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 #47586

Exploring the Capability of Generative AI as an Engineering Lab Report Assessment Assisting Tool

Dr. Dave Kim, Washington State University-Vancouver

Dr. Dave Kim is Professor and Mechanical Engineering Program Coordinator in the School of Engineering and Computer Science at Washington State University Vancouver. His teaching and research have been in the areas of engineering materials, fracture mechanics, and manufacturing processes. In particular, he has been very active in pedagogical research in the area of writing pedagogy in engineering laboratory courses. Dr. Kim and his collaborators attracted close to $1M in research grants to study writing transfer of engineering undergraduates. For technical research, he has a long-standing involvement in research concerned with the manufacturing of advanced composite materials (CFRP/titanium stack, GFRP, nanocomposites, etc.) for marine and aerospace applications. His recent research efforts have also included the fatigue behavior of manufactured products, with a focus on fatigue strength improvement of aerospace, automotive, and rail structures. He has been the author or co-author of over 200 peer-reviewed papers in these areas.

Dr. Charles Riley P.E., Oregon Institute of Technology

Dr. Riley has been teaching mechanics concepts for over 10 years and has been honored with both the ASCE ExCEEd New Faculty Excellence in Civil Engineering Education Award (2012) and the Beer and Johnston Outstanding New Mechanics Educator Award (2013).

Exploring the Capability of Generative AI as an Engineering Lab Report Assessment Assisting Tool

Abstract

Since ChatGPT's public launch in November 2022, considerable discussion and changes have occurred in higher education. Active educational research related to generative artificial intelligence (GAI) has been conducted in various areas, including student learning, ethics, and assessment. Although many authors have raised concerns about the impact of GAI, particularly a large language model (LLM), in writing education, the systematic studies related to the ethical use of GAI are limited. While grounded in the ethical adaptation of GAI in grading and feedback for engineering lab writing, we focus on GAI's capability to assist with engineering lab report assessment. Lab report grading is time-consuming for lab instructors and teaching assistants. Moreover, constructing impactful feedback can be challenging for many reasons. In this pilot study, we used Copilot and ChatGPT 4o to conduct evaluation and feedback on student lab reports of past courses when the instructors did not use generative AI technologies. The study space was limited to the two engineering labs in two institutions: strength of materials for mechanical and civil engineering students at a 4-year public polytechnic university and engineering materials for mechanical engineering students at a 4-year R1 university. GAI tools were asked to generate scores, overall reviews, suggestions, or improvement tips. We compared the evaluation scores and feedback of each student lab graded by instructors or graduate teaching assistants with those from GAI tools. The comparative analysis results will be discussed to answer how the GAI tool's evaluation results align with scores and feedback by instructor/TAs regarding accuracy and clarity.

1. Introduction

Lab education is essential in college engineering as it offers students hands-on experience with critical technical skills, such as operating equipment, conducting experiments, and developing problem-solving and critical thinking abilities [1]. Often, lab courses are offered in the early phase of engineering majors to provide students with hands-on experience and a foundational understanding of core engineering principles. For engineering labs, a range of assessment methods exists and includes lab reports, quizzes and exams, post-lab assignments, lab practicals, and instructor observations. Among these, lab reports are the most dominant assessment method for evaluating students' learning from the labs. Indeed, lab report writing aligns well with the "write to learn" approach - an active learning approach - by encouraging students to reflect on their learning from the labs, reinforce key engineering concepts, and synthesize their technical knowledge through practical application [2]. At the same time, lab writing exercises support students in becoming effective communicators by engaging them in purposeful documentation for a technical audience [3]. For this reason, lab reports are often used to assess ABET's engineering student learning outcomes 3 and 6 [4] in many engineering programs. Although lab reports are one of the predominant tools for evaluating engineering students' achievement in experimentation and communication skills, assessing lab reports is challenging for engineering educators. Lab report graders must accurately and precisely assess the student's achievement in various aspects while providing productive feedback [5]. Timely return of lab reports further complicates the assessment process because of their length and the time required to grade them effectively. Lab report grading is burdensome for many lab instructors, including teaching assistants, and many engineering lab instructors have sought ways to get assistance in lab report grading.

Since ChatGPT's public launch in November 2022, many generative artificial intelligence (GAI) chatbots have been introduced to the public [6]. While each chatbot service has its niche areas, they commonly generate human-like responses to queries. Although many educators raised concerns about the impact of GAI chatbots on writing education [7], many studies have been published on the ethical adaptation of GAI chatbots in writing education. One of the areas GAI chatbots can contribute is the evaluation of students' written reports. Zhou et al developed an automatic scoring system based on ChatGPT for the Automatic Control Theory experiment course at Wuhan University [8] to show the system's high exactitude and reliability. GAI chatbot's capability to provide feedback scripts on any submission can assist lab report grading in engineering lab courses; however, there are many unknowns related to using GAI chatbots for lab report assessment, including the ethics of such use, the accuracy of the grades, and the quality of the feedback.

Among many GAI chatbots, we focus on two commercially available and widely used chatbot platforms: Microsoft 365 CoPilot and ChatGPT-4o. This pilot research work is designed to answer the following research questions:

1. How can the two GAI chatbots be ethically used for student lab report grading in engineering labs?
2. How do the chatbots' assessment scores align accurately with those of the lab instructors or teaching assistants?
3. How does the chatbot's feedback differ from that of lab writing assessors?

2. Methods of Approach

2.1 Study area

This study was conducted at two universities: Oregon Institute of Technology (OIT), a polytechnic university offering ABET-accredited programs in civil, electrical, mechanical, and renewable energy, and Washington State University Vancouver (WSUV), a branch campus of a research-one (R1) land grant university offering ABET-accredited programs in electrical and mechanical engineering. OIT had around 650 students and 30 faculty members in its engineering programs, offering multiple engineering lab courses, including sophomore-level lab courses in civil engineering taught by faculty and supported by undergraduate teaching assistants. WSUV's engineering programs, with about 350 students and 15 faculty members, included junior-level mechanical engineering courses in the study, all taught by graduate teaching assistants supervised by instructors. Table 1 presents the basic information for the participating lab courses from the two institutions we studied.

