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 #48290

Analyzing Feedback of an AI tool for formative feedback of Technical Writing abilities

Dr. Sean P Brophy, Purdue University at West Lafayette (COE)

Dr. Sean Brophy is a learning scientist, computer scientists and mechanical engineering who design learning environments