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 #47476

Using Generative AI Prompts for Summative and Formative Feedback on Engineering Writing Assignments

Dr. Stephany Coffman-Wolph, Ohio Northern University

Dr. Stephany Coffman-Wolph is an Assistant Professor at Ohio Northern University in the Department of Electrical, Computer Engineering, and Computer Science (ECCS). Previously, she worked at The University of Texas at Austin and West Virginia University Institute of Technology (WVU Tech). She is actively involved in community outreach with a goal of increasing the number of women in STEM and creating effective methods for introducing young children to CS concepts and topics. Dr. Coffman-Wolph's research interests include: Artificial Intelligence, Fuzzy Logic, Software Engineering, STEM Education, and Diversity and Inclusion within STEM.

Dr. Abigail Clark, Ohio Northern University

Abigail Clark is an assistant professor of mechanical engineering at Ohio Northern University. She holds a PhD in Engineering Education from The Ohio State University. She also holds degrees in Mechanical Engineering from Ohio State and Ohio Northern University. Prior to her time at OSU, she worked at Battelle Memorial Institute in Columbus, Ohio. Her research interests include pre-college engineering education, informal engineering education, and identity development.

Dr. J. Blake Hylton, Ohio Northern University

Dr. Hylton is an Assistant Professor of Mechanical Engineering and Coordinator of the First-Year Engineering experience for the T.J. Smull College of Engineering at Ohio Northern University. He previously completed his graduate studies in Mechanical Engin

Dr. Bryan Alan Lutz, Ohio Northern University

Bryan A. Lutz (he/they) is an Assistant Professor of Rhetoric and Composition at Ohio Northern University. His research examines how activists, advocates, and public and private organizations use technology and writing to define an identity, argue, and act to solve (or make) problems. He teaches organizational communication, academic writing, and professional writing courses. Dr. Lutz has published with the journals Computers and Composition Online, The Journal of Critical Thought and Praxis, The Journal of Interactive Technology and Pedagogy, and the Journal of Contemporary Rhetoric, as well as with the academic publishers Pearson, Routledge, and the ACM Digital Library. Dr. Lutz also serves as chair of the Ohio Council of Writing Program Administrators, as a board member for Grey Matter Media, and as a communications consultant with non-profit groups and private businesses.

Gabriel Mott, Ohio Northern University

Using Generative AI prompts for summative and formative feedback on engineering writing assignments

Introduction

Purpose

The practice and evaluation of technical writing in an engineering course context has long been a subject of discussion. While recognized as valuable to student development, there is a tension of time and attention between traditional technical content and technical writing content, both on the side of the students, who have only so much bandwidth to dedicate to a course, and the instructor, who necessarily must minimize the assessment burden wherever possible and has only limited lecture time available. Technical writing most commonly makes its way into the engineering coursework through the avenue of laboratory courses and cross-disciplinary design courses, such as capstone and first-year engineering. In the case of first-year engineering especially, the scope and scale of those courses creates even greater time pressure on the evaluation of writing content - a pressure which may unfortunately be relieved by reducing technical writing evaluation to a superficial or cursory treatment.

This work explores the efficacy of AI tools as an alternative means of alleviating this pressure. Born of a desire to improve technical writing feedback and evaluation without dramatically increasing the assessment burden on course instructors, evaluation of a first-year engineering technical writing assignment is explored. The study team brought together two engineering faculty members who are familiar with the course and assignment under study, an English faculty member with expertise in technical writing, a Computer Science faculty member familiar with the use of various AI tools, and a student researcher familiar with both technical writing conventions and statistical analysis.

Background

There is a growing body of literature on using AI as a tool supporting assessment. Working at Aalborg University, Lindsay and Jahromi [1] explored using Natural Language Process (NLP) to assign pass/fail grades to a 2000-word reflective essay. The researchers were motivated to use AI because of the labor-intensive nature of grading the essays, which they calculated as "well over 500 hours of pass/fail summative assessment work within a very short timeframe" for their 1500 students who completed the task. Of the 1500 submissions, Lindsay and Jahromi used 80% of the data as a training set and 20% as test data. They used the NLTK Python library to code classifiers based on the desired sections of the essay before using a Convolutional Neural Network made up of weighted connections between nodes, and adjusting the weightings between the nodes to train the classifier. The researchers found an inverse relationship between the amount of data used to train their system and the likelihood of false positives and false negatives, "Reducing the training set reduces the overall human effort required, even considering the remarking of false negatives; but it does so at the cost of the introduction of false positives as a consequence." Suresh et al. [2] likewise explored using an AI system and Graph-based techniques to automate the process of evaluating student writing. Collected from Kaggle's Automated Student Assessment Prize (ASAP), they likewise used 80% of the data for training reserving 20% as test data before performing a Kappa test to determine if the feedback was consistent. They concluded that the AI model "has the potential to accurately grade essays and

provide valuable feedback to students," but this conclusion was supported by comparing AI's feedback to itself, not by a standard of assessment or feedback used by humans.

Studies are emerging that demonstrate the efficacy of AI as support for teaching and grading practices. Furze et al. [3] conducted a pilot study at the British University Vietnam (BUV) exploring how AI could support students and teachers in enacting their university's curriculum. Motivated by a 70% increase in "instances of students' academic misconduct with AI," they piloted an Artificial Intelligence Assessment Scale (AIAS) as a "flexible framework for incorporating GenAI into educational assessments." *Assessment*, in this case, refers to when students produce work that thinks critically about a given topic. The AIAS consists of five levels, No AI (such as hand-written exams or class discussions), AI-assisted idea (such as brainstorming for essays or converting class notes into conceptual outlines), AI-assisted editing (prompting AI to revise written work), AI Task Completion with Human Evaluation (prompt AI to sort data and have a human interpret the data), and FULL AI (exploring prompts or writing human responses to fully AI-generated content). Using the model as a guide, faculty would create lessons across the five assessment levels to "maximize learning opportunities while reducing instances of academic misconduct." Furze et al. found a reduction in academic misconduct cases related to GenAI, while noting significant increases in student attainment across the university and in individual module passing rates. Zhao et al. saw similar success using AI as an assessment tool for teachers. The researchers recruited 279 students Chinese students in Grades 7 or 8 who had more than five years of experience writing in English, and tasked students with a picture-cued writing test that was assessed for two main qualities: demonstrating a suitable level of vocabulary to describe the picture while also describing a wide variety of visual content (ex. a sports scene, a nature scene, a daily life scene, etc.). The research team included two K12 domain experts to develop a rubric that would be used for six human reviewers and the AI before calculating the results using six measures. Zhao et al. [4] found that the AI "can grade student responses for picture-cued tasks as fairly as human raters," though they list several discrepancies where the AI judged students' writing based on enumerated details rather than interpretations of the relationship between humans and objects in each pictured scene, which were valued more by the human graders.

Limitations and Scope

This work is best understood in context of its limitations. First, AI tools, such as ChatGPT and Copilot are relatively new, and constantly evolving in their abilities. Secondly, the sampled student work represented a small portion of available data and thus is not representative of the whole set. Finally, while every effort was made to ensure that the evaluators were consistent with the application of the rubric, it is simply not realistic to expect people with varying expertise to be completely consistent. While these limitations were important to acknowledge, they do not limit the importance of this work. We see potential not only for optimizing writing support but also for fostering student-involved negotiations on how AI can aid the writing process.