Table 1: Participating engineering laboratory courses in the study

Case	Institution, Semester/Quarter	Major	Course	Topic	Term	Labs analyzed in this study	Labs taught by	Lab report evaluated by	Student sample number
1	4-year public polytechnic college (OIT), Quarter	Civil	ENGR 213	Strength of Materials	Fall 2022	Lab 3 Tensile testing of polymers	Instructor	Undergrad teaching assistant	10
2	4-year public polytechnic college (OIT), Quarter	Civil	ENGR 213	Strength of Materials	Fall 2024	Lab 3 Tensile testing of polymers	Instructor	Undergrad teaching assistant	10
3	4-year public college (WSUV), Semester	Mecha-nical	MECH 309	Engineering Materials	Fall 2022	Lab 4 Tensile tesing	Graduate teaching assistant	Graduate teaching assistant	20

4					Fall 2016	Lab 4 Tensile tesing	Graduate teaching assistant	Faculty instructor	20

2.2 Student lab report sample collection and their evaluation by the instructors

We collected students samples randomly in ENGR 213 (Strength of Materials) in Fall of 2022 and 2024 (*n=10 for each case*), and MECH 309 (Engineering Materials) in Fall of 2016 and Fall of 2022 (*n=20 for each case*). The course delivery modes were in-person. All the lab reports were evaluated by TAs or the faculty instructor, and they did not use GAI during their assessment. The rubrics used by the lab writing assessors when assessing all the sample lab reports collected are included in Appendix A.

2.3 Evaluation processes by generative AI chatbots

All the information related to personal identifying information, institution information, course information, and dates were removed from individual samples. All the report samples are stored in PDF before submission to GAI chatbots. The following are the prompts used in the chatbot for each sample evaluation:

- I need your help assessing lab reports for an engineering lab course, which is a college {sophomore} course about {tensile testing of polymers}.
- I have my own rubric to share with you. Please assess each lab report using this rubric. (Upload the lab report rubric in PDF)
- I need your scores for each category with a short assessment description. You can give scores between {100 and 0}, for example, {85 or 24}.
- Assess this report (Upload a student sample in PDF for CoPilot) OR Assess these reports. (Upload a set of up to ten student samples in PDF for ChatGPT4o).

We changed the content in {   } to align with each lab's topic and rubric scoring system. CoPilot did not accept multiple file uploading, while ChatGPT4o could accept uploading up to ten PDF files. The users can determine the outcome formats as descriptions or tables, which were downloadable Excel files.

3. Results and Discussion

3.1 Ethical Considerations on Lab report Assessment by Generative AI Chatbots

The two universities in the study area have licenses from Microsoft to access Microsoft's artificial intelligence (AI) companion, CoPilot, within the university's network authentication. With the university-authenticated sessions, CoPilot ensures privacy and data security by not making chat history accessible to users or university IT administrators and discarding prompts and responses when the web browser closes, the chat topic resets, or the session times out. CoPilot does not access any data within the university's Microsoft 365 environment, and the user information is removed at the start of each session solely to verify eligibility for Copilot access. Additionally, no data is sent to external providers, and prompts and responses are not used as training data for large language models [9]. The use of Copilot works the same as TurnItIn, a web-based plagiarism detection tool, to submit a student's work to the system within university's authenticated sessions.

WSU has published guidance on data stewardship and artificial intelligence on June 21, 2023 [10]. It stated that "Users of ChatGPT and similar artificial intelligence (AI) technology or AI programs must avoid integrating, entering, or otherwise incorporating any non-public institutional data or information, including but not limited to personal identifying information or research data." Only data and information designated as "public" under the school policies may be used with AI tools. The school policy [11] defines public information examples as widely distributed materials and public research publications. OIT is preparing their guidelines for data stewardship on AI technologies.

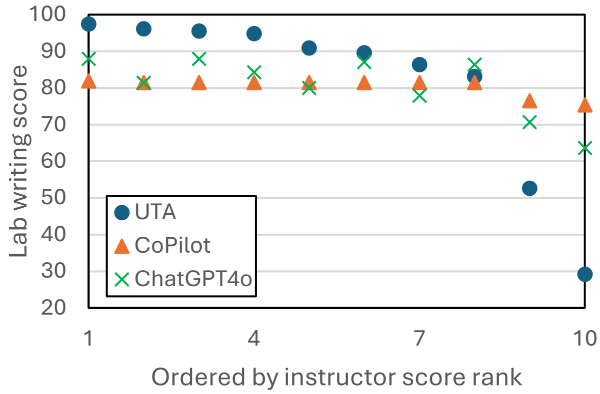
To avoid potential risks associated with the use of ChatGPT-4.0o in this study, we removed all the information related to personal identifying information, institution information, course information, and dates. The technical information included in lab reports can be considered as widely distributed materials that are not restricted from being disclosed to the public. Therefore, the sample preparation was conducted before submitting the samples to GAI chatbots.

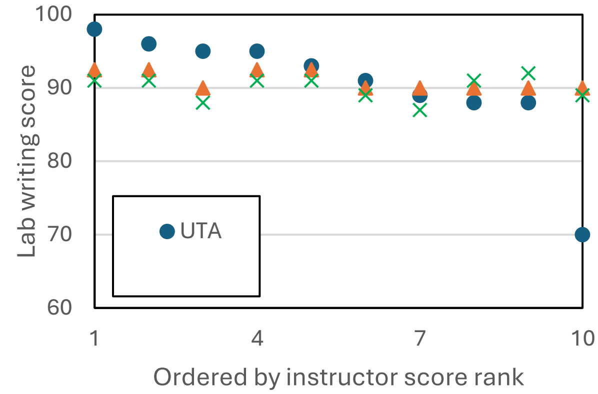
WSU's IRB considered this study as Not Human Subject Research (NHSR), because there was no "interaction or intervention" with living individuals; therefore, 45CFR46.102 (e)(1)(i) does not apply. In addition, the researcher is not utilizing identifiable data or biospecimens about a living individual; therefore, 45CFR46.102 (e)(1)(ii) does not apply. Instead, IRB noted that the authors have a responsibility to oversee the project and ensure the ethical principles outlined in the Belmont Report are upheld.

