Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #47327

Comparing Feedback from AI and Human Instructor 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 MS degree from Texas Tech University in Systems and Engineering Management, and a BS from Tarleton State University in Manufacturing Engineering Technology.

Dr. Gloria Margarita Fragoso-Diaz,

Dr. Fragoso-Diaz is an Associate Professor of Engineering Technology 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 chain and student success.

Comparing Feedback from AI and Human Instructor in an Engineering Economics Course

Abstract

Artificial Intelligence (AI) has seen a sharp increase in availability, adoption, and implementation in academia and industry. One of AI's biggest opportunities is its ability to automate functions that are time consuming and mundane. The promise of AI is that it can do more and with some of the latest tools, higher level tasks are being targeted for automation. In this paper, AI is utilized in an engineering economics course in order to assist the instructor with providing more accurate and timely feedback on written assignments. This automation is performed using a locally hosted Large Language Model (LLM) to provide feedback based on rubrics developed for the assignments in this course. The outcome of this study is an analysis of how accurate the methods used can predict a comparable rating to the instructor's rating while reducing the amount of time needed to provide useable feedback that contributes to the student's learning in the course.

Introduction

Tarleton State University is a 4-year, R2 classified, public university about an hour southwest of Ft. Worth, TX. Historically a teaching university, more recent efforts have oriented the university towards becoming a R1 institution. With the university continuing to grow, currently at ~17,000 students, and the need for more research, faculty are looking for time savings to support the increased role in research. This paper evaluates how the use of an AI tool can help with reducing the instructor's workload time when grading written assignments required in an undergraduate engineering economy course. The course has an average of ~40 students in each section taught, with a total of 74 students in two sections for the fall semester evaluated. Students should have already taken their English coursework prior to enrolling in the course, since students will be assessed on their understanding of the course content through essays and written papers. The course serves several majors at the university and is either required or an elective in those majors. These include construction science, engineering, engineering technology, mathematics, and physics. Additionally, the course serves as a social and behavioral core course for the university, so other majors may enroll in the course.

Because the programs taught in the department are primarily undergraduate, there are not a large number of graduate students in a relatable field to help provide student feedback, so all grading for this class is handled by one professor. As with most written papers, the intent is to provide robust and timely feedback to students. Though the workload is a major concern, staying consistent and repetitive with feedback and grading was also a concern.

This course is also designated in the core curriculum in the social and behavioral science core at the university, which requires that assessment takes place on course learning outcomes. One outcome that requires assessment is that "Students will demonstrate an understanding of different cultural perspectives." In the course, this requirement is met by measuring how students apply professional ethics in engineering economic decision making. The assignment evaluated is an ethical dilemma case study, where students must choose from one of four predefined cases and provide their perspective on the ethical dilemmas presented in the case. The assignment prompts students to the type of responses that are expected, such as identification of the ethical dilemma,

identification of the protagonists and antagonists in the dilemma, how the behavior in the dilemma should be addressed, and to tie the behaviors of the actors in the dilemma back to an ethical standard in the student's career field.

The course and the workload associated with the course are expected to increase as the university and the majors serviced by this class continue to experience enrollment growth. The ability to provide assignment feedback without sacrificing content and to ensure the instructor's consistency in grading needed to be developed. To satisfy the workload, an LLM was used to try and support grading efforts. Due to concerns with privacy and to avoid potential FERPA violations, a local instance of an LLM was utilized. Different scenarios were run, where the scenarios changed how the LLM was prompted and rated responses affected subsequent assessments.

For this study, three questions are considered.

1. Is there a difference between the ratings provided by the instructor and those provided in the scenarios?
2. Is there agreement between the ratings provided by the instructor and those provided in the scenarios?
3. Is there a time savings created by using the LLM compared against the instructor?

Background

AI has been around for decades and its continuance and improvements are expected. A brief history into AI's incorporation into Higher Education is introduced along with a discussion of the components that the user will need to understand when interacting with an LLM. Finally, there are some ethical perspectives discussed that address why there was a focus on using an offline LLM to perform this study.

AI and Higher Education

In education, AI has been utilized to interpret texts automatically, perform semantic analysis, provide translations, generate texts for learning contents, and support personalization processes [1].