Methods

Context

This study occurred at a small, private institution, located in the rural midwestern United States. Ohio Northern University (ONU) includes a well-established engineering college, which houses six majors (Civil and Environmental engineering, Computer Engineering, Computer Science,

Electrical Engineering, Engineering Education, and Mechanical Engineering). The college has an enrollment of approximately 700 students. ONU utilizes a common first year approach [5] via a two-semester course sequence (Foundations of Design 1 and 2). All students are required to take the first semester course, and all students, except computer science majors, take the second semester course. These courses focus on building foundational engineering skills, including introduction to the engineering design process via hands-on projects, skills such as computer-aided modeling, teamwork, technical communication, and others.

One long-standing assignment in this course is the "One-Minute Engineer", which occurs during the first semester course [6], [7]. This assignment is part of the technical communication content in the FYE curriculum. In this assignment, students first identify a goal within the United Nations' Sustainable Development goals [8] and identify an engineering-connected topic within that goal. For example, a student may select the goal of "Zero Hunger" and the topic of "Genetically Modified Organisms" within that goal. Once approved by the instructor, the students research their topic and write a one-to-two-page memo regarding the topic, get feedback from the campus writing center, revise their memo and give a one-minute presentation to their classmates regarding their topic. Following their presentation, students evaluate their presentation and write a reflection focused on both their writing process and presentation. The current assessment approach for the OME assignment primarily focuses on the technical aspects of the report, with an admittedly limited and superficial evaluation of the writing components.

Rubric

Given that the existing evaluation methods did not sufficiently capture the writing aspects desired, it was necessary to develop a new rubric built-to-purpose for this study and, ideally, to be deployed in future iterations of the assignment. The full rubric is reported in appendix 1. Rubric development was informed by time-honored theories in rhetoric and composition and by the context of our small liberal arts university. Memos are a well-established genre of professional communication and as such, memos can be expected to follow established conventions and moves. Conventions describe recurring patterns of organization, such as an introduction, body, and conclusion, as well as features such as including a to, from, subject line, and date. Moves are clauses or sentences that perform a particular communicative function. For example, an introduction paragraph typically includes a sentence that refers to the document itself and explains its purpose. Ex. This memo outlines a new procedure for distinguishing counterfeit currency from legal tender.

John Swales' [9] developed the concept of moves by combining rhetorical analysis with corpus linguistics to identify common patterns in academic and professional writing. Swales [10] later defined moves as an essential aspect of genres in academic and research settings. Expanding on the work, Bhatia [11] extended Swales' model to study professional writing genres like corporate disclosure documents, promotional texts, and legal cases, determining that defining a topic, a purpose, and a main point as essential moves of professional introductions. Later, Peacock [12] investigated the move structure of research article introductions across seven disciplines - physics, biology, language and linguistics, environmental science, business, law, and public and social administration and found three common moves across the abstracts of these seven disciplines. Such moves can be taught to novice writers of a particular genre [11], [13] even when English is the learner's second language [14] or when teaching English abroad [15].

Another useful concept when teaching technical and scientific writing is plain style. Contrary to the grand style of writing for professional oration, which utilizes rhythm, parallelism, grandiose phrases and vocabulary, and abundant similes and metaphors, plain style stresses “precise and concise sentences in audience-appropriate vocabulary” as well as “the importance of rhetorical arrangement and good document design” [16, p. 285]. To achieve concision, plain style asks for writing that uses few and precise verbs, avoiding colloquialisms, using metaphor rarely when communicating to layperson readers, and limiting modifiers such as adjectives and adverbs. To second what level of vocabulary is appropriate, the concept of fluency foregrounds how specialized vocabulary can clearly and economically communicate specialized knowledge between authors and readers sharing a vocation in science and tech. This does not mean relying heavily on jargon, but strategically utilizing “strong vocabulary and longer sentence constructions” when writing is a component of problem-solving in science and tech [17].

At our small liberal arts university, students may take specialty courses in technical and science writing offered by the English department. These courses teach genres like emails, memos, proposals, and reports, which are evaluated using rubrics informed by Swales and his contemporaries. The English faculty member on our research team reviewed Engineering’s OME Memo assignment and existing rubric to develop a rubric with specific, measurable criteria that could be applied both by humans and by AI to evaluate not only the technical content but also the writing itself. In other words, the concepts used for teaching how to write effective memos now serve as a specific measure of evaluation within the rubric.

At the bottom of the rubric, there are two headings that solicit faculty to list the aptitudes and opportunities for improvement. This is a space for summative feedback to the student on what was done well and what can be done for improvement. This is where faculty members can provide around two to three suggestions that would significantly improve the writing to better meet the standard defined by the rubric. This practice is informed by the concept of Minimal marking, developed by Richard Haswell and Nancy Sommers. Sommers [18] emphasizes respecting students’ writing and writing process by taking time to compliment them on what was done well before providing critiques. Sommers also emphasizes that students can be overwhelmed by correcting every error that is on the page. Instead, minimal marking is an efficient and effective strategy for providing feedback in support of students’ writing development in a course and across curricula. According to Haswell [19], there are a few key components of the approach. Instructors should focus comments on more substantial writing issues rather than aggregating surface-level mistakes. The rubric needs to perform the work of evaluation and as such, the instructor should not spend their time justifying the rubric scores in their summative feedback. Instead, students should be encouraged by the rubric to correct their own errors, and research has shown that students typically correct 60-70% of their own errors when using this system. Minimal marking helps students master "threshold errors,” those they are close to competency on while teachers can allocate more time to reinforcing learning through successful problem-solving. By implementing minimal marking, teachers can provide more efficient feedback while simultaneously developing students' self-editing skills and focusing attention on higher-level writing concerns.

It is difficult to eliminate ambiguity when assessing writing and providing feedback. Moreover, reducing writing to conventions and moves will not recognize when a student may have thought creatively with their assignment and achieved an unforeseen rhetorical effect. Caution should be used when defining such criteria for students so they do not view writing tasks as simply "built" without any creative license or necessary variability based on the unique combinations of author, audience, and context. But our measures provided an ideal that was understandable between our faculty member and student who are proficient in the study of writing, our engineering faculty who are specialized in their respective fields, and our AI that knows only the input we can provide in our prompts.

Human Scoring Approach

We began by creating a norming procedure that guided all human reviewers through the evaluation process. Using Google Docs, a separate folder was created that made 40 anonymized memos available to all parties. The writing evaluators discussed the rubric criteria in detail and drafted a coding scheme using Google Sheets that would define the different Likert scores possible on the rubric. The writing evaluators then performed a norming session where they evaluated three memos independently and discussed scores and feedback. When necessary, the writing evaluators revised the scheme to account for variability in the sample memos. Once substantial agreement was reached, the two Engineering faculty also met for a norming session guided by the coding spreadsheet developed by the writing evaluators. Then, all four human evaluators independently evaluated the 40 memos and provided feedback as if they were planning to submit the feedback directly to students. Once the memos were evaluated, they were made available for statistical calculation and analysis. All storage of and interactions with student material followed IRB approved procedures.

