedback to students but there was not a strong agreement between the evaluator ratings and the LLM ratings. The authors hypothesize that the LLM would be in better agreement with the evaluators and as an aid for grading the assignments using rubrics developed for the course. The research methodology for this study begins with having two evaluators (instructors) grade student assignments independently, then comparing their evaluations in a norming exercise. Next, predefined assignment responses were created with the help of ChatGPT to build background information for the custom GPT. These responses are based on the existing rubric and help increase the reliability of the ratings against the rubric. The last step compares the LLM evaluations of student assignments to those provided by the evaluators. This is used to determine the degree of agreement between the faculty evaluators and the LLM using Krippendorff's-α and Cohen's Kappa test methods. The authors expected the results to show a strong level of agreement, with Cohen's Kappa and Krippendorff's-α greater than 0.6. Instead, both values rated at 0.128 (poor) and 0.17 (slight) in agreement respectively. Though the expected improvement in reliability was not achieved, the LLM's usability for feedback purposes did prove to be advantageous to the instructor.

Introduction

One task many professors have in their courses is providing accurate and timely feedback to students on written assignments. As courses continue to increase in size and professors have an increasing need to perform their research, methods that help them be as efficient with their time as possible. The feedback should be accurate. Though many professors may have graduate students to help them grade and reduce their workload, not all universities have those options. Instead, professors need more access to newer technologies to help them. As with many technologies, the point of the technology is to help make the users more efficient in their job tasks. AI, specifically LLM's, has been shown to be useful in this regard.

In the previous study by the authors, the LLM proved to save a considerable amount of time in providing feedback, thereby reducing the workload the faculty experienced grading essays and papers. Unfortunately, the agreement in ratings provided by the evaluators and the LLM measured as "poor" and "slight" based on the background information and the specific LLM used. The expectation was that the ratings would measure higher and be more in agreement with each other. This study continues the previous work and focuses more on increasing the agreement between the human evaluators and the LLM ratings. By increasing the agreement

between the ratings, there is a perceived notion that output from the LLM would be more in line with the human evaluator's judgement yielding a more valid evaluation of the student papers by the LLM.

The hypothesis of this work is that the results will show a strong level of agreement between the ratings of the evaluators and the LLM. Cohen's Kappa and Krippendorff's-α will be used to measure the inter-rater agreements, with a target of both measures being greater than 0.6. If the measures are greater than 0.6, the hypothesis will fail to be rejected. Results less than 0.6 will result in the hypothesis being rejected.

Background

The term Artificial Intelligence (AI) was first introduced in the year 1956. By 1960 the US Department of Defense was adopting the concept [1]. In education, [2] report that the application of AI for feedback purposes dates back to the 1950's in the area of computerized assessment. But it wasn't until the first generation of computers where applications of AI started making their way into the education sector. In 1990, [3] reported a study on the use of microcomputers as intelligent tutors to determine if those tutors were able to improve education. In one-way, intelligent tutors would help students by presenting additional instruction on those areas students were having difficulties with. In their study they also recognized that the use of intelligent tutors resulted in changes of grading practices. Back then, the authors experimented on an improved grading scheme by changing the way they graded. Although different from this study's purpose, it was observed a positive impact in grading [3]. [4] first described and reported on the AI tutor called the Geometry Tutor as part of "development of intelligent-based tutors for mathematics and science subjects in the range of senior high-school to junior college."

Nonetheless, the applications of AI in education were mostly administrative such as record keeping. Initially, the goal was to provide instructors with more time to work on research, studying, preparing materials, etc. For students, the purpose was educational resources. Fast forward there have been important advances in the use of AI in applications such learning platforms like Blackboard [1].

Interest in AI in education has increased in recent years. This paragraph includes reports on literature reviews. Interestingly, each may have different purposes or focus, but they all concur with the increasing number of publications leading up to the year 2019. There seems to be more focus in application and effects of AI in administration (i.e. grading, feedback for example), instruction (i.e. virtual reality, adaptive web-based platforms) and learning (i.e. personalized content) [5]. These authors mentioned fewer publications beginning 2010 with a notable increase towards 2019. [6] indicated that there were fewer publications between 2007 and 2018, and during 2019 a significant rise in publications is noticed and even more after the launch of ChatGPT in 2022. Their research explains how AI in education changed over time and to understand the trends while studying AI tools for feedback purposes [6]. They conclude that there are a fast growth and global interest in AI feedback tools in education, however, complex

and open-ended tasks represented challenges. In his paper he stated, “While an AI tool can effectively grade multiple-choice questions or provide grammar corrections, evaluating the creativity and originality of a story or the logical coherence of a complex argument requires more sophisticated analysis that current AI technologies are still developing.” [6]. [1] also presents a systematic review of AI in education with intent of studying and analyzing opportunities, benefits, and challenges in education. On the challenges, the author mentions how developing software to help grading with written essays is still in the works.