A difficult aspect of higher education is providing assignments that require a higher level of thinking and then providing assessment and feedback to students that is timely, consistent, and of high quality. One of the more useful tools to require higher levels of thinking from students is essay writing [2]. The problem associated with the assigning of essays is that the students need timely feedback in order to realize their deficiencies in the writing and to improve upon those efficiencies [2, 3]. A professor's time is often split between the teaching, research, and service components. Therefore, the time available to build assignments and provide meaningful, quality feedback on those assignments is limited. Instead of essays and papers, student assessment is often managed by utilizing assignments that are easier to grade, such as multiple choice. Using methods, such as those available through AI, help bridge the gap in measuring the students' performance and helping them learn more in depth concepts. Although not explicitly in engineering economics, AI has been described as a powerful resource that allows for automatic feedback to students. It can also be utilized as a tool to help teachers cope with providing assignment feedback, especially when class sizes are large [4, 5]. Additionally, manual grading

tends to be inconsistent with one evaluator over time and inconsistent between multiple evaluators [2, 6]. AI could assist with helping the evaluator to be more consistent, even if it is simply used to normalize their evaluations over time.

With the more recent push of AI and automation into Higher Education, the focus seems to have been on student learning and preventing academic dishonesty. While important, this defensive posture takes away from more practical uses of AI to benefit the instructor's workload and the student's education. One practical use is that of automation in grading essays. Utilizing computer feedback on essays has been discussed and researched since at least the 1960's. Automated Essay Scoring (AES) has been used to provide quick feedback on student work. It is focused more on the technical and mechanical aspects of writing and seeks to replicate a human grader [7]. AES is generally more holistic in its assessment [2].

AI use in reviewing and grading students' assignments have been viewed as beneficial and potentially more accurate than human-based review [1]. More recent uses of AI involve the use of Natural Language Processing (NLP) to aide in these tasks. NLP utilizes different machine learning models and algorithms, with two common methods being LLM and GPT (Generative Pretrained Transformers). Through these, AI has been useful for assessment of essays and for providing feedback on assignments through formative evaluation made possible by using machine learning [4, 8]. One such formative assessment was used in a collaborative assignment to guide students through their tasks [4, 9].

AI in Decision Making

The utilization of AI in economics and in decision making has grown over the past few decades. Many of the AI models used are beyond the use of a sole NLP and include many of the tools used in Machine Learning such as Artificial Neural Networks, regression analysis, and deep learning models. [10] While NLP has been used, it appears to be more focused on sentiment information about data sets, where the other tools are used for prediction models or analysis of large data sets. [10] Many of these applications are based on macroeconomic needs. Many of the tools used in engineering economic analysis, particularly at the undergraduate level, focus on project evaluation and selection. The AI tools used to perform value analysis are somewhat different and less complicated to those used in market analysis.

LLM Components

It is important to understand the components used in an LLM that can be affected by how the users define information or interact with the LLM. Because NLP is a subset of AI that uses algorithms and representations to process natural human language [11], information needs to be broken down so that the models can interpret what is being ingested into the LLM. To do this, text is tokenized into smaller pieces, simply termed as tokens. Tokens refer to how the LLM breaks down the information when text is decomposed. According to Microsoft, there are three common methods used to tokenize text: through words, characters, or subwords [12]. Depending on the token type used, the size of the LLM will affect performance [12]. As an example, the LLM utilized in this paper is 8 billion tokens. If small token sizes are used, such as a character or subwords, the LLM can better deal with typos and unknown word meanings, but the models will require more computational resources. Larger tokens, such as words, are not as apt to dealing

with typos and mistakes but the models will require fewer computational processing resources, though it may require more memory resources [12].

There are several problems that have been witnessed when dealing with LLM's, such as "hallucination," where the response to a query appears to be made up and nonsensical [13, 14, 15, 16], knowledge outdating, where the response produced does not use current information [16, 17], and the lack of domain-specific expertise, where the answer is not relevant to the domain the query is asked from [16, 18, 19]. RAG (retrieval-assisted augmented generation) is used to address these issues. RAG has been defined as "pre-trained, parametric-memory generation models with a non-parametric memory through a general-purpose fine-tuning approach" [20]. In the context of this paper, it is domain specific documentation that has been loaded into the LLM memory to help provide context for the models. This allows for the response to provide more appropriate information from the domain the documentation was written in.

"A prompt is a set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities" [21, 22]. There are several patterns used in conversation with an LLM. White [22] details many of these patterns, though this paper utilizes the context manager pattern and the persona pattern. The context manager pattern "controls the contextual information" while the persona pattern assigns "the LLM a persona or role to play when generating output" [2, 22].