AI Tools

OpenAI's ChatGPT, a deep learning application, was trained on the Microsoft Azure AI supercomputers [20]. It is what is referred to as a large "multimodal model" that accepts not only text inputs but also images. However, the system emits text outputs. The concept behind ChatGPT and other systems is to mimic human-level performance on various tasks. The latest version, ChatGPT 4, is more stable, creative, and able to be more nuanced in the interactions between human and the AI system [20] than previous versions. Despite various upgrades, ChatGPT 4 still has the limitation of potentially "hallucinating" and suffers from reasoning errors [20] - thus not fully reliable. Additionally, training data mostly ends in 2021 and does not incorporate learning [20], thus making ChatGPT 4 inaccurate about current events.

Microsoft Copilot is another "AI companion" and also a "multimodal model" that also accepts both images and text while producing text outputs [21]. Of course, Microsoft copilot utilizes the Bing search service which provides (1) current information available on the web and (2) verifiable citations [21]. Microsoft Copilot follows Microsoft's "AI Principles" [21] and "Responsible AI Standard" [22] - which monitors misuse of Copilot and "identify, measure, and mitigate potential risks" [22] associated with the Copilot product. The Copilot was built on OpenAI's ChatGPT models, but, though Copilot does use Azure OpenAI service, the two systems do not share any information [23].

Both products have numerous similarities. Both are large language model bots trained to hold conversations, answer questions, and provide feedback to the users [24]. The key difference is the purpose. ChatGPT was designed to be more general whereas Copilot was designed to increase work productivity and to interact with various Microsoft tools (Bing, the Office Suite, Office 360) [24].

Analysis Methods

We develop a validation approach to test the strength of our rubric, and likewise the strength of AI tools. A sample of 40 blinded memos from the 2023-2024 academic year were scored according to our rubric by eight total "evaluators." Four were human and four were AI. The human evaluators consisted of two engineering experts (Engineering A, Engineering B) and two language experts (Language A, Language B). Engineering B contains the first 20 of 40 memos.

ChatGPT and Microsoft Copilot were both run twice on the entire set of memos, providing four total reviewers (OpenAI A, OpenAI B, Copilot A, Copilot B). Both models had their settings adjusted to avoid the effects of local memory when reviewing memos for the second time. We seek to identify the intra-evaluator (Engineering A v. Engineering B), inter-evaluator (Engineering v. Language v. OpenAI v. Copilot) and inter-type (Human v. AI) grading reliability arrived at by our rubric.

We begin intra-evaluator assessment by calculating the individual score differences for every criterion in every memo (always A - B). Second, we calculate the total score for each memo per evaluator, the net difference (the sum of all differences) for each memo per evaluator, and the total difference (the sum of the absolute value of differences) for each memo per evaluator. Third, we test the intra-evaluator correlation using a simple linear model. Significance is taken from the associated F-statistic, not from the slope coefficient. Finally, we calculate the Intraclass Correlation Coefficient (ICC) using the *irr* package for interrater reliability and agreement in R. Significance in these metrics indicates some form of similarity/correlation between the scoring behavior of the evaluators.

One limitation to this approach is that the human evaluators lend themselves to a more robust post-hoc analysis than the AI evaluators. Individual human evaluators may be inherently more generous or conservative in their grading. It is therefore possible that a comparable relative evaluation of the memos provides misleadingly different results (although our rubric has been designed to mitigate these issues). This is not the case with the AI evaluators; OpenAI A and OpenAI B are in no way different large language models, and there is no relationship between the "A" test of Memo 1, Memo 2, etc. For this reason, we calculate significance in grading differences using a paired t-test for the human evaluators only. Here, significance indicates a difference between the scoring behavior of the evaluators.

We begin inter-evaluator assessment by calculating a single score set for each evaluator. In most cases, we average the values of the "A" and "B" evaluators for each criterion per memo. For memos 21-40, the aggregate "Engineering" score is deferred to Engineering A. Statistical review is then conducted by repeating the above methods for each evaluator pair. Our inter-type assessment is conducted identically to the inter-evaluator assessment, with aggregate type scores being the average of the type's two aggregate evaluator scores.

Results and Discussion

Quantitative

The results are divided into two sections - intra-evaluator and inter-evaluator analysis. The first section explores the consistency of evaluators within a given type (language, engineering, AI) and inter-evaluator analysis explores the various combinations of types, averaging scores within a given type group. Full data tables are included in appendix 2 while summary tables are reported in the following sections for ease of interpretation. As a general criteria, statistical significance is defined as $p<0.05$ and technical significance is defined as a difference of at least 1.5 on a given rubric row (out of 10 points). This basis guided what aspects of the dataset warranted further reflection and discussion. In discussing these results, human reviewers are generally treated as the "control," assumed to be closer to the "true" score of the memo than the AI whenever a significant difference between the two types arises.

Intra-Evaluator

Across most measures, the Engineering and Language evaluators were both less internally consistent than either of the AI evaluators. Exceptions include the Header score, on which OpenAI was much less consistent than either of the human evaluators. Language evaluators were slightly more consistent than the Engineering evaluators, and Copilot was generally more consistent than OpenAI.

For Language, the differences between the scores for the Introduction, Topic Sentences, Quantitative Information, Sources, Figure, Header, and Grammar criteria were statistically significant. None of these were technically significant, although a few criteria were technically significant while lacking statistical significance. These can be seen in Table 4 in appendix 2. Engineering had a mostly different set of statistically significant criteria differences: Topic Sentences, Figure, Visuals, Precision, Definition, and End Matter. Of these, Figure, Visuals, Precision, and Definition are technically significant. Engineering also saw a statistically significant difference in total scores.

Linear Model (LM) and Intraclass Correlation Coefficient (ICC) calculations provide similar results. Formatting and Header values for these tests could not be validly estimated for Engineering because all 40 memos received scores of 10 from one of the evaluators. For all four evaluators, both tests have statistically significant correlations for most criteria (OpenAI's ICC scores are evenly split 9-9 for significance). These can be seen Table 5 in appendix 2. Only a few criteria have consistent technical significance in addition, however. These are shown in Table 1, below.

Table 1. Selected Linear Model (LM) Adjusted R-Square and Intraclass Correlation Coefficient (ICC):

Criterion	Language		Engineering		OpenAI		Copilot	
	LM	ICC	LM	ICC	LM	ICC	LM	ICC
Body: SDG	0.347***	0.595***	0.168**	0.45**	0.421***	0.662***	0.428***	0.654***
Body: Topic Sentences	0.319***	0.454***	0.433***	0.536***	0.0927**	0.344**	0.507***	0.705***
Body: Quantitative	0.499***	0.624***	0.596***	0.752***	-0.0231	-0.0621	0.642***	0.803***
Figure	0.090**	0.232*	0.0685	0.177	0.830***	0.914***	0.885***	0.943***
Total	0.633***	0.485***	0.135*	0.075	0.325***	0.567***	0.652***	0.795***

****<=0.01; **<=0.05; *<=0.1*

Sustainable Development Goals (SDGs) may be more highly correlated than other criteria in part because they have been clearly defined prior to the creation of our rubric. AI's stark superiority in consistency on grading the Figure criterion compared to the human evaluators may also be due in part to the figure's distinctiveness in the memo pdf. However, the large difference in Figure scores found during Inter-Evaluator analysis (Table 6, in appendix 2) indicates that while internally consistent, AI scoring of figures is highly erroneous. This is consistent with the findings of P. Sharma et al. [25]. The outlier of insignificance in OpenAI's evaluation of Quantitative Information is unexpected and highly noteworthy.