3.2 Quantitative Analysis Results: Evaluation score comparisons

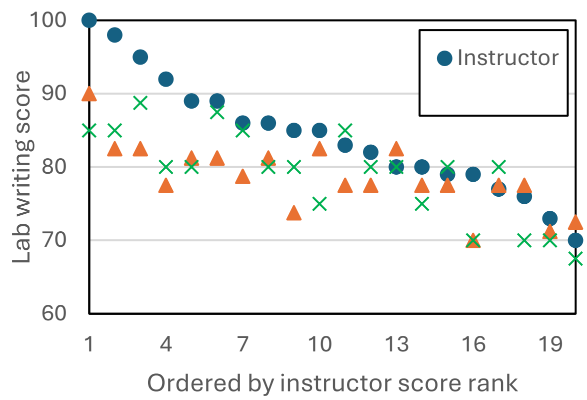
3.2.1 Lab report score and rank comparisons

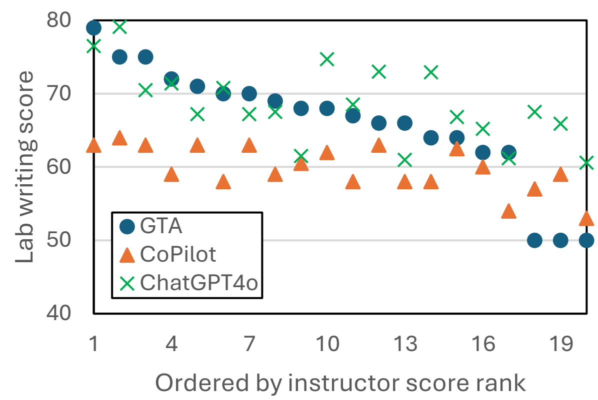
Figures 1 compare the GAI-generated scores to those assigned by human assessors (UTA, GTA, and faculty), with results ordered according to the human assessors' rankings. Assessment results of lab writing assessors and the two GAI chatbots were broadly consistent, and their correlation coefficients are computed with scores and rankings. Pearson correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two variables [12]. Scatterplots in Figure 2a show a strong positive relationship between assessor-assigned scores and GAI-generated scores. The Pearson correlation coefficients for the comparison of scores generated by GAI models and human assessors for all student samples ($n=60$) were +0.74 for CoPilot and +0.82 for ChatGPT-4o. The correlations were statistically significant ($p<0.001$), highlighting the reliability of these GAI models in aligning with human assessments. Scatterplots in Figure 2b present student sample's ranks by human assessors and GAI chatbots for each lab. Spearman correlation is a non-parametric measure of rank correlation that assesses the strength and direction of the monotonic relationship between two variables [12]. Spearman correlation coefficients were calculated to compare the rankings of scores generated by GAI models and human assessors. The coefficients were +0.79 for CoPilot and +0.76 for ChatGPT-4o, indicating a strong agreement in rank ordering between GAI-generated scores and human-assigned scores. The Spearman correlation results are also statistically significant ($p<0.001$), suggesting that GAI tools are not only compatible with human assessors in absolute scoring but also in reflecting the rank of student performance within a lab group.





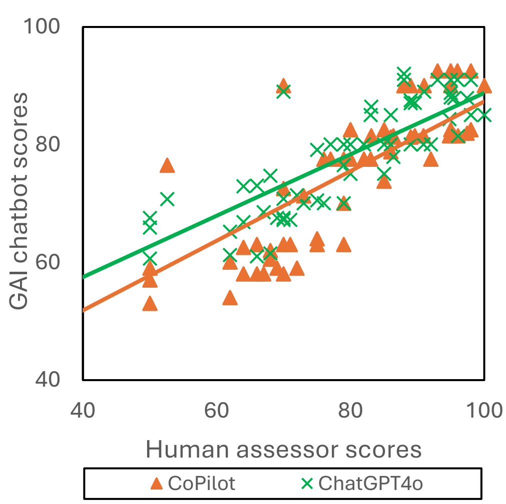
(a) ENGR 213 (Fall 2022) (b) ENGR 213 (Fall 2024)

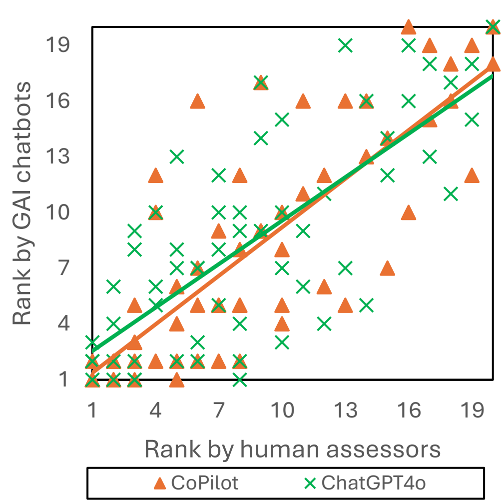




(c) MECH 309 (Fall 2016) (d) MECH 309 (Fall 2022)

Figure 1. Lab report evaluation scores by human assessors and two GAI chatbots





(a) human assessor scores vs GAI scores (b) human assessor ranks vs GAI ranks

Figure 2. Scatterplots of human assessor and GAI assessment results for all student samples (*n=60*)

The GAI tools' compatibility with human assessor assessment results depends on rubric structure and language. We used MECH 309's Fall 2022 samples to investigate discrepancies between human assessor scores and GAI-generated scores within the lab using one rubric. The lab's rubric, shown in Table A.2, had four criteria: 1) technical background, 2) tables and figures, 3) data analysis and comparisons, and 4) structure and conventions. Table 2 presents average scores and % difference for each rubric criterion. The smallest % difference was observed in Criterion 1) technical background, as this criterion requires evaluating consistent elements such as the lab's purpose, context, and technical background, making it a well-defined task. In contrast, the largest discrepancy occurred in Criterion 3) data analysis and comparisons. CoPilot and ChatGPT-4o demonstrated 20% and 21% differences compared to the human assessor, respectively, reflecting challenges in handling complex, context-dependent tasks where GAI tools are prone to factual inaccuracies. Criterion 2) tables and figures resulted in the highest difference between the two GAI chatbots: CoPilot (19%) versus ChatGPT-4o (8%). This high discrepancy may be due to CoPilot's relatively underdeveloped multimodal capability. The improved multimodal capability of ChatGPT-4o compared to earlier versions may reflect better compatibility with the TA scores in this criterion. Lastly, Criterion 4) structure and conventions, which assesses report organization, showed a 14% difference with CoPilot and a 6% difference with ChatGPT-4o, with ChatGPT-4o aligning most closely with the TA scores despite the criterion's broad nature.