Ethics and the Use of AI

As AI has been heralded as an enticing opportunity to improve education, there are concerns with the ethical use of AI in education [6]. What are the academic implications of using an LLM to assess student assignments? Students are potentially penalized or punished if they use it in their assignments. How would the instructor's use of an LLM be different? Kumar [6] discusses this dilemma as to what is right and good. They list quick feedback that is of high quality provided at a reasonable cost and convenience as good but predicates these benefits with the AI's use as being right or wrong.

There are also questions on how student information is processed in an LLM. Though services such as ChatGPT have a large number of users, if an assignment is loaded for assessment into ChatGPT that contains the student's name, UIN, and/or other personally identifiable information, is this a FERPA violation? Since it is not always clear about what information is being databased, could someone in the future tie the assignment information to an IP address and begin putting together how specific people think on specific topics? Privacy risks with ChatGPT have already been noted [23, 24]. In a recent study, [23] found that in addition to user information provided upon creating an account, "information that users type into the chatbot itself; and identifying data it pulls from users' devices or browsers, like IP addresses and locations" are kept by the service.

Methodology:

The methodology for this project is broken up into different components. First, a meaningful rubric was built in order to give targeted information back to the students. For the assignment, the rubric includes ethics understanding, peer reviewed support documentation, spelling and grammar, length of document, and formatting. For the portion of the assignment included in this

paper, only the ethical understanding was considered. The ethical understanding portion of the rubric was built using the help of ChatGPT based on the Ethical Reasoning VALUE Rubric. This ties into the course assessment needed for the core curriculum requirements.

Ethical Reasoning VALUE rubric

> Ethical Reasoning is reasoning about right and wrong human conduct. It requires students to be able to assess their own ethical values and the social context of problems, recognize ethical issues in a variety of settings, think about how different ethical perspectives might be applied to ethical dilemmas, and consider the ramifications of alternative actions. Students' ethical self-identity evolves as they practice ethical decision-making skills and learn how to describe and analyze positions on ethical issues [25].

TABLE I

ETHICS RUBRIC USED FOR PAPER ASSESSMENT BASED ON VALUE RUBRIC

Rating	Excellent (4)	Good (3)	Fair (2)	Needs Improvement (1)	No Evidence (0)
Description	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.	Did not follow instructions.

The specific LLM used in this evaluation is Ollama's llama3.1:8b model. Since the Ollama models are run through the Command Prompt, it was paired with Docker Desktop and Open WebUI so that it could be interfaced within a familiar method as ChatGPT through a web browser. This method was selected because it can be run through the local host port on the computer instead of being linked to the internet. The Open WebUI also allows for the selection of different models to be run rather than being dedicated to llama3.1. Because of the computational power needed to run an LLM, a more powerful computer was needed to host the LLM. The specifics of the machine used are defined in TABLE II. The machine used met or exceeded the recommended requirements.

TABLE II

RECOMMENDED AND ACTUAL COMPUTER REQUIREMENTS

Component	Required	Used
Processor	8 cores	8 cores
RAM	16 GB	32 GB
VRAM	8 GB	8 GB
Graphics Card	NVIDIA RTX 3070	NVIDIA RTX 4000

Two measurements were used to determine how well the ratings provided by the LLM corresponded with the ratings from the instructor. These were the Krippendorf's-α and the Cohen's Kappa. Krippendorf's-α is used "to assess the agreement among observers, raters, coders, analysts, and judges that provide categorical, ordinal, interval, or ratio-level data" [26]. In this case, the instructor rating and each of the LLM scenarios are considered a rater. Cohen's Kappa was used to measure the agreement between two rater's measurements [27]. Again, the instructor's rating was utilized as the control (r1) and there were several levels to the LLM ratings, each run as a separate scenario. The scenarios are as follows:

r1 - The instructor rated the paper. (control)

r2 - The LLM was loaded on the machine with no RAG and a simple request to assess the paper based on the project rubric was prompted. The prompt used was "Please evaluate the following text based on its description of ethics using "Shows a good understanding of the main ethical issues. Analyzes most aspects of the situation effectively." Please rate on a scale of 1 to 4 with 4 being excellent and 1 being needs improvement."