Inter-Evaluator

Aggregating the "A" and "B" scores for each evaluator made the distinction between human evaluators more distinct. Between the A and B scores, 7 of the 18 criteria saw statistically significant differences for the Language evaluator and a different 7 of 18 were significant for the Engineering evaluator. After aggregating results, 13 of the 18 criteria saw statistically significant differences between the Language and Engineering evaluators. Consistency between the aggregated scores of OpenAI and Copilot is roughly comparable to the pre-aggregation A versus B consistency of the human evaluators. 7 of 18 criterion differences were statistically significant, although none were technically significant. The average criteria score difference was also much lower for the two AI evaluators than the two human evaluators. This is consistent with the expectation that even across platforms, LLMs are capable of replicating results when given identical prompting [26].The technically significant differences between Language and Engineering evaluators are shown below in Table 2.

Table 2. Selected differences between language, engineering, OpenAI, and Copilot evaluators

Criterion	Human				AI			
	Language	Engineering	-	||	OpenAI	Copilot	-	||
Body: Background	2.3125	4.2125	-1.9 ***	2.25	5.4	5.6	-0.2	0.975
Style: Precision	5.45	7.8	-2.35 ***	2.6	6.75	6.6	0.15	0.575
Style: Definition	5.325	7.8	-2.475 ***	2.775	5.9625	6.9125	-0.95 ***	1.2
Total	97.45	115.1125	-17.6625 ***	27.262 5	109.4875	112.05	-2.5625	19.0875

****<=0.01; **<=0.05; *<=0.1*

It is unsurprising to find Precision and Definition as the two most inconsistent criteria grades between the Language and Engineering evaluators, as these entail the most pronounced difference in experience between the two fields. LM and ICC scores at this level of analysis were generally comparable for human evaluators and AI evaluators. As with the prior round of analysis, Figure and SDG scores stand out as being highly correlated by both tests. This can be seen in Table 7 in appendix 2.

Aggregating the evaluator scores into "Human" and "AI" type scores yielded some of the most significant scoring differences of the analysis. 12 of 18 criteria had significant paired differences and 10 of 18 had significantly different adjusted R-square values, although only 3 of 18 had significantly different ICC values. These are included in Table 8 of appendix 2, and the technically significant criterion differences are provided in Table 3 below.

Table 3. Selected comparison between human and AI evaluations

Criterion	Human	AI	-	||	LM	ICC
Body: Background	3.2625	5.5	-2.2375 ***	2.475	-0.02314	-0.507
Figure	6.475	3.93125	2.54375 ***	2.70625	0.08403 **	-0.0243
Design: Header	7.9375	6.375	1.5625 ***	1.9125	-0.02534	-0.261
Total	106.28125	110.76875	4.4875 **	24.8875	0.1946 ***	0.422 ***

****<=0.01; **<=0.05; *<=0.1*

The continued significance of the Figure criterion is unsurprising, given the figure's distinctiveness in the memos and what is known about LLM processing of images. LLMs do well with processing simple images, so where the human evaluators may find more detailed images more useful, and thus rate them higher, LLMs may struggle to interpret and evaluate those same images [25]. The significance of the Background criterion may be explained by the human evaluators' observations. In grading the memos, the human evaluators often deducted points from both the Introduction and Background criteria because the student decided to provide the requirements of these two criteria in one section instead of dedicating a section to each. In addition to almost certainly losing valuable information from at least one of the memo components, this practice makes the difference between the Introduction and Background of a memo more difficult for AI systems to detect; LLM text analysis is contingent largely on the proximity of words, and it struggles to parse nuanced distinctions such as the one present here [27]. This resembles prior research which demonstrates human difficulty in distinguishing human-generated and AI-generated abstracts and background sections in academic papers [28].

It is also noteworthy that, in aggregate, AI scores differed from human scores by less than 5% of the total point value, roughly ⅓ the difference observed between the engineering and language evaluator groups and between individual human evaluators. LM and ICC scores confirm that although their scores significantly diverge at several points, human and AI evaluation of our memos remain significantly correlated at the macro level. In all, these results indicate the validity of our rubric and in the use of AI systems to evaluate memos using it.

Qualitative

Since the rubrics solicited typed feedback from both human and AI reviewers, this provided a dataset that could be qualitatively analyzed. It would be difficult to do so with all 40 memos, which would have 60 corresponding end comments to compare and contrast. Thus for this paper, a random sample of six memos were chosen and their corresponding rubrics were collected into a spreadsheet so they could be reviewed: Memos 1, 7, 12, 16, 30, 38. The feedback was then examined with the following questions in mind: what type of feedback was provided to students? What are the similarities and differences between reviewers? Is the feedback useful to students?

The concept of minimal marking directs reviewers to let the rubric scores guide students in the presence or absence of essential elements, freeing the reviewers to focus feedback on what was done well and what would be most beneficial for the students to revise to improve the overall quality of the memo. With the exception of the English faculty member, this was the first time our three human reviewers had been prompted to provide feedback in this way, so perhaps it is no surprise that there was only one instance of 100% agreement between all human and AI reviewers on what a student should do to improve. This occurred in the first memo when all noted the absence of an introduction. Minimal marking would generally direct the reviewers to allow the rubric to communicate such an absence to the student, but it's possible that including an introduction may have been the more substantial revision in this individual case.

There were notable instances of agreement between reviewers. The English faculty member and the Poly Sci undergrad were in agreement with each other in eleven instances across the six memos. A few examples will illustrate how the feedback compares. First, the two reviewers would note when the student could improve concision in their writing. The English faculty

member would rewrite an example sentence a note the exact reduction in word count. For example, “Overall AI in its young age is on the right track to being a majorly helpful and sustainable tool that can aid in the work of not only engineers, but many other careers as well. In terms of how it affects people, AI when developed as a tool makes the jobs of engineers much more manageable” (56 words) was suggested to be revised to “While AI is in its infancy, it is a sustainable tool that can not only optimize engineering, but all careers” (20 words). Second, both reviewers would comment if errors in punctuation and grammar were frequent enough to compromise the effectiveness of the memo, and in many cases provide advice about grammatical rules. In one instance, both reviewers noted a student had included a “reflection” section in their memo was neither a convention nor a move of memos, and that contained information that was not solicited by the assignment prompt (all other reviewers ignored the additional section). In nearly all instances, there as an earnest effort to be clear and specific on how the writer might improve the memo signaled with phrases like “You could improve by,” “I recommend,” or “It would help to” followed by explicit suggestions.

Microsoft Copilot and Chat GPT-4 were largely in agreement with each other, but this unfortunately was not an encouraging result. AI’s feedback tended to cover the entirety of the rubric. Instead of offering specific advice, AI favored reiterating rubric criteria prefaced with phrases like “Ensure that you” or “Be sure to include” so as to sound like direct feedback, but rarely contain advice specific to the memo being examined. Thus agreements are reached simply through the volume of feedback that was provided and by reiterating rubric criteria. Not all of the feedback would be useless to students. For example, AI and Engineering faculty were in often in agreement when students failed to include conventions and moves in their memos.