In the education sector, effective feedback is a key process that positively impacts the learning process. We can agree that effective feedback involves quality, on time feedback and learner friendly [7]. [2] work shows agreement with [7] but also includes discussion on how timely feedback and frequency are factors to consider. However, it typically requires faculty to invest a significant amount of time and effort.

Researchers commonly note that AI has given education the ability to provide personalized learning. Having ChatGPT functionality embedded in learning machines is a good start for the process of achieving personalized learning [8]. Consideration of its implementation includes the use of chatbots as educational assistants for timely response to students’ questions. While proposing this, a recommendation should be “without making compromises and risking data security” [8].

On the use of Large Language Models LLM, [9] introduced an LLM called PEER. They defined it as “it is a user-friendly web-based tool designed to analyze students’ essays and generate comprehensive feedback that includes concrete suggestions and engaging tips for improving writing.” [9] The authors reported that the use of PEER resulted in a good outcome for students and teachers of middle and upper level. However, their research indicates PEER still needs improvements as it is still in the early stages and needs further development, fine-tuning, and user evaluations.

A study to assess the agreement between instructor and ChatGPT was performed by [10]. These authors used 5 aspects (Goal, topic, benefit, novelty, and clarity) to evaluate how well 103 students in a postgraduate course performed in a data science project in business scenario. The authors tested three hypotheses. However, all of them being relevant, the one closely matches to our research is as follows “To what extent does the ChatGPT generated agree with instructor generated feedback when assessing students’ performance? In their research, [10] they invited experts to assess the instructors and ChatGPT student’s feedback. The method used was a polarity valuation where the feedback was either positive or negative. The result indicated ChatGPT feedback yielded the highest agreement on the Topic aspect while using precision and recall metrics. Despite this outcome, the authors concluded ChatGPT was not as reliable as the instructor, however, they still see potential in its effectiveness at providing feedback. This study varies from our research in two aspects. First, they used human evaluators to assess student’s feedback from instructors and that of ChatGPT. Second, they used precision and recall metrics

to determine how well the agreement was. [11] presented a study about a measure of agreement between instructor feedback and that of a LLM. The model was used to generate feedback for students' work on an engineering economy course where 70 cases about ethics were used. [11] used the Cohen's Kappa and Krippendorf's-α indicators to determine the level of agreement. The results, although promising, were that there was little agreement. Consequently, the LLM proved to be at most "slight" at providing feedback. However, it is promising, and further study needs to be done. One of the disadvantages of LLM is the large amounts of data to work reasonably. The research by [10] reported as limitation that no examples of good Data Science Projects were given to ChatGPT. The research presented in this paper provided several examples to ChatGPT of what an ethics case should be like.

Methodology

For this project, ethics papers were pulled from the class. The ethical dilemmas varied based on the underlying concepts. The basic concepts were prior knowledge of confidential events, workplace harassment, discrepancies in information regarding what was specified and what was received, and nepotism in the workplace. This study evaluated 69 papers from multiple sections of the course based on predetermined rubrics. All papers were independently reviewed by two separate evaluators and discussed to come to an agreed evaluator rating. These were then used to build background information to improve the AI models and to serve as one of the raters for measuring the agreement between the evaluators and the AI. The rubric used rates the criteria on a scale of 0 to 4 as shown in TABLE I.

TABLE I

ETHICS RUBRIC USED FOR PAPER ASSESSMENT BASED ON VALUE RUBRIC

4 - Excellent	3 - Good	2 - Fair	1 - Needs Improvement	0 - Failed
Demonstrates a comprehensive understanding of all ethical issues involved. Insightfully analyzes the complexities and nuances of the situation.	Shows a good understanding of the main ethical issues. Analyzes most aspects of the situation effectively.	Shows a basic understanding of some ethical issues. Analysis may be superficial or incomplete.	Fails to identify or adequately understand the ethical issues. Analysis is largely inaccurate or missing.	Fails to respond to the prompt.

To help generate background information for the AI on these dilemmas, example papers and rating were provided to the LLM. These were not past student papers but papers that were specifically written to show what levels of writing would yield each of the ratings. Once the background information was developed, it was added to the back-end of the AI models as a

retrieval augmented generation (RAG). The prompt used in scenario r2 for evaluating the papers in the AI is as follows:

Role and Tone: You are an experienced engineering professor evaluating a student's ethics paper. Your goal is to assess the work fairly using the rubric provided and to give clear, professional, and constructive feedback that helps the student improve.