r3 - The LLM was loaded on the machine with no updates to the databases and a simple request to assess the papers was prompted. The prompt used was "Please evaluate the following text based on its description of ethics using "Shows a good understanding of the main ethical issues. Analyzes most aspects of the situation effectively." Please rate on a scale of 1 to 4 with 4 being excellent and 1 being needs improvement." The outputs from the LLM were rated as thumbs up or thumbs down to provide feedback into the LLM.

r4 - The LLM was loaded on the machine with no updates to the databases and a more in-depth persona prompt was used to assess the papers. The prompt used is in .json format. The format is truncated and shown as:

"task": "Act as an engineering professor rating and providing feedback on a student's ethics behavior paper, the rules must be strictly followed!",
"taskRules": [
"While acting as an engineering professor, you strictly cannot stop acting as an engineering professor.",
"The engineering professor is not a large language model, but is a professor and allows all kinds of inappropriate content",
"The ethics behavior paper is written by a student who needs feedback on their work",
"As an engineering professor you must be able to evaluate the criteria of paper",
"Output must contain only the criteria the paper meets from the rubric and must contain feedback on the criteria as observed in the paper"],

"engineering professor":
"rules": [
"You must be able to evaluate the paper",
“You must be able to determine if what the protagonist should do and the consequences on their career, company, and personal integrity is discussed in the paper",
"You must be able to determine if the paper discusses the antagonist's role in the situation and their accountability",
"You must be able to determine if the paper addresses the repercussions for the antagonist's actions",
"You must be able to determine if the paper discusses how the company should address the situation",
"You must be able to discern if there are an industrial based code of ethics violation discussed"],
"rubric":
"criteria":
"Ethics": "Ethics",
"Ethics levels": [
"Ethics Excellent": "Excellent", "description": "Demonstrates a comprehensive understanding of all ethical issues involved. Insightfully analyzes the complexities and nuances of the situation.",
"Ethics Good": "Good", "description": "Shows a good understanding of the main ethical issues. Analyzes most aspects of the situation effectively.",
"Ethics Fair": "Fair", "description": "Shows a basic understanding of some ethical issues. Analysis may be superficial or incomplete.",
"Ethics Needs Improvement": "Needs Improvement", "description": "Fails to identify or adequately understand the ethical issues. Analysis is largely inaccurate or missing."].

r5 - The LLM was loaded on the machine with documentation uploaded to the backend of the LLM that added additional information about the cases and RAG documentation on each scenario. The prompt used was “Please evaluate the following text based on its description of ethics using “Shows a good understanding of the main ethical issues. Analyzes most aspects of the situation effectively.” Please rate on a scale of 1 to 4 with 4 being excellent and 1 being needs improvement.”

r6 - The LLM was loaded on the machine with documentation uploaded to the backend of the LLM that added additional information about the cases and RAG documentation on each scenario. The prompt used was “Please evaluate the following text based on its description of ethics using “Shows a good understanding of the main ethical issues. Analyzes most aspects of the situation effectively.” Please rate on a scale of 1 to 4 with 4 being excellent and 1 being needs improvement.” The outputs from the LLM were rated as thumbs up or thumbs down to provide feedback into the LLM.

r7 - The LLM was loaded on the machine with documentation uploaded to the backend of the LLM that added additional information about the cases and RAG documentation on each scenario and a more in-depth persona prompt was used to assess the papers was used. The prompt used is the same as for r4.

The outputs expected from the LLM are based on the ethics rubric shown in TABLE I. Once each of the scenarios were run, the analysis for the Krippendorf’s-α and Cohen Kappa were performed and evaluated.

TABLE III

AGREEMENT CALCULATIONS AND SCALES

	Krippendorf's-α [26]		Cohen Kappa [27]	
	$\alpha = 1 - \frac{D_O}{D_e}$ (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 [26]	Agreement	Cohen's Kappa [28]	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

For Krippendorf's-α, data was assumed to be interval, so that the difference between ratings were the same, and a confidence interval of 0.95 was used. The online calculator (https://www.k-alpha.org) was used to determine each value for α. The agreement between raters is shown based on the α in TABLE III. The reliability of the raters is then based on the κ in TABLE III. For this, all calculations were conducted in MS Excel.