The key shortcomings of this kind of feedback are numerous. They can overwhelm students with a full aggregate or errors rather than focusing the student on key areas of improvement. The feedback was often not tailored to the student's submission and instead included statements that were reiterations of rubric criteria. The statements were prefaced with phrases like “Make sure to” and “Double check your” before reiterating rubric criteria, often verbatim. Humans would at times highlight or reference the rubric’s language when suggesting improvements. English and Poli Sci Undergrad may have used rubric language as supplement for their lack of expertise in Engineering. For example, “listing the strengths and weaknesses of the technology” requires knowledge they could not explicitly direct students to accomplish. The Engineering faculty preferred this method of feedback for its expediency or because the are not specialized in providing feedback on writing. In either case, this kind of feedback should not read like justifications of rubric scores rather than suggestions for improvement, or the integrity of the feedback is lost.

Engineering faculty were rarely in agreement with their typed feedback. Perhaps this is because Engineering faculty would rely heavily on the rubric to direct students on areas of improvement rather than offer detailed suggestions of their own. However, one key feature of their feedback not present in the feedback of other reviewers were be questions about the accuracy of content. For one example, an Engineering faculty disputed a student's claim about the amount of CO_2 produced while generating electricity. In another instance, the Engineering faculty questioned the student about chemical composition as a way to help the student elaborate on the benefits and drawbacks of the technology described in the memo. This content expertise is unsurprising given

the Engineering faculty's expertise, but the other reviewers would sometimes compensate. English faculty offered suggestions to improve accuracy at the level of correctly citing source material, noting when sources were misquoted or when summaries misrepresented the cited information. For example, the English faculty noted when a student should have cited a quote within a quote, but student cited information as the finding of a study rather than the source's author citing another work as part of their literature review. In instances when the student would fail to connect their project to a UN SDG, both human and AI would occasionally suggest a specific UN SDG relevant to the student's work.

There were seven instances where the two Engineering faculty were in agreement with either the English faculty or the Poly Sci undergrad. In three notable instances, the human reviewers were unsure if the student had included a introduction to their memo, or a background section, or both. This is because a student might have used a heading to denote "background," but the moves within the paragraph were typical of introductions. In each case, the human reviewers asked guiding questions and advised the student to better distinguish their introductions from their background sections. It is encouraging that human reviewers were agreement on how to advise students to improve. AI reviewers, by contrast, might note the absence of clear introduction, but neither Copilot nor OpenAI phrased their feedback in terms of better utilizing moves. In some cases, the AI might praise a student for a strong background section even though three or more human reviewers felt the background section and introductions were conflated or indistinguishable.

There were two rubric criteria within which English and Engineering evaluators often found themselves in agreement. First, on three of the memos, three or more human evaluators agreed that the student needed to better distinguish between an introduction section and a background section. Engineering evaluators would remark on the presence or absence of particular sections, or ask a question about better distinguishing between the two. The Writing evaluators would provide advice on how to better include certain moves that were customary to those respective sections. AI struggled with providing students advice in this regard primarily because the template provided to students included a "Background" heading but not an "Introduction" heading, so no matter what moves were attempted by the student, the AI was prone to regard every paragraph as a background paragraph regardless of moves present. Second, both Writing and Engineering evaluators might remark on the effectiveness of an image included in the memo, and offer advice on a better figure or image to include that might strengthen the memo's content. AI would only reiterate the need for a strong figure, but could not recognize the efficacy of an image, nor advise students accordingly.

Albeit rarely, both humans and AI would make errors in assessment. However, there were notable differences in the frequency and type of errors. The differences can be sorted as false positives or false negatives. Humans could sometimes compliment a student for something that was done well where the student had substantial room for improvement. One example is synthesis, where students would combine sources within single paragraphs to support a topic sentence. Human reviewers were generally more generous whenever just two sources were included in a paragraph. In other words, they could provide students with false positives of competency where there was actually room for growth. AI, by contrast, may at times incorrectly praise students. In two instances, AI praised students for including a UN SDG when no such goal

was present or cited. But AI was more prone towards criticality. AI provided false negatives in eight instances. It would direct students to include a figure when the student had already included one that was descriptively labeled, or it would direct students to include signal phrases referring to the figure when such phrases were actually present. The presence of false negatives is a fascinating finding given that AI is generally more consistent than humans in its scoring of student memos.

Conclusions and Future Work

Formative assessments provide feedback that students can use to improve their learning, while summative assessments provide feedback that helps students understand how well students performed relative to an assignment and in pursuit of a course's learning goals. False positives from humans or AI can undermine formative assessments by assuring them a standard has been met when there is actually room for improvement. By contrast, false negatives may be more damaging. Whether from humans or from AI, false negatives threaten to undermine summative assessments because they can cast doubt on how well students' submissions are being fairly assessed even though the scores, especially in the case of AI, are reliable. This study demonstrated the possibility of both false positives and false negatives from both human and AI evaluators but noted them to be more prominent in the AI case.

Looking at specific assessment content, a major finding was the AI evaluators' struggle in effectively evaluating the substance and use of supporting imagery. It is perhaps unsurprising that AI would struggle to advise students on the presence, absence, or efficacy of images. This finding corroborates what Zhao et al.[4] discovered when they compared the feedback provided by human and AI evaluators. They found that humans were more attuned to how well their students understood how elements contained within an image would relate to each other and convey a deeper meaning about the overall significance of a scene, while AI tended to enumerate elements and count how many times a student might define or describe the total elements contained therein.

From the perspective of the language evaluators, the abundance of rubric criteria required the evaluators to work hard at defining each number on the Likert scale in our rubric. Sometimes, that meant remarking on the presence of absence of certain moves. Other times, it required shades of nuance regarding efficacy of the elements when taken in aggregate. There will always be subjectivity involved when quantifying writing, but we achieved some success with defining our criteria clearly so that Engineering evaluators and AI could follow our lead.

The engineering evaluators generally found the complete, writing-centric rubric to be a fairly heavy assessment instrument and cognitively demanding to use. Evaluation was significantly slower, by orders of magnitude, versus the slimmed down (and less valuable, from a writing perspective) prior rubric. However, with the possible ability of the AI to deploy the rubric, those concerns are largely assuaged as the faculty member would not be the one regularly deploying the rubric. With refinement, this approach represents a promising possibility to incorporate more robust writing feedback without requiring a substantially greater time or cognitive investment from the instructor.

Based on the findings discussed above, it appears that, for the most part, the AI is able to deploy the rubrics with a fairly close but imperfect accuracy, especially when considering qualitative feedback. Perhaps the best path forward is to explore how to scaffold AI into the writing process so that students regard AI as an imperfect partner in their writing process. One way might be to utilize AI comments at the drafting stage as a tool for guiding students (and perhaps tutors and professors) to revise their drafts before they are submitted for evaluation. AI then can be used to score memos, thereby freeing valuable time for professors to provide meaningful feedback on how students might improve a memo of this kind when students encounter them in future classes or their careers. This is especially true when providing feedback on images and figures. Until AI, or our prompting, is strengthened in this regard, then humans will need to be the primary stewards of students' writing process.