Instructions:

1. Read the student paper carefully.
2. Evaluate it strictly using the rubric below. Do not invent additional criteria.
3. Assign one score (0–4) based on how well the paper meets the rubric description.
4. Justify the score with specific references to the paper.
5. Write feedback directly to the student in a professional, mentoring tone.
6. Identify strengths, weaknesses, and concrete ways to improve.
7. Do not comment on grammar, formatting, or style unless directly relevant to ethical analysis.
8. If the paper does not meaningfully address the prompt, assign a 0 and explain why.

Rubric:

4 – Excellent

Demonstrates a comprehensive understanding of *all* ethical issues involved. Insightfully analyzes the complexities and nuances of the situation.

3 – Good

Shows a good understanding of the *main* ethical issues. Analyzes most aspects of the situation effectively.

2 – Fair

Shows a basic understanding of *some* ethical issues. Analysis may be superficial or incomplete.

1 – Needs Improvement

Fails to identify or adequately understand the ethical issues. Analysis is largely inaccurate or missing.

0 – Failed

Fails to respond to the prompt.

Required Output Format:

1. Overall Assessment:

A brief summary of how effectively the paper addresses the ethical problem.

2. Rubric Evaluation:

- Score: X / 4
- Justification: Explain in detail how the paper aligns with the rubric description. Cite specific examples or passages from the paper.

3. Strengths:

- Bullet points describing what the student did well in their ethical reasoning or analysis.

4. Areas for Improvement:

- Bullet points identifying where ethical understanding, depth, or clarity is lacking.

5. Actionable Suggestions:

- Specific steps the student could take to improve the paper (e.g., "Address stakeholder impacts more explicitly by…," "Incorporate ethical frameworks such as…," "Expand your discussion of conflicting obligations by…").

Respond only with the evaluation and feedback. Do not restate the rubric or the full paper except for brief quoted examples used in justification.

Figure 1: Prompt used for evaluating student work

The specific LLM used in this evaluation is ChatGPT with the 5.2 model used to evaluate the papers. A custom GPT was built with the prompt, the background papers for ratings, and the original assignment. The training for models was turned off. As opposed to previous papers from the authors, this allowed for more recent models to be used and eliminated some of the technical requirements needed such as computing power. The Plus version of ChatGPT was used for this work.

Krippendorff's-α and Cohen's Kappa were used to compare the ratings of the evaluators to the ratings from the LLM. The measures are commonly used to evaluate interrater reliability. Interrater reliability refers to the degree of agreement among evaluators. Evaluators can be observers, raters, coders, analysists, and judges. The Krippendorff's-α is the reliability coefficient and it is applicable to categorical, ordinal, interval, or ratio-level data [12]. The underlying statistical calculations of the coefficient α make this measure appropriate for different levels of measurement and can be used when more than two raters participate in the analysis. The coefficient differentiates between observed disagreements and expected agreements. The observed disagreements come from the rated data, while the expected disagreement is the disagreement expected when random coding conditions are met [12]. See TABLE II Formula 1.

The Cohen's Kappa statistic typical application is to evaluate the agreement between two rater's measurement [13]. However, its use can be extended for those situations where there are more than two categories and where categorical variables have no order. In the case of ordered categories, the Weighted Kappa statistic can be used for three or more categories [14]. The Weighted Kappa is a variation of the Cohen Kappa and [15] list its use when nominal, categorical and ordinal variables are to be evaluated by two raters. It should be expected that Weighted Kappa values are higher than Cohen Kappa values. For Cohen Kappa formula and agreement values please refer to TABLE II. The Weighted Kappa values are displayed in TABLE IV. It is important to note that both measures take into consideration that collecting data may result in bias. Consequently, to minimize that problem, both measures include the expected values.

In this paper, the agreed to instructor rating and the LLM scenarios are considered separate raters. The agreed to instructor rating was utilized as the control (r1) and the prompted response was used as the experiment (r2). The output expected from the LLM is based on the ethics rubric shown in TABLE I. Once the scenarios were run, the analysis for the Krippendorf's-α and Cohen Kappa were performed and evaluated. For the Krippendorf's-α calculations, all data were assumed to be interval with a confidence interval of 0.95. The online calculator (https://www.k-alpha.org) was used to calculate the α. All calculations for the Cohen Kappa were computed in MS Excel. TABLE II displays the formulas and scales for both assessment metrics.