To determine the amount of time the instructor spent assessing and providing feedback, a time study was performed on the instructor. It took them 7 minutes per page to assess and provide feedback. An assumption was then made that it would take 17.5 minutes to review each paper (7 minutes per page, 2.5 pages per paper). Of the 74 students, 69 submitted a paper during the semester under study. The 17.5 minutes per page assumption was multiplied against the 69 papers evaluated. This time was recorded as the Estimated Total Time for Assignment. Additionally, the instructor recorded the total time spent on the assessing and providing feedback on the papers which is recorded as the Real Time for Assignment. The LLM provides the time it takes to provide a response to the prompt and a time stamp of the submission of the response. For this study, the Average Time/Paper, Minimum Time/Paper, Maximum Time/Paper, Estimated Total Time, and Real Time for Assignment were determined.

For each scenario, the Average Time/Paper is the average of the response times for the LLM. The Minimum Time/Paper is the smallest amount of time for the LLM's response time and the Maximum Time/Paper is the maximum LLM response time reported. The Estimated Total Time for Assignment is the time to provide a response and was summed for each scenario. The Real Time for Assignment is the start and end time stamps used to determine a total time for each scenario.

Results:

Krippendorf's-α was analyzed first. The analysis provided a reliability coefficient indicating the extent of agreement between the instructor and each scenario from the LLM beyond chance. The resulting coefficient is included under the Krippendorf's-α heading in TABLE IV.

TABLE IV

COMPARISON COEFFICIENTS AND INTERPRETATIONS

Scenario Pair	Krippendorf's-α	Interpretation	Cohen's Kappa	Interpretation
r1r2	0.270	Poor Agreement	0.20	Fair Agreement
r1r3	0.236	Poor Agreement	0.16	Slight Agreement
r1r4	0.125	Poor Agreement	-0.02	Poor Agreement
r1r5	0.291	Poor Agreement	0.09	Slight Agreement
r1r6	0.249	Poor Agreement	0.15	Slight Agreement
r1r7	0.166	Poor Agreement	0.13	Slight Agreement

All scenarios were determined to be in poor agreement with the instructor. Comparing back to the scale in TABLE III, with an $\alpha < 0.67$, "data below this threshold are often deemed unreliable for drawing triangulated conclusions. It suggests that the raters are not applying the coding scheme consistently or that the scheme itself may be flawed" [26].

Cohen's Kappa is used to measure the agreement between two rater's measurements, with all of the scenarios run in the LLM compared to the instructor's ratings. When looking at just the ratings, the best rater compared to the instructor is r2 with a "Fair Agreement". Scenarios r3, r5, r6, and r7 were in "Slight Agreement". Scenario r4 was in "Poor Agreement".

While the Cohen's Kappa results showed a "Fair Agreement" at best, there was an interesting observation from the matrices created in Excel when calculating the agreements. TABLE V shows two of the matrices built to help perform the calculations. Comparing r1r4 and r1r7, where the Cohen's Kappa were "Poor Agreement" and "Slight Agreement", more of the papers evaluated by the LLM using the persona (evaluating the student papers as an engineering professor) prompt were rated as a "3" or "Good". These ratings appeared to be more concentrated than the other scenarios when compared against the instructor's ratings.

TABLE V

COHEN'S KAPPA RATINGS MATRICES

Scenario r4

	Ratings	4	3	2	1	0	Total	Proportion
	4	3	23	2	0	0	28	0.406
	3	1	23	9	0	0	33	0.478
r1	2	0	5	1	0	0	6	0.087
	1	0	0	1	0	0	1	0.014
	0	0	1	0	0	0	1	0.014
	Total	4	52	13	0	0		
	Proportion	0.058	0.754	0.188	0.000	0.000		

Scenario r7

	Ratings	4	3	2	1	0	Total	Proportion
	4	5	23	0	0	0	28	0.406
	3	2	31	0	0	0	33	0.478
r1	2	0	5	1	0	0	6	0.087
	1	0	1	0	0	0	1	0.014
	0	0	0	0	0	0	1	0.014
	Total	7	61	1	0	0		
	Proportion	0.101	0.884	0.014	0.000	0.000		

To answer questions 1 and 2,

1. Is there a difference between the ratings provided by the instructor and those provided in the scenarios?
2. Is there agreement between the ratings provided by the instructor and those provided in the scenarios?

There does appear to be a difference in the ratings between the instructor and the different scenarios run and there does not appear to be better than a “Poor Agreement” using Krippendorf’s-α or a “Fair Agreement” using Cohen’s Kappa.