Table 2. Average scores and %difference comparisons for each rubric criterion in MECH 309's 2022 samples.

	Criterion 1) Technical background		Criterion 2) Tables and Figures		Criterion 3) Data analysis and comparisons		Criterion 4) Structure and conventions	
	Ave. score	Ave. %diff	Ave. score	Ave. %diff	Ave. score	Ave. %diff	Ave. score	Ave. %diff
TA	17.6	-	16.3	-	14.5	-	17.3	-
CoPilot	16.4	9%	13.7	19%	13.3	20%	15.3	14%
ChatGPT4o	17.0	8%	15.6	8%	15.5	21%	16.3	6%

The compatibility of GAI tools with human assessment results appears to depend on the expertise of the lab writing assessors. As shown in Figure 3, Spearman correlation coefficients from the instructor (+0.82 for CoPilot and +0.83 for ChatGPT4o) are the highest, while those for UTA assessment are the lowest (+0.52 for CoPilot in 2022 and +0.46 for ChatGPT4o in 2024). Notably, MECH 309's instructor and GTA used the same rubric, and the Spearman coefficients for GTA are lower than those for the instructor. These findings suggest that the disciplinary expertise of human assessors significantly influences the compatibility of GAI-generated results with human evaluations.

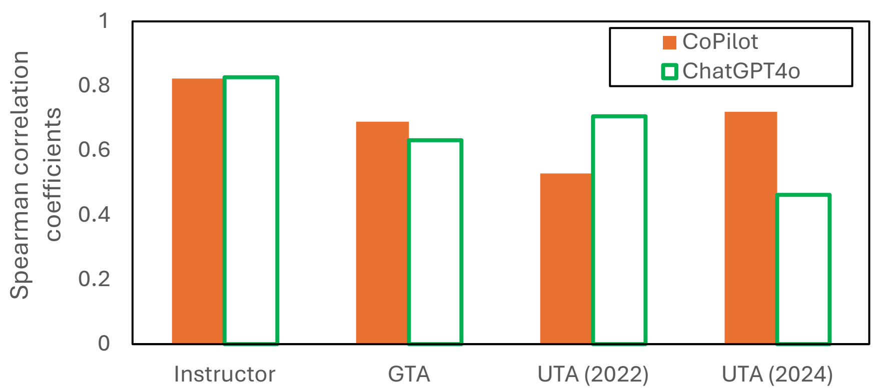


Figure 3. Spearman correlation coefficients for instructor (MECH 309's Fall 2016), GTA (MECH 309's Fall 2022), and UTAs (ENGR 213's Fall 2022, Fall 2024)

3.2.2 Intrarater reliability of GAI

We tested the two GAI chatbots' intrarater reliability, which refers to the consistency of scores or assessments given by the same evaluator across multiple instances under similar conditions. During the analysis with the MECH 309's Fall 2022 samples, we picked three student samples 1, 5, and 7 in the engineering materials lab and conducted the GAI assessment process repeatedly for samples 21 to 23, respectively. As shown in Table 3, the raw scores of the three student samples were compared between the two assessments in different sequences. CoPilot's scores of the two assessments for the same sample were identical in terms of scores and comments, word by word. However, ChatGPT's scores in Criterion 3 and 4 were lowered by 1 in the same samples during the later assessments. Also, ChatGPT's feedback became longer and slightly more detailed at the later assessment. The sample feedback is included in Table A.3 in the Appendix. This suggests that ChatGPT-4o may have less intrarater reliablility and its scores may become lower when the assessment progresses.

Table 3. Intrarater reliability test results by assessing the same samples twice in different orders.

	CoPilot				ChatGPT4o			
Criteria	1	2	3	4	1	2	3	4
Sample 1	18	15	14	17	18	17	16	17
Sample 21	18	15	14	17	18	17	16	17
Sample 5	16	14	13	15	17	16	**16**	16
Sample 22	16	14	13	15	17	16	**15**	16
Sample 7	14	12	13	14	14	12	13	**14**
Sample 23	14	12	13	14	14	12	13	**13**

3.3 Qualitative Analysis Results: Reasons for score discrepancies between the TA and the GAI chatbots

3.3.1 Student samples under-rated by GAI chatbots

We selected a few samples that were under-rated by GAI chatbots to investigate potential discrepancies compared to human assessors. The 4th ranked student sample in a 2016 MECH 309 lab received the 10th rank by both GAI chatbots, which focused more on room for improvement. ChatGPT4o's overall assessment was "Good use of data but limited discussion on results," while CoPilot's overall assessment was "The report meets expectations in most areas but has room for improvement in clarity, precision, and supporting evidence." The sample shows a few distinct characteristics compared to other lab reports with similar human assessor scores. First, it was one of the short lab reports, containing less than 5,000 words. Although it is a short report, it contains all the necessary technical elements related to the lab content. Other top 5 ranked samples contained more than 6,500 words. Second, the 4th ranked sample had five data tables and a graph containing three stress-strain curves with yield strength, ultimate tensile strength, and fracture strength values. The student added one unique graph to relate hardness and toughness of the materials tested. This graph was unique among the cohort, and the human assessor valued that the student did additional analysis.