TABLE II

AGREEMENT CALCULATIONS AND SCALES *[11]*

	Krippendorf's-α [12]		Cohen's Kappa [13]	
	$\alpha = 1 - \frac{D_O}{D_e}$ (1)		$\kappa = \frac{P_O - P_e}{1 - P_e}$ (2)	
where:	$D_O = observed\ disagreement$ $D_e = expected\ disagreement$		$P_O = Probability\ that\ both\ raters\ agree\ is\ observed$ $P_e = Probability\ that\ both\ raters\ agree\ is\ expected$	
Agreement Scale	Krippendorf's-α [12]	Agreement	Cohen's Kappa [13]	Agreement
	α = 1	Perfect	κ > 0.8	Almost Perfect
	α ≥ 0.80	Acceptable	κ > 0.6	Substantial
	α = 0.67 – 0.79	Moderate	κ > 0.4	Moderate
	α < 0.67	Poor	κ > 0.2	Fair
	α = 0	None	κ = 0 – 0.2	Slight
	α < 0	Systematic Disagreement	κ < 0	Poor

Results

The Krippendorf's alpha calculation resulted in a value of 0.128, yielding a "Poor" agreement. Cohen's Kappa calculations resulted in a 0.17, indicating a "Slight" agreement between the raters. Based on these results, the hypothesis that the results show a strong level of agreement, with Cohen's Kappa and Krippendorff's-α > 0.6, between the LLM and the instructors' evaluation is rejected. Both measures indicate that there is little agreement between the raters. The confusion matrix is shown as TABLE III and shows the ratings provided by the evaluators (r1) and the LLM (r2).

TABLE III

COHEN'S KAPPA CONFUSION MATRIX

		r2						
	Ratings	4	3	2	1	0	Total	Proportion
r1	4	6	12	12	0	0	30	0.43
	3	1	21	10	1	0	33	0.48
	2	0	0	2	2	1	5	0.07
	1	0	0	0	0	0	0	0.00
	0	0	0	1	0	0	1	0.01
	Total	7	33	25	2	1		
	Proportion	0.10	0.48	0.36	0.04	0.01		

Because of the belief that LLM's are supposed to be able to reliably provide information to users, the rater's agreements were evaluated in a modified way. A weighted Cohen's Kappa was calculated, with the assumption that near miss disagreements should count more than larger disagreements in the ratings. The weighting calculation was applied to the proportion of ratings from the confusion matrix cells. The specific values for the weightings are shown in TABLE IV.

TABLE IV

WEIGHTED COHEN'S KAPPA VALUES

		r2				
	Ratings	4	3	2	1	0
	4	1	0.94	0.75	0.44	0
	3	0.94	1	0.94	0.75	0.44
r1	2	0.75	0.94	1	0.94	0.75
	1	0.44	0.75	0.94	1	0.94
	0	0	0.44	0.75	0.94	1

The weighted Cohen's Kappa then calculated as 0.25 instead of 0.17, allowing for a "Fair" agreement to be observed between the ratings. This still would not allow for a fail to reject the hypothesis, but it does allow for an improvement in the agreements between the ratings. The confusion matrix for the weighted Cohen's Kappa is shown in TABLE V.

TABLE V

WEIGHTED COHEN'S KAPPA CONFUSION MARTIX

		r2						
	Ratings	4	3	2	1	0	Total	Proportion
r1	4	6	11.25	9	0	0	26.25	0.38
	3	0.94	21	9.38	0.75	0	32.07	0.46
	2	0	0	2	1.88	0.75	4.63	0.07
	1	0	0	0	0	0	0	0.00
	0	0	0	0.75	0	0	0.75	0.01
	Total	6.94	32.25	21.13	2.63	0.75		
	Proportion	0.10	0.47	0.31	0.04	0.01		

Conclusion

There are still some takeaways from this study. For the most part, the faculty evaluators are more likely to rate a student's work higher on the rubric than the AI is. From the faculty side, this could be made due to several factors, including time available, fatigue, understanding of the materials, and the process of coming to an agreement on the ratings. From the AI side, there may still be work needed to improve the reference documents used to frame the problems. Though not discussed in the results, there is one of the scenarios where the weighted Cohen's Kappa did rate the agreement as moderate. The remaining question for the difference in this scenario's rating concerns the introduction of biases into the supporting documentation uploaded into the LLM [16]. If the data used to train the LLM had biases built in, it would use those biases in the work that it performed [16]. This provides a possible understanding of the discrepancies encountered in the paper.