References

[1] E. Lindsay and M. N. Sabet Jahromi, "The development of an artificial intelligence classifier to automate assessment in large class settings: Preliminary results," in *2023 ASEE Annual Conference & Exposition Proceedings*, Baltimore , Maryland: ASEE Conferences, Jun. 2023, p. 44085. doi: 10.18260/1-2--44085.

[2] V. Suresh, R. Agasthiya, J. Ajay, A. A. Gold, and D. Chandru, "AI based automated essay grading system using NLP," in *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2023, pp. 547–552.

[3] L. Furze, M. Perkins, J. Roe, and J. MacVaugh, "The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI supported assessment," 2024, *arXiv*: 2403.14692.

[4] R. Zhao, Y. Zhuang, D. Zou, Q. Xie, and P. L. Yu, "AI-assisted automated scoring of picture-cued writing tasks for language assessment," *Educ. Inf. Technol.*, vol. 28, no. 6, pp. 7031–7063.

[5] X. Chen, C. Brawner, M. Ohland, and M. Orr, "A Taxonomy of Engineering Matriculation Practices," in *2013 ASEE Annual Conference & Exposition Proceedings*, Atlanta, Georgia: ASEE Conferences, Jun. 2013, p. 23.120.1-23.120.13. doi: 10.18260/1-2--19134.

[6] J.-D. Yoder, B. K. Jaeger, and J. K. Estell, "One Minute Engineer, Nth Generation: Expansion to a Small Private University," in *2007 Annual Conference & Exposition Proceedings*, Honolulu, HI, 2007.

[7] J. K. Estell, L. Laird, and J.-D. Yoder, "Engineering Personified: An Application Of The One Minute Engineer," in *2008 Annual Conference & Exposition Proceedings*, Pittsburgh, Pennsylvania: ASEE Conferences, Jun. 2008, p. 13.518.1-13.518.18. doi: 10.18260/1-2--3155.

[8] United Nations Department of Economic and Social Affairs, Sustainable Development. Accessed: Jan. 14, 2025. [Online]. Available: https://sdgs.un.org/goals

[9] J. M. Swales, "Aspects of article introductions," University of Aston in Birmingham, Aston ESP Research Report No. 1, Language Studies Unit, 1981.

[10] J. M. Swales, in *Genre Analysis: English in Academic and Research Settings*, Cambridge, UK: Cambridge university press, 1984, pp. 77–86.

[11] V. K. Bhatia, *Analysing genre: Language use in professional settings*. Routledge, 1993.

[12] M. Peacock, "Communicative moves in the discussion section of research articles," *Systems*, vol. 30, no. 4, pp. 479–479, 2002.

[13] T. Dudley-Evans, "The teaching of the academic essay: Is a genre approach possible," in *Genre in the Classroom: Multiple Perspectives*, 2002, pp. 225–235.

[14] A. Cheng, "Understanding learnes and learnign in ESP genre-based writing intstruction," *Engl. Specif. Purp.*, vol. 25, no. 1, pp. 58–75, 2014.

[15] H. Lee, "The effect of the genre-based approach on KFL advanced learners' writing and reading," PhD Dissertation, SOAS University of London, London, UK, 2023.

[16] J. Buehl, "Syle and the professional writing curriculum: Teaching sylistic fluency through science writing," in *The Centrality of Style*, Anderson, SC: Parlor Press, 2013, pp. 279–308.

[17] M. Abdel Latif, "What Do We Mean by Writing Fluency and How Can It Be Validly Measured?," *Appl. Linguist.*, vol. 34, pp. 99–105, Mar. 2013, doi: 10.1093/applin/ams073.

[18] N. Sommers, "Responding to student writing," *Coll. Compos. Commun.*, vol. 33, no. 2, pp. 148–156, 1-82, doi: 10.2307/357622.

[19] R. Haswell, “Minimal Marking,” *Coll. Engl.*, vol. 45, no. 6, pp. 600–604, 1983, doi: 10.58680/ce198313616.
[20] OpenAI, “ChatGPT,” openai.com. Accessed: Jan. 13, 2025. [Online]. Available: https://openai.com/index/gpt-4-research/
[21] K. Davis, “Overview of Copilot,” learn.microsoft.com. Accessed: Jan. 12, 2025. [Online]. Available: https://learn.microsoft.com/en-us/copilot/overview
[22] “Copilot in Bing: Our approach to Responsible AI - Microsoft Support,” support.microsoft.com. Accessed: Jan. 12, 2025. [Online]. Available: https://support.microsoft.com/en-us/topic/copilot-in-bing-our-approach-to-responsible-ai-45b5eae8-7466-43e1-ae98-b48f8ff8fd44
[23] K. Davis, “Frequently asked questions about Copilot,” learn.microsoft.com. Accessed: Jan. 12, 2025. [Online]. Available: https://learn.microsoft.com/en-us/copilot/faq
[24] R. Wells, “Microsoft Copilot Vs. ChatGPT - Which Should I Use For Work?,” Forbes. Accessed: Jul. 18, 2024. [Online]. Available: https://www.forbes.com/sites/rachelwells/2024/01/18/microsoft-copilot-vs-chatgpt-which-should-i-use-for-work/
[25] P. Sharma *et al.*, “A Vision Check-up for Language Models,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 14410–14419. doi: 10.1109/CVPR52733.2024.01366.
[26] A. Myers, “Can AI hold consistent values? Stanford researchers probe LLM consistency and bias,” Stanford HAI. Accessed: Jan. 13, 2025. [Online]. Available: https://hai.stanford.edu/news/can-ai-hold-consistent-values-stanford-researchers-probe-llm-consistency-and-bias
[27] D. Loureiro, K. Rezaee, M. T. Pilehvar, and J. Camacho-Collados, “Analysis and Evaluation of Language Models for Word Sense Disambiguation,” *Comput. Linguist.*, pp. 1–57, May 2021, doi: 10.1162/coli_a_00405.
[28] I. A. Huespe *et al.*, “Clinical Research With Large Language Models Generated Writing—Clinical Research with AI-assisted Writing (CRAW) Study,” *Crit. Care Explor.*, vol. 5, no. 10, p. e0975, Oct. 2023, doi: 10.1097/CCE.0000000000000975.