The 9th ranked student sample in the 2016 MECH 309 lab received the 17th and 14th rank from CoPilot and ChatGPT-4o, respectively. CoPilot and ChatGPT-4o's overall comments were "The report is well-structured with a clear introduction, body, and conclusion. However, the report could benefit from more detailed explanations and a more thorough discussion of the results," and "Good visuals and structure; more in-depth data interpretation required," respectively. The human assessor scores for this sample may be over-rated, as the report omits several critical aspects of the lab. Notably, the sample lacks the construction of stress-strain curves and the accompanying discussion. Additionally, it fails to define and analyze the mechanical properties of the tested materials, which is a fundamental requirement of the lab. It may be possible that the human assessor over-valued the student's effort in comparing the lab data with reference data from external sources, despite these significant omissions.

The 3rd ranked sample by the UTA in the 2024 ENGR 213 lab received the 5th and 10th rank from CoPilot and ChatGPT-4o, respectively. CoPilot and ChatGPT-4o's overall comments were "The report is generally accurate, with clear information, consistent units, and error-free calculations. The report is well-organized, with smooth transitions, a professional tone, and consistent formatting." and "This report demonstrates a good understanding of tensile testing and data analysis. Enhancing statistical analysis, refining the discussion, and offering more specific recommendations would improve its overall quality.," respectively. Like the sample discussed above, this sample also appears to be overrated by the UTA for unknown reasons. The stress-strain diagrams are incomplete, only plotting the linear portion for determination of the modulus of elasticity, and the comparison of measured and published values is incomplete, lacking statistical analysis where appropriate. It may be that the UTA was working quickly or otherwise missed identifying these deficiencies, while the chatbot was more explicit in identifying these deficiencies in the results and analysis criterion: "Results are presented with stress vs. strain graphs for each material, but there is limited comparison to expected values and statistical analysis." This is a very nuanced and accurate statement that was simply missed by the UTA.

3.3.2 Student samples over-rated by GAI chatbots

A few over-rated student samples by GAI chatbots are analyzed to characterize the discrepancies between the human assessors and the GAI chatbots. The 12th ranked student sample in the 2022 MECH 309 lab received the 4th and 6th ranked by both GAI chatbots with a summary comment of "Detailed procedures and results; could improve on connecting results to theoretical concepts" by ChatGPT-4o. This report features well-written introduction and experimental procedures sections, which comprise nearly half of its total pages. The sample's results section includes only data tables and graphs, lacking descriptions of lab data analysis and interpretation results. Moreover, the conclusion is limited to a single sentence, making it insufficient to place this report in the top third of the samples within the lab. The 13th ranked student sample in a 2016 MECH 309 lab is ranked at 5th and 7th by CoPilot and ChatGPT-4o, respectively. The report has over 6000 words in the report with well-structured introduction, experimental procedures, results sections. This sample has two critical errors: first, it was verbose, and second, the conclusion section was too long. Even with the area of improvement, this sample could be within the top third of the samples within the lab. The human assessor might have over-focused on the errors when evaluating the sample.

3.4 Qualitative Analysis Results: Feedback provided by TAs and GAI Chatbots.

We compiled feedback from the human assessors and the two GAI chatbots for each lab report to analyze their quality. Notably, many report samples contained no feedback from the human assessors. According to our qualitative analysis, the human assessors' feedback was often either too brief or overly direct, as shown in Table 4. This may have led students to focus on surface-level error correction rather than fostering a deeper understanding of how writing instruction supports their progress. The quality of assessor comments was inconsistent in two key ways: some lacked specificity (e.g., "objective statement unclear"), while others were vague or imprecise (e.g., "add more detail"), potentially diminishing students' confidence in the assessment and hindering their writing development. Additionally, assessor comments contained typo errors (e.g., "poast" for paste), which could make students confused, undermine their confidence in the feedback, and detract from the assessment credibility. In contrast, GAI provided feedback in a range of formats and lengths for each lab sample. Table 4 shows that GAI demonstrated flexibility in its feedback styles, allowing users to adjust the level of detail, ranging from concise bullet points to comprehensive paragraphs of information. The feedback by the GAI chatbots was heavily dependent on the style and language of the assessment rubrics. They effectively addressed key rubric points in a manner that was both constructive and encouraging. GAI chatbots' feedback is related to the human assessor's. For example, the GTA's feedback

“Could have more technical background” was articulated by CoPilot as “The report introduces the objective and technical background, but the explanation is somewhat brief and lacks depth. It could benefit from a more detailed discussion of the technical background and a clearer connection to the audience’s expectations.”

Table 4. Feedback comparisons between human assessors (instructor, GTA, and UTA) and GAI chatbots

	Human assessor	Copilot in Microsoft 365	ChatGPT-4o
UTA (CE 212) in 2022	Add more detail about the request for the lab and the testing to be done.; Define sections.; Figure suggestion - remove title, center, and don’t screen shot; Copy and poast into the Word doc	The report could be better organized, with smoother transitions, a more professional tone, and consistent formatting.	The report is somewhat organized, but transitions between sections are abrupt, and the tone is inconsistent (informal in places). Formatting is cluttered, and the document lacks polish in its presentation.
GTA (MECH 309)	Could have more technical background.	The report introduces the objective and technical background, but the explanation is somewhat brief and lacks depth. It could benefit from a more detailed discussion of the technical background and a clearer connection to the audience’s expectations.	The introduction is brief and lacks detail about the significance of the properties being measured or their relevance in real-world applications. A more thorough explanation of why the material’s properties are being tested and how they relate to product design would improve this section.
Faculty (MECH 309)	Objective statement unclear; I like your data analysis skills, but your discuss is very minimal. I would like read more of your findings from discussion; Any discussion here?	• Introduction: Clearly state the objective and hypothesis. Provide sufficient background information to contextualize the experiment. • Results and Discussion: Present data clearly in tables and figures. Discuss the significance of the results, comparing them to theoretical values and previous studies. Include all relevant data, such as hardness values, and provide detailed analysis.	• Strengths: Effective use of graphs to visualize stress-strain data; logical presentation of material properties. • Areas for Improvement: Enhance the introduction to provide a more robust objective statement. Include more detailed analysis in the discussion, especially on anomalies. Refine figure labels and ensure consistent formatting throughout.