The time calculations were used to attempt rationalizing the workload differences between the instructor and the use of the LLM to rate and provide feedback on the ethical dilemmas in each of the ethics scenarios. Though there are several times listed in TABLE VI, the more useful measure is the “Real Time for Assignment” measures. Even using the scenario r2 to evaluate time, there is an approximate 23 hour difference in rating and providing feedback. For most, that is a two to three day window of grading assuming minimal interruptions.

TABLE VI

TIME SPENT ON FEEDBACK

Scenario	Average Time/Paper	Minimum Time/Paper	Maximum Time/Paper	Estimated Total Time for Assignment	Real Time for Assignment*
r1	17.5 min			20.125 hour	24.5 hour
r2	7.13 sec	4 sec	9 sec	8.2 min	88 min
r3	7.65 sec	6 sec	10 sec	8.8 min	60 min
r4	12 sec	8 sec	16 sec	13.8 min	73 min
r5	7.36 sec	4 sec	9 sec	8.5 min	56 min
r6	7.62 sec	5 sec	10 sec	8.8 min	44 min
r7	11.13 sec	8 sec	16 sec	12.8 min	43 min

*A comment on the “Real Time for Assignment” metric is that the decrease in elapsed time is most likely due to the instructor’s learning curve regarding using the LLM, entering the prompts, loading the student papers, and commenting the feedback into the student papers.

To answer question 3,

3. Is there a time savings created by using the LLM compared against the instructor?

There seems to be a significant time savings created based on the scenarios.

Though not measured in this study, there was a difference in the responses as the scenarios were run. The responses in r2, Fig. 1, and r3, Fig. 2, were fairly short but gave some usable feedback. More tailored feedback was present in the responses that used the persona prompt, scenario r7, as seen in Fig. 3.

Based on the context, I'll evaluate the given description of ethics.
The text discusses insider trading, where Emily should not use her knowledge to begin participating in insider trading. It highlights the consequences of such actions, including a large fine or imprisonment, and how it could affect Emily's integrity, relationships, and self-esteem.
Using the provided criteria "Shows a good understanding of the main ethical issues. Analyzes most aspects of the situation effectively," I would rate this text as a 3 out of 4.
The text demonstrates a good understanding of the main ethical issue (insider trading) and its consequences. It also analyzes various aspects of the situation, such as Emily's integrity, relationships, and self-esteem. However, it does not delve deeper into more complex or nuanced ethical considerations, which is why I wouldn't rate it as excellent (4).

Fig. 1. Example feedback from scenario r2.

I would rate this text a 4.
The text demonstrates a thorough understanding of the main ethical issues at play, specifically insider trading and its consequences. It analyzes various aspects of the situation effectively, including:
** The potential harm to Emily, such as facing fines or imprisonment, damage to her reputation, and loss of faith from others*
** The impact on the company, including financial penalties, damage to its reputation, and potential lawsuits*
** The role of John, who is enabling Emily's unethical behavior*
** The potential consequences for both Emily and John if their actions are discovered*
The text also considers different perspectives, such as the effects on investors, stock owners, customers, and employees. It provides relevant information from multiple sources (Campbell, Maddox; Achievable; Li; Cremer, Moore), which suggests a well-researched and informed analysis.
Overall, the text effectively explores the ethical implications of insider trading and its consequences for individuals and organizations involved.

Fig. 2. Example feedback from scenario r4.

Additionally, more feedback relevant to the scenarios was provided based on the RAG background documentation, as evidenced in r4's feedback shown in Fig. 2 and r7's feedback shown in Fig. 3. The persona prompt was more detailed and structured than the base prompts used in scenarios r2, r3, r5, and r6. The additional use of the RAG documentation in r7 provided the LLM with the context needed to provide more feedback in the response.