Appendix 1

Rubric: Engineering Memorandum ENGR 1041

Genre

Is there an introduction, background, body, and conclusion section?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10

Not Attempted Needs Work Proficient Excellent

Introduction

Does the memo have a clear introduction? Does the introduction state a topic, a purpose, and a main point?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10

Not Attempted Needs Work Proficient Excellent

Body

Does the memo include a relevant background section? Does the background section provide information like times, dates, people, and events?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10

Not Attempted Needs Work Proficient Excellent

Does the content of the memo reference the UN Sustainable Development Goal and explain the topic's societal impact?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10

Not Attempted Needs Work Proficient Excellent

Are body paragraphs organized around clear topic sentences and transitions? Are paragraphs unified around main ideas?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10

Not Attempted Needs Work Proficient Excellent

Does the memo include technical information that clearly connects the topic to the field of engineering? Is the technology described in necessary detail, and are its capabilities and limitations described?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10

Not Attempted Needs Work Proficient Excellent

Does the memo provide quantitative information (e.g., measurements, costs, ratings, dates, etc.)? Are claims and statements supported with evidence?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10

Not Attempted Needs Work Proficient Excellent

Does the memo provide two or more reliable sources provided with in-text citations and references? Does the memo avoid dubious reference sources such as blogs, forums, or Wikipedia?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

Figures

Does the memo include at least one figure? Are figures centered in the document? Are figures numbered in order of how they appear in the document? Are figures labeled and do the labels provide sufficient details? Are figures referenced with signal phrases in the body of the memo? Is there a citation provided for each figure?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

Conclusion

Does the conclusion include sentences that restate the main point, thank the reader, and describe future implications for the research? Where applicable, is the author(s)'s contact information provided?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

Design

Does the memo contain visual cues such as headings, subheadings, bullet points, or icons? Are headings clear, descriptive, and demonstrate parallel structure? When necessary, are data visualizations included? Do the visuals enhance the reader's comprehension of the content?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

Does the memo follow block formatting? Are margins set to 1 inch?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

t

Is there a clear memo heading? Is the memo addressed to the instructor?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

Style

Is the writing concise by limiting adjectives and adverbs, and using few and precise verbs? Are needless prefaces absent? Are extraneous and unnecessary details absent?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

Is the document fluent in its language, using precise terms when necessary and defining jargon and other technical terms for layperson readers? Are definitions provided either in parenthetical, full-sentence, or extended formats (ex. glossary or appendix)?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

Are sentences varied in their grammatical structure? Is the memo free of grammatical errors that attenuate the reader's comprehension of the memo? Is the punctuation correct?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

In-text Citations

Does the memo provide in-text citations in IEEE format? Are all citations numbered in square brackets referring to the full citation listed in the references? Do citations follow numerical order?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

End Matter

Are references listed in IEEE format at the end of the memo? Where necessary, is there an appendix with relevant supplementary material?

0_____1_____2_____3_____4_____5_____6_____7_____8_____9_____10
Not Attempted Needs Work Proficient Excellent

Overall aptitudes and opportunities for improvement listed below:

Evaluation notes listed below:

Appendix 2

*Table 4. Average grading outcomes for both iterations of the four evaluators. "A" denotes the average value for evaluator A, "B" denotes the average value for evaluator B, "-" denotes the average difference between evaluators A and B, and "| |" denote the average absolute value difference between evaluators A and B. The final row gives the average difference and average absolute difference between all criteria scores, excluding the total score. Significance is as follows: ***<=0.01; **<=0.05; *<=0.1*

Criterion	Language				Engineering				OpenAI	Copilot
	A	B	-	| |	A	B	-	| |	| |	| |
Genre	4.725	5.15	-0.425 *	1.075	5.775	6.4	-0.2	1.4	0.475	0.85
Introduction	2.85	1.5	1.35 ***	1.35	2.275	4	-1.6	3.6	1.875	1.25
Body: Background	3.45	1.175	2.275	2.625	3.625	6.15	-2.35	2.35	0.8	0.55
Body: SDG	3.35	2.975	0.375	1.175	3.65	4.4	-0.7	1.3	1.35	1.325
Body: Topic Sentences	4.975	4.275	0.7 ***	1.15	5.8	6.55	-0.95 **	1.25	0.5	0.4
Body: Technical	5.2	3.525	1.675	1.975	5.1	5.85	-0.7	1.6	0.5	0.425
Body: Quantitative	5.55	4.25	1.3 ***	1.8	4.275	5.05	-0.7	1.4	0.85	0.4
Body: Sources	6.525	5.55	0.975 ***	1.675	5.825	6.2	-0.55 *	0.95	0.95	0.425
Figure	7.325	6.075	1.25 ***	2.1	5.85	7.5	-1.6 **	2.2	0.8	0.575
Conclusion	4.8	2.625	2.175	2.475	4.325	4.45	0	0.8	1.2	1.225
Design: Visuals	5.75	4.85	0.9	0.95	6.075	8.15	-2.1 ***	2.9	0.75	1.225
Design: Formatting	8.25	8.775	-0.525	2.025	9.925	10	-0.15	0.15	0.525	0.975
Design: Header	7.05	7.475	-0.425 **	0.825	8.2	10	-1.65 ***	1.65	2.25	0.9
Style: Precision	5.35	5.55	-0.2	1.85	7.35	8.95	-1.8 ***	2.2	0.7	0.5

Style: Definition	6.375	4.275	2.1	2.6	7.175	9.25	-2.5 ***	3	0.825	0.725
Style: Grammar	6.975	6.3	0.675 **	1.375	7.9	8.2	-0.4	2.2	0.875	0.55
In Text Citations	7.825	7.9	-0.075	0.925	8.95	7.5	0.4	2.1	1.4	1
End Matter	8.45	7.9	0.55 *	1.45	8.4	9.55	-1 **	1.5	1.1	0.4
Total	104.775	90.125	14.65 ***	29.6	110.475	128.15	-18.55 ***	32.55	18.025	13.7
Mean	–	–	0.814	1.633	–	–	-1.0139	1.808	1.00139	0.761

*Table 5: Linear Model (LM) Adjusted R-Square and Intraclass Correlation Coefficient (ICC) values comparing the "A" and "B" tests for each evaluator. Significance is as follows: ***<=0.01; **<=0.05; *<=0.1*

Criterion	Language		Engineering		OpenAI		Copilot	
	LM	ICC	LM	ICC	LM	ICC	LM	ICC
Genre	-0.010	-0.157	-0.0516	0.0617	-0.0169	0.0828	0.0751**	0.317**
Introduction	0.276***	0.316**	-0.0469	0.0314	0.0506*	0.272**	0.306***	0.554***
Body: Background	0.081***	-0.0346	-0.0183	-0.235	0.00790	0.176	0.552***	0.709***
Body: SDG	0.347***	0.595***	0.168**	0.45**	0.421***	0.662***	0.428***	0.654***
Body: Topic Sentences	0.319***	0.454***	0.433***	0.536***	0.0927**	0.344**	0.507***	0.705***
Body: Technical	0.0689*	0.0913	0.0425	0.274	-0.0144	0.0883	0.240***	0.517***
Body: Quantitative	0.499***	0.624***	0.596***	0.752***	-0.0231	-0.0621	0.642***	0.803***
Body: Sources	0.205***	0.354**	0.125*	0.334*	0.0520**	0.287**	0.310***	0.565***
Figure	0.090**	0.232*	0.0685	0.177	0.830***	0.914***	0.885***	0.943***
Conclusion	-0.02364	-0.361	0.276**	0.518***	-0.00142	0.122	0.186***	0.454***
Design: Visuals	0.231***	0.27**	0.0442	-0.301	0.0519*	0.26**	0.403***	0.621***
Design: Formatting	0.126**	0.298**	NA	0	0.00464	0.182	0.0534*	0.275**
Design: Header	0.289***	0.508***	NA	-0.244	-0.0249	0.0396	0.417***	0.65***
Style: Precision	-0.0249	0.0463	-0.0297	-0.119	0.00209	0.175	0.162***	0.435***
Style: Definition	0.0846**	0.0721	-0.0410	-0.17	0.266***	-0.0895	0.300***	-0.0366
Style: Grammar	0.223***	0.413***	-0.0235	0.177	0.0643*	0.302**	0.459***	0.691***
In Text Citations	0.707***	0.848***	0.175**	0.483**	0.0783**	0.328**	0.0877**	0.239*
End Matter	0.0170	0.179	-0.0255	-0.264	0.0955**	0.352**	0.245***	0.425***
Total	0.633***	0.485***	0.135*	0.075	0.325***	0.567***	0.652***	0.795***