4. Discussion on the potential of GAI-assisted lab writing assessments and limitations of the research

Our quantitative and qualitative analysis results with GAI-assisted lab writing assessments are consistent with prior research [13-16], showing mixed agreement regarding the compatibility of quantitative assessments of GAI chatbots, yet improved quality of written feedback by GAI chatbots.

The lab report genre is well established, so most GAI chatbots need little guidance and can describe the genre and audience expectations accurately without additional clarification. However, one of the biggest benefits of using GAI as described here is providing it with the instructor's specific expectations, in this case via a rubric, which ensures that expectations are reiterated in the qualitative feedback. If students use the rubric effectively to craft their lab report then they would be provided feedback that is consistent in structure and content.

Once the preliminary analysis was conducted as described in the methods section, additional querying of the dataset was conducted with each chatbot, for instance to request summary comments or common errors or themes that could be useful to an instructor in providing general feedback to a class. The chatbots were also queried for good and poor examples of writing in particular sections, which is a feedback approach that many instructors use but that requires time and attention to collect during grading.

A particularly useful prompt used here was "how are the reports I have provided the same and how are they different," which results in useful statements of general features or analysis or content as well as significant differences in quality, length, tone, and even conclusions. Another is "how do the writing styles vary between the reports," which results in a list of the reports with descriptions of their tone and formality, clarity and structure, and use of technical language. ChatGPT-4o even offered specific examples from the reports, which then quoted conversational phrasing like "it's hard to say exactly" and "Overall, though." Specific examples like this are very helpful when providing general feedback to a class and were generated quickly and accurately by ChatGPT-4o. Chatbots are becoming more helpful in general, often suggesting ways they can assist or prompts that the human user might find valuable.

This study has some limitations. First, the sample size of lab samples and the participating courses were small. The study was performed in two lab courses from two engineering majors (civil and mechanical) at two institutions, limiting the generalizability of our study's results. Furthermore, we did not investigate systematically about the effect of GAI prompts on the assessment outcomes, which is a future research topic.

5. Conclusion

After conducting student lab report sample evaluations (*student sample n = 60*) using Copilot and ChatGPT-4o in strength of materials lab and engineering materials lab courses, we have drawn the following conclusions.

1. GAI-generated scores showed strong positive relationships with those assigned by human assessors, with Pearson correlation coefficients of +0.74 for CoPilot and +0.82 for ChatGPT-4o, and Spearman correlation coefficients of +0.79 and +0.76, respectively, reflecting good agreement in both scores and rankings.
2. The assessment outputs from GAI chatbots were highly influenced by the style and language of the rubrics. GAI scores aligned more closely with assessor scores in rubric dimensions focused on the introduction and lab writing conventions, while showing lower compatibility in dimensions related to data analysis and discussion writing.
3. Both GAI chatbots demonstrated nearly perfect interrater reliability.
4. GAI feedback effectively addresses key aspects, aligning closely with the rubric points emphasized by the assessors in a constructive and supportive manner.
5. GAI prompting can produce useful summaries, themes, example text, and general feedback, useful for sharing with a whole class.

6. References:

[1] L. D. Feisel and A. J. Rosa, "The role of the laboratory in undergraduate engineering education," *Journal of Engineering Education*, vol. 94, no. 1, pp. 121-130, 2005.

[2] C. D. Troy, R. R. Essig, B. K. Jesiek, J. Boyd, and N. T. Buswell, "Writing to learn engineering: Identifying effective techniques for the integration of written communication into engineering classes and curricula (NSF RIGEE project)," in *the Proceedings of the 2014 American Society for Engineering Education Annual Conference*, 2014.

[3] J. Parkinson, “The student laboratory report genre: A genre analysis,” *English for Specific Purposes*, vol. 45, pp. 1-13, 2017.

[4] ABET, “Criteria for accrediting engineering programs, 2024-2025,” 2024. [Online]. Available: https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2024-2025/. [Accessed: Jan. 1, 2025].

[5] D. Kim, R. Sandry, A. Taran, A. Yurov, and J. Lynch, “Investigating engineering laboratory course assignments and assessments across four institutions and a case study on their impact on students’ lab report writing,” in *Proc. ASEE Annu. Conf. & Expo.*, Baltimore, MD, USA, 2023.

[6] H. Wang, A. Dang, Z. Wu, and S. Mac, “Generative AI in higher education: Seeing ChatGPT through universities’ policies, resources, and guidelines,” Comput. Educ.: Artif. Intell., vol. 7, no. 1, 2024, doi: 10.1016/j.caeai.2024.100326.

[7] Wang T.H., Jack. Wang JTH (2023) Is the laboratory report dead? AI and ChatGPT, Microbiology Australia, 44(3), 144–148. doi:10.1071/MA23042

[8] Zhou, X., Hu, W., Lei, Z., Liu, GP. (2024). Implementation and Evaluation of an Automatic Scoring System for Experimental Reports Based on ChatGPT. In: Auer, M.E., Cukierman, U.R., Vendrell Vidal, E., Tovar Caro, E. (eds) Towards a Hybrid, Flexible and Socially Engaged Higher Education. ICL 2023. Lecture Notes in Networks and Systems, vol 899. Springer, Cham. https://doi.org/10.1007/978-3-031-51979-6_45

[9] WSU Information Techniology Services, “Microsoft Copilot for Web, WSU’s new AI Tool,” 2024. [Online]. Available: https://its.wsu.edu/2024/08/12/microsoft-copilot-for-web-wsus-new-ai-tool/ [Accessed: Jan. 1, 2025].

[10] R.J. Wolcott, “New guidance available on AI and data stewardship,” 2023 [Online]. Available: https://news.wsu.edu/news/2023/06/21/new-guidance-available-on-artificial-intelligence-and-data-stewardship/ [Accessed: Jan. 1, 2025].

[11] WSU Information Technology Services, “WSU Information Security Roles, Responsibilities, and Definitions,“ [Online]. Available: https://policies.wsu.edu/prf/index/manuals/business-policies-and-procedures-manual/bppm-87-01/ [Accessed: Jan. 1, 2025].

[12] S. Boslaugh, Statistics in a Nutshell: A Desktop Quick Reference, 2nd Ed., O'Reilly Media, 2012.