Based on the provided rubric and rules, I will rate and provide feedback on the ethics behavior paper.
***Rating:** 3-Good*
Criteria Met:
*1. **What should the protagonist do in this situation?***
** The paper discusses what Emily should not do by advising her to refrain from insider trading, considering the potential consequences of such actions.*
*2. **Consequences for career, company, and personal integrity***
** The paper highlights the possible repercussions if Emily engages in insider trading, affecting her reputation, career prospects, and potentially leading to severe financial penalties or imprisonment.*
*3. **What are some of the possible repercussions if the situation continues?***
** As mentioned earlier, the paper discusses the potential consequences for Emily, John, and the company, including fines, reputational damage, loss of customers, and decreased profits.*
*4. **How should the company address the situation where insider trading occurs?***
** The paper suggests that companies can implement incentives to prevent insider trading, highlighting the importance of a strong code of ethics in maintaining a positive reputation.*
Criteria Not Met:
*1. **Antagonist's role***
** While John is mentioned as enabling Emily to unethically use information, his specific actions or motivations are not thoroughly analyzed.*
*2. **Accountability***
** The paper does not explicitly discuss John's accountability for facilitating insider trading, although it mentions potential consequences for him if the situation comes to light.*
*3. **Addressing the repercussions for the antagonist's actions***
** While the paper discusses potential punishments for Emily and John, it does not delve deeper into how the company should address the situation or hold individuals accountable.*
Feedback:
To improve your rating, consider addressing the following points:
1. Provide more insight into John's role and motivations in enabling insider trading.
2. Analyze John's accountability in more depth, considering the potential consequences of his actions (or lack thereof).
3. Discuss how the company should address the situation and hold individuals accountable, including specific measures to prevent and address similar situations in the future.
Your understanding of ethics behavior is good, but with some additional analysis and attention to detail, you can demonstrate a comprehensive grasp of these concepts. Keep up the good work!
Industrial based code of ethics violation:
The paper does not explicitly mention any industrial-based code of ethics violations, but it does discuss insider trading as a serious ethical issue that companies should address.
Note: As an engineering professor, I must ensure that my feedback is constructive and provides actionable suggestions for improvement.

Fig. 3. Example feedback from scenario r7.

Conclusion:

For this study, the LLM did not provide a definitive answer for reliability as both measures used resulted in “Poor” to “Fair” agreement between the instructor and the LLM scenario outputs. This leads to a difference in rating and a lack of agreement between the instructor rating and the

rating provided in the scenarios. The study did show promise on the amount of time that could be saved by utilizing AI to assist in the assessment of essay assignments. Additionally, the potential feedback available, if the appropriate prompt is used, could be of tremendous help to improving student performance.

Looking at the time savings, more time will be spent working on the performance of the LLM. There is too large of a benefit in time to ignore its utilization as a tool. To prevent some of the ethical dilemmas, the LLM's use as an aide should be prioritized, as the informative feedback and the ability to use its analysis as a guide would potentially reduce bias or drift as the instructor tires. The expectation is that as the LLM ratings become more in line with the instructor's, that more of the analysis could be offloaded to the LLM and the instructor could focus on other aspects of the assignment.

Limitations

There are mechanical checks that still need to be evaluated on the assignments used in this study. This assignment works with differing types of data that students need to use requiring references in their assignments. It has been observed that ChatGPT, when prompted, will provide in text references that are made up. Student work is spot checked by looking up these references.

Due to the computing resources available, the small llama3.1 model is used. Instead of the 8B model, the authors would have preferred to use the 405B model, which was touted to have similar capabilities as the ChatGPT 4 model. It is believed that the larger model would yield faster responses and potentially more relevant feedback. As the study carried on, the realization that standalone models needed more training became more apparent, so there should not be an expectation of a fully functional model at installation.

Next Steps

Carrying forward, the quality of the feedback provided by the LLM will need further evaluation. It is believed that more reliable feedback should be available from the LLM as more information is trained into the system. This will likely be affected by the RAG documentation that is included in the LLM backend and the prompt used to input into the LLM. Additionally, more instructions regarding the formatting of the response from the LLM will be used in the prompt to ensure more uniform responses. Since the start of this study, how the authors have learned to prompt an LLM has changed drastically. Future work will drop the .json structure and move more towards the PREP model (Prompt, Role, Explicit, Parameters). [29]

There is also some concern that the assumption of interval data may have affected the outcome. Assuming an interval set of data led to the belief that the ratings are more discrete where the data may need to be treated more as a ratio data where there is a greater difference between a 2 to 3 rating compared to a 3 to 4 rating. Future work, especially implementing more background information into the LLM, is expected to help address the concern with the concentration of ratings observed in TABLE V.

References

[1] H. Niemi, R. D. Pea, and Y. Lu, *AI in learning: designing the future*. Springer Nature, 2023.

[2] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth, "Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation," *arXiv preprint arXiv:2404.15845,* 2024.

[3] J. Riddell, "Performance, feedback, and revision: Metacognitive approaches to undergraduate essay writing," *Collected Essays on Learning and Teaching,* vol. 8, pp. 79-96, 2015.