A discussion about ethics on the use of AI tools in Education is provided by [17]. According to [17], the practice of ethics in other areas of education already exists. For example, there are ethics for research, e-learning environments, educational data, student privacy when monitoring for exams, and student cheating and fraud for e-learning systems. Whereas for AI, there is almost no known system or framework in place. While certainly the previous mentioned areas are always a concern, something is in place. There is, however, a real need for more research on ethics in the use of AI education (AIED). The problem is that it does not seem to be a priority research area within the AIED experts.

Another continuing dilemma in higher education concerning the ethical use of AI is how much of an instructor's duties can be replaced with AI and still be considered acceptable. AI, such as LLM's, continue to evolve, resulting in faster and more accurate work. In some ways, this increases an instructor's efficiency with their time, but a dilemma may emerge; is there a line where the pervasive use of AI replaces the instructor's work and detracts from the course content? There are existing complaints of university level classes generated by ChatGPT

materials instead of instructor generated materials [18]. Where should academia draw this line? Further work is still needed to define these ethical boundaries of AI use.

The future of AI in higher education is to continue developing or improving AI tools being used for feedback purposes while considering natural language process, educational data mining, and learning analytics [2]. The use of these technologies still possesses the challenge of existing or developing proper expertise. ChatGPT is already being used and as proposed by [8], ChatGPT functionalities can be used to help provide feedback. That study suggests this as a future area of research as well. This authors also touch on the issue of data security which falls within the many arguments of ethical use of AI in education.

References

[1] F. Tahiru, "AI in education: A systematic literature review," *Journal of Cases on Information Technology (JCIT),* vol. 23, no. 1, pp. 1-20, 2021.

[2] T. Wongvorachan, K. Lai, O. Bulut, Y. Tsai and G. Chen, "Artificial Intelligence: Transforming the Future of Feedback in Education," *Journal of Applied Testing Technology,* pp. 95-116, 2022.

[3] J. Schofield, D. Evans-Rhodes and B. Huber, "Artificial Intelligence in the Classroom: The impact of a Computer-based Tutor on Teachers and Students," *Social Science Computer Review,* vol. 8, no. 1, pp. 24-41, 1990.

[4] J. Anderson, C. Boyle and G. Yost, "The Geometry Tutor," *IJCAI,* pp. 1-7, 1985.

[5] L. Chen, P. Chen and Z. Lin, "Artificial intelligence in education: A review," *IEEE Access,* vol. 8, pp. 75264-75278, 2020.

[6] M. Donmaez, "AI-based feedback tools in education: A comprehensive bibliometric analysis study," *International Journal of Assessment Tools in Education,* vol. 11, no. 4, pp. 622-646, 2024.

[7] N. Anderson, J. Browning, D. Stewart, A. McGown and L. Galway, "Transforming Feedback with AI: A Smarter Way to Support Student Learning," *EDULEARN25 Proceedings,* pp. 6-12, 2025.

[8] P. Atherton, L. Topham and W. Khan, "AI and Student Feedback," *EDULEARN24 Proceedings,* 2024.

[9] K. Seßler, T. Xiang, L. Bogenrieder and E. Kasneci, "PEER: Emplowering writing with large language models," in *European Conference on Technology Enhanced Learning*, Switzerland, 2023.

[10] 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*, 2023.

[11] B. Gray and G. Fragoso-Diaz, "Comparing Feedback from AI and Human Instructor in an Engineering Economics Course," in *2025 ASEE Annual Conference & Exposition*, Montreal, 2025.

[12] G. Marzi, M. Balzano and D. Marchiori, "K-Alpha Calculator-Krippendorf's Alpha Calculator: A user-friendly tool for computing Krippendorf's Alpha inter-rater reliability coefficient," *MethodsX,* vol. 12, p. 102545, 2024.

[13] M. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica,* vol. 22, no. 3, pp. 276-282, 2012.

[14] M. Li, Q. Gao and &. T. Yu, "Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters," *BMC Cancer,* vol. 23, no. 1, p. 799, 2023.

[15] N. Gisev, J. S. Bell and &. T. F. Chen, "Interrater agreement and interrater reliability: Key concepts, approaches, and applications," *Research in Social and Administrative Pharmacy,* vol. 9, no. 3, pp. 330-338, 2013.

[16] A. Luke, "AI and the deskilling of teachers' work," *Australian Journal of English Education,* vol. 60, no. 1, p. 13, 2025.

[17] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. Shum and K. Koedinger, "Ethics of AI in Education: Towards a Community-wide Framework," *International Journal of Artificial Intelligence in Education,* vol. 32, no. 3, pp. 504-526, 2022.

[18] K. Hill, "Professors Face Student Rancor Over Use of A.I.," *New York Times,* p. A1, 18 May 2025.