[13] I. T. Awidi, "Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool," *Computers and Education: Artificial Intelligence*, vol. 6, no, 3, 100226, 2024.

[14] J. Escalante, A. Pack, and A. Barrett, "AI-generated feedback on writing: insights into efficacy and ENL student preference," *International Journal of Educational Technology in Higher Education*, vol 20, article no. 57, 2023.

[15] M. Zuckerman, R. Flood, R. J. B. Tan, N. Kelp, D. J. Ecker, J. Menke, and T. Lockspeiser, "ChatGPT for assessment writing," *Medical Teacher*, vol. 45, no. 11, pp. 1224–1227, 2023.

[16] D. Foung, L. Lin, and J. Chen, "Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices," *Computers and Education: Artificial Intelligence*, vol. 6, 100250, 2024.

8. Appendix

Table A.1 Rubric for the Strength of Materials labs

Scoring system: 1.1 pts = exceeds expectations; 1 pts = acceptable; 0.5 pts = needs work; 0 pts = missing/incomplete	
Introduction (technical) Reiterate request Purpose/goals of the document Summary of contents and organization of the document	**Introduction (writing)** Tone according to audience (peers, professor, professional, client) Avoid conversational language Focus on technical/professional presentation Remain objective (fact vs opinion)
Methods (technical) Time and location of testing Theories/topics explored List of equipment used Test configuration Steps involved in testing Steps involved in analysis	**Methods (writing)** Descriptive narrative Enumerated list of steps Pictures or schematics of test setup Reference sources, like lab manuals or test standards, can provide details, be cited, and be reiterated or not

Organization, if multiple tests	
Results and Analysis (technical) No raw data. Data synthesized for reporting into Tables and Figures. Comparisons to expected/published/design values. Statistical/comparative analysis.	**Results and Analysis (writing)** Focus is effective, complete and concise presentation of results Tables (label/title above) Figures (label/title below) Plots, photographs, sketches with annotation Tables and figures referenced in text (e.g. see Table 1 for…) Use summary graphics for ease of interpretation
Discussion (technical) Summary of expected results Explanation of unexpected results Interpretation of results Comparisons made, referencing specific data Sources and impact of errors (systematic vs random error) Suggestions to improve results Future work	**Discussion (writing)** Be specific Avoid generalizations Reference primary sources (results values, trends, analysis) Reference secondary sources (theory, published sources)
Conclusions (technical) Summary of methods. Summary of generalized results. Note specific outliers. Relevance of work. Potential applications	**Conclusions (writing)** Be as general as possible, given evidence provided. Avoid overstating results. Avoid being overly critical of theory, sources, or errors that occurred.
References and Appendices (technical) Author, title, publication, date. Raw data, raw data tables, hand or example calculations, spreadsheet calculations	**References and Appendices (writing)** Formatted according to a particular and consistent style guide (APA, IEEE, Chicago, ASCE, etc.). Cited in the text of the document. Organized and labelled clearly. Not formally presented, but organized. Cited in the body of the report. Page numbers not necessary.
Overall (technical) Technical information is clear and accurate. There is a response to all requests. Units are present on all values requiring them. Calculations and analysis are error free.	**Overall (writing)** Organization: Engineering lab reports are organized documents with a clear introduction, body, and conclusion. Structure: The body includes methods, results, and discussion (IMRADC). Further subsections may be necessary depending on the complexity of the experiment. Transitions between sections are smooth, not abrupt.

	Tone: The tone should be professional and technical, making use of industry-relevant terminology and methods. Formatting should be consistent, uncluttered, and unambiguous. Audience should be addressed specifically, whether it is your peers, a professor, or a hypothetical or real client.

Table A.2 Rubric for the Engineering Materials lab (Tensile testing lab)

	Exemplary (20)	**Meet Expectation (15)**	**Need improvement (10)**
Lab objective 3. Introduce the objective and technical background in the report.	Analyze the technical audience's expectations and the context for the lab report. Provide purpose, context, and technical background proficiently.	The writer's understanding of the context and audience supports a generally successful report. Attention to purpose, context, and technical background are generally appropriate, with some lapses.	Little to no awareness of the audience's needs and the context. The purpose, context, and technical background provided are too basic or inadequate.
Lab objective 4. Use table(s) and plot(s) to determine mechanical properties.	The writer uses effective strategies to use graphic/table forms when communicating lab data/results. Graphic/table forms are stand-alone and professional. They contain all required features to follow standard conventions and include useful captions. Figures, tables, and illustrations are correctly and usefully labeled.	When communicating lab data/results, strategies using graphic/table forms were generally appropriate, with lapses. Graphic/table forms are generally appropriate; however, they contain minor errors. Figures, tables, and illustrative materials are labeled.	The writer fails to use effective graphic/table forms when communicating lab data/results. Graphic/table forms contain little or no required features. Multiple errors are found in the graphics/tables. Figures, tables, and illustrative materials are not labeled.
Lab objective 5. Compare the data among test materials to verify engineering principles.	The writer analyzes lab data using appropriate methods (comparative studies for this lab) professionally. The writer draws significant technical knowledge from an in-depth analysis consistent with the complexity of the experimentation.	Lab data analysis is generally appropriate; however, the analysis methods have some lapses, or the analysis results of lab data are not well aligned with the complexity of the experimentation.	The writer fails to analyze lab data. The writer's lab data analysis is limited, and the data analysis methods have significant errors. Sometimes, the writer may "let the data do the talking."
Lab objective 6. Provide a research paper by establishing solid and consistent control of conventions	The writer provides a purposeful structure that clearly articulates the experiment's purpose as a whole document. The report has a well-structured introduction, body, and conclusion. Each of these three parts (intro, body, conclusion) well functions in one report.	The writer provides a structure (intro, body, and conclusion) generally appropriate for a lab report as a whole document. Generally, each part (intro, body, conclusion) relates to the primary purpose of the report.	The report's structure (intro, body, conclusion) may be inappropriate, incomplete, or missing. The writer made significant errors in the functions of these three parts (intro, body, conclusion).

for a technical audience.			