*Table 6: Average grading outcomes for the composite scores of the four evaluators. "-" denotes the average difference between evaluator, and "| |" denotes the average absolute value difference between evaluators. The final row gives the average difference and average absolute difference between all criteria scores, excluding the total score. Significance is as follows: ***<=0.01; **<=0.05; *<=0.1*

Criterion	Human				AI			
	Language	Engineering	-	| |	OpenAI	Copilot	-	| |
Genre	4.9375	5.825	-0.8875 ***	1.0125	5.8375	5.925	-0.0875	0.7875
Introduction	2.175	2.675	-0.5	1.425	3.3125	4.275	-0.9625 ***	1.7375
Body: Background	2.3125	4.2125	-1.9 ***	2.25	5.4	5.6	-0.2	0.975
Body: SDG	3.1625	3.825	-0.6625 ***	0.9125	4.8	4.2875	0.5125 **	1.0875
Body: Topic Sentences	4.625	6.0375	-1.4125 ***	1.6375	6.025	6.05	-0.025	0.575
Body: Technical	4.3625	5.275	-0.9125 ***	1.3625	5.875	6.0625	-0.1875 *	0.5125
Body: Quantitative	4.9	4.45	0.45 *	1.125	5.675	5.8	-0.125	0.625
Body: Sources	6.0375	5.9625	0.075	1.125	7.3	7.5375	-0.2375	0.7875
Figure	6.7	6.25	0.45 *	1.275	3.875	3.9875	-0.1125	0.6875
Conclusion	3.7125	4.325	-0.6125 ***	1/2375	4.7	4.8875	-0.1875	1.4625
Design: Visuals	5.3	6.6	-1.3 ***	1.45	6.4	5.4375	0.9625 ***	1.5375
Design: Formatting	8.5125	9.9625	-1.45 ***	1.45	9.6125	8.7125	0.9 ***	1.25
Design: Header	7.2625	8.6125	-1.35 ***	1.425	6	6.75	-0.75	2.55
Style: Precision	5.45	7.8	-2.35 ***	2.6	6.75	6.6	0.15	0.575
Style:	5.325	7.8	-2.475	2.775	5.9625	6.9125	-0.95 ***	1.2

Definition			***					
Style: Grammar	6.6375	8	-1.3625 ***	1.4625	7.1375	7.3	-0.1625	0.7375
In Text Citations	7.8625	8.85	-0.9875 ***	1.5125	7.35	7.975	-0.625 ***	1.2
End Matter	8.175	8.65	-0.475 *	1.225	7.475	7.95	-0.475 **	0.8
Total	97.45	115.1125	-17.6625 ***	27.2625	109.4875	112.05	-2.5625	19.0875
Mean	–	–	-0.981	1.515	–	–	-0.142	1.0604

*Table 7: Linear Model (LM) Adjusted R-Square and Intraclass Correlation Coefficient (ICC) values comparing the composite evaluator scores for each evaluator type. Significance is as follows: ***<=0.01; **<=0.05; *<=0.1*

Criterion	Human		AI	
	LM	ICC	LM	ICC
Genre	-0.01181	-0.16	0.1119**	0.251*
Introduction	0.2268***	0.471***	0.1391**	0.318**
Body: Background	0.05479*	-0.0397	0.07253*	0.251*
Body: SDG	0.6426***	0.748***	0.6593***	0.788***
Body: Topic Sentences	0.1629***	0.0972	0.09314**	0.329**
Body: Technical	0.18***	0.321**	0.2507***	0.497***
Body: Quantitative	0.6141***	0.779***	0.163***	0.352**
Body: Sources	0.1172**	0.269**	0.03979	0.24*
Figure	0.365***	0.595***	0.8677***	0.934***
Conclusion	0.003653	0.0902	-0.01074	0.117
Design: Visuals	-0.02054	-0.194	0.1453***	0.159
Design: Formatting	-0.01899	-0.192	-0.01343	-0.287
Design: Header	0.3779***	0.288**	-0.02618	-0.0274
Style: Precision	-0.02631	-0.468	-0.0004985	0.152
Style: Definition	-0.02307	0.326*	0.126**	-0.0574
Style: Grammar	0.3736***	0.239*	0.1411***	0.395***
In Text Citations	0.4458***	0.616***	0.1288**	0.319**
End Matter	-0.000261	-0.178	0.08861**	0.22*
Total	0.2754***	0.0464	0.3382***	0.45***

*Table 8: Average grading outcomes for the composite scores of the two evaluator types and Linear Model (LM) Adjusted R-Square and Intraclass Correlation Coefficient (ICC) values comparing the composite scores of the two evaluator types. "-" denotes the average difference between evaluator, and "| |" denotes the average absolute value difference between evaluators. The final row gives the average difference and average absolute difference between all criteria scores, excluding the total score. Significance is as follows: ***<=0.01; **<=0.05; *<=0.1*

Criterion	Human	AI	-	| |	LM	ICC
Genre	5.38125	5.88125	-0.5 ***	0.8375	-0.02421	-0.089
Introduction	2.425	3.79375	-1.36875 ***	1.73125	0.1379 **	0.184
Body: Background	3.2625	5.5	-2.2375 ***	2.475	-0.02314	-0.507
Body: SDG	3.49375	4.54375	-1.05 ***	1.6375	0.4148 ***	0.541 ***
Body: Topic Sentences	5.33125	6.0375	-0.70625 ***	0.85625	0.05404 *	0.061
Body: Technical	4.81875	5.96875	-1.15 ***	1.3	0.08223 **	-0.073
Body: Quantitative	4.675	5.7375	-1.0625 ***	1.6	0.07517 **	0.0945
Body: Sources	6	7.41875	-1.41875 ***	1.41875	0.2133 ***	-0.155
Figure	6.475	3.93125	2.54375 ***	2.70625	0.08403 **	-0.0243
Conclusion	4.01875	4.79375	-0.775 ***	1.1875	0.0143	0.03
Design: Visuals	5.95	5.91875	0.03125	1.09375	-0.01772	0.0934
Design: Formatting	9.2375	9.1625	0.075	0.7	-0.01416	0.104
Design: Header	7.9375	6.375	1.5625 ***	1.9125	-0.02534	-0.261
Style: Precision	6.625	6.675	-0.05	0.65	0.151 ***	0.351 **
Style: Definition	6.5625	6.4375	0.125	0.925	0.1217 **	-0.0221
Style: Grammar	7.31875	7.21875	0.1	0.85	0.0896 **	0.319 **
In Text Citations	8.35625	7.6625	0.69375 *	2.10625	0.08848 **	0.206 *
End Matter	8.4125	7.7125	0.7 ***	0.9	-0.02131	-0.123
Total	106.28125	110.76875	4.4875 **	24.8875	0.1946 ***	0.422 ***
Mean	–	–	0.249	1.383	–	–