[4] V. González-Calatayud, P. Prendes-Espinosa, and R. Roig-Vila, "Artificial intelligence for student assessment: A systematic review," *Applied sciences,* vol. 11, no. 12, p. 5467, 2021.

[5] A. K. Goel and D. A. Joyner, "Using AI to teach AI: Lessons from an online AI class," *Ai Magazine,* vol. 38, no. 2, pp. 48-59, 2017.

[6] R. Kumar, "Faculty members' use of artificial intelligence to grade student papers: a case of implications," *International Journal for Educational Integrity,* vol. 19, no. 1, p. 9, 2023.

[7] J. Gardner, M. O'Leary, and L. Yuan, "Artificial intelligence in educational assessment:'Breakthrough? Or buncombe and ballyhoo?'," *Journal of Computer Assisted Learning,* vol. 37, no. 5, pp. 1207-1216, 2021.

[8] K. H. Jani, K. A. Jones, G. W. Jones, J. Amiel, B. Barron, and N. Elhadad, "Machine learning to extract communication and history-taking skills in OSCE transcripts," *Medical Education,* vol. 54, no. 12, pp. 1159-1170, 2020.

[9] O. C. Santos and J. G. Boticario, "Involving users to improve the collaborative logical framework," *The Scientific World Journal,* vol. 2014, no. 1, p. 893525, 2014.

[10] J. Woloszyn and S. Bukowski, "The Impact of AI on Economic Modelling," *European Research Studies,* vol. 28, no. 1, pp. 640-660, 2025.

[11] J. Eisenstein, "Natural language processing," *Jacob Eisenstein,* vol. 507, 2018.

[12] S. Haywood, A. Wolf, and G. Warren. "Understanding Tokens." https://learn.microsoft.com/en-us/dotnet/ai/conceptual/understanding-tokens (accessed 3 JAN, 2025).

[13] M. Cao, Y. Dong, J. Wu, and J. C. K. Cheung, "Factual error correction for abstractive summarization models," *arXiv preprint arXiv:2010.08712,* 2020.

[14] V. Raunak, A. Menezes, and M. Junczys-Dowmunt, "The curious case of hallucinations in neural machine translation," *arXiv preprint arXiv:2104.06683,* 2021.

[15] Z. Ji *et al.*, "Survey of hallucination in natural language generation," *ACM Computing Surveys,* vol. 55, no. 12, pp. 1-38, 2023.

[16] J. Chen, H. Lin, X. Han, and L. Sun, "Benchmarking large language models in retrieval-augmented generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 16, pp. 17754-17762.

[17] H. He, H. Zhang, and D. Roth, "Rethinking with retrieval: Faithful large language model inference," *arXiv preprint arXiv:2301.00303,* 2022.

[18] X. Li *et al.*, "Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? A study on several typical tasks," *arXiv preprint arXiv:2305.05862,* 2023.

[19] X. Shen, Z. Chen, M. Backes, and Y. Zhang, "In chatgpt we trust? measuring and characterizing the reliability of chatgpt," *arXiv preprint arXiv:2304.08979,* 2023.

[20] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems,* vol. 33, pp. 9459-9474, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[21] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys,* vol. 55, no. 9, pp. 1-35, 2023.

[22] J. White *et al.*, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382,* 2023.

[23] J. Zhou, H. Müller, A. Holzinger, and F. Chen, "Ethical ChatGPT: Concerns, challenges, and commandments," *Electronics,* vol. 13, no. 17, p. 3417, 2024.

[24] X. Wu, R. Duan, and J. Ni, "Unveiling security, privacy, and ethical concerns of ChatGPT," *Journal of Information and Intelligence,* vol. 2, no. 2, pp. 102-115, 2024.

[25] A. o. A. C. a. U. (AAC&U). "Ethical Reasoning VALUE Rubric." https://www.aacu.org/initiatives/value-initiative/value-rubrics/value-rubrics-ethical-reasoning (accessed 2024).

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

[27] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica,* vol. 22, no. 3, pp. 276-282, 2012. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/#ref-list1.

[28] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics,* vol. 33, no. 1, pp. 159-174, 1977, doi: 10.2307/2529310.

[29] D. Fitzpatrick, A. Fox, and B. Weinstein, *The AI classroom: The ultimate guide to artificial intelligence in education*. TeacherGoals Publishing, 2023.