Table A.3 Intrarater testing results from ChatGPT-4o

Sample 5	Sample 22 (A repeat of Sample 5)
Here's the assessment of the fifth lab report based on your rubric: Lab Objective & Technical Background (17/20) The introduction explains the purpose of the lab clearly, focusing on tensile strength and toughness. While it outlines the importance of these properties, it could benefit from a more detailed explanation of their relevance to engineering applications, particularly in relation to the specific materials being tested. Use of Tables and Plots (16/20) The report provides stress-strain curves and a table comparing hardness values. However, the figures lack detailed captions and explanations, making it harder to fully understand the results at a glance. More emphasis on explaining what the figures reveal about the materials' behaviors would improve clarity. Comparison of Data & Verification of Engineering Principles (16/20) The comparison between the materials' tensile strength and toughness is solid, and the report correctly highlights the trade-off between these properties. However, the analysis could be more in-depth, particularly in discussing how the experimental results align with theoretical expectations and manufacturer specifications. Technical Writing & Structure (16/20) The report is well-organized, with clear sections for methods, results, and conclusions. However, there are some grammatical issues and areas where the writing could be refined for better flow and clarity. The transition between the results and discussion sections could be smoother, and more direct links between the data and conclusions would strengthen the report. Total Score: 65/80 This is a good report that covers the essential content but could be improved by adding more detailed analysis and refining the data presentation. Enhanced clarity in writing and deeper exploration of the engineering principles would elevate the overall quality.	Here's the assessment of the twelfth lab report based on your rubric: Lab Objective & Technical Background (17/20) The introduction clearly outlines the purpose of the lab, focusing on tensile strength and toughness of Aluminum 6061-T6, Delrin, and Nylon. However, the discussion could benefit from more context regarding real-world applications and the significance of these tests in engineering design. Use of Tables and Plots (16/20) The report includes several useful tables and stress-strain graphs for each material. However, the presentation could be improved with more detailed captions and clearer explanations of how the graphs support the conclusions. The hardness values are presented well, but the report would benefit from better integration of data into the discussion. Comparison of Data & Verification of Engineering Principles (15/20) The comparison between materials is informative, but the analysis lacks depth. More discussion of how the experimental results align with theoretical expectations or manufacturer data would enhance this section. The report touches on the inverse relationship between toughness and yield strength but does not deeply analyze the implications of these findings. Technical Writing & Structure (16/20) The writing is mostly clear, but there are some minor grammatical issues and awkward phrasings that detract from readability. The structure is logical, though transitions between sections could be smoother. The conclusions drawn are generally sound but could be more thoroughly supported by the experimental data. Total Score: 64/80 This report presents the necessary information but would benefit from more in-depth analysis and clearer presentation of data. Improved data interpretation, particularly in relation to engineering principles, and more fluid technical writing would elevate the overall quality of the report.

Table A.4 Feedback from GAI tools for one lab report sample in a Strength of Materials lab.

Copilot
Introduction **Technical: 16** **The introduction provides a brief overview of the testing process and the materials tested but lacks a detailed summary of the document's contents and organization.** **Writing: 15** **The tone is somewhat conversational and less formal than expected for a technical report.** **Methods** **Technical: 17** **The methods section describes the testing process and equipment used but lacks detailed steps and references to figures or sources.** **Writing: 16** **The narrative is descriptive but could benefit from a more structured, enumerated list of steps and inclusion of schematics or pictures.** **Results and Analysis** **Technical: 16** **Data is presented with some synthesis into figures, but there is limited comparison to expected values and statistical analysis.** **Writing: 15** **Results are presented, but the clarity and conciseness could be improved. Figures are included but not well-labeled or referenced in the text.** **Discussion** **Technical: 16** **The discussion interprets some results and mentions potential errors but lacks detailed comparisons and references to specific data.** **Writing: 15** **The writing is somewhat general and could be more specific, with better references to primary and secondary sources.** **Conclusions** **Technical: 16**

The conclusions summarize the findings but are somewhat general and lack detailed discussion of relevance and potential applications.

Writing: 15

The conclusions are general and could avoid overstating results and being overly critical.

References and Appendices

Technical: 16

References are mentioned but not detailed or formatted according to a specific style guide.

Writing: 15

References are not clearly cited in the text, and the organization could be improved.

Overall

Technical: 16

The report is generally accurate but could benefit from clearer information, consistent units, and error-free calculations.

Writing: 15

The report could be better organized, with smoother transitions, a more professional tone, and consistent formatting.

Total Score: 199 out of 260

If you need further details or have more reports to assess, feel free to share!

76.5%

ChatGPT-4o

Assessment of Sample02 (Based on Provided Rubric)

1. Introduction
Score: 18
The introduction clearly outlines the purpose, materials, and objectives of the experiment. However, it could better connect the experiment to broader engineering concepts or practical applications.

2. Methods
Score: 17
The methods section is detailed and well-organized, listing the equipment and procedures. Including schematics or photos of the setup would enhance clarity.

3. Results and Analysis
Score: 19

Results are well-documented with clear graphs, tables, and explanations. Calculations are accurate, and data are compared to design values effectively. More focus on error analysis could improve this section.

4. Discussion
Score: 18
The discussion links results to material behavior (e.g., ductility, brittleness) and uses external sources (e.g., MatWeb) effectively. It could delve deeper into the reasons for discrepancies between measured and expected values.

5. Conclusion
Score: 17
The conclusion summarizes key findings well but could be more specific about the implications of the results and provide concrete recommendations for future work.

6. References and Appendices
Score: 16
References are used, but citation formatting is inconsistent. The appendices include detailed data but lack organization for easier navigation.

7. Overall
Score: 18
The report is well-structured, with smooth transitions and professional tone. Formatting is generally clear, though minor inconsistencies in headings and labeling detract slightly from presentation quality.

Total Assessment Summary

Overall Score: 123/140
This report demonstrates a strong understanding of the lab experiment and results interpretation. Improvements in visual aids, citation formatting, and deeper analysis of discrepancies would elevate it further.

Let me know if you'd like a more detailed review of any specific section!
87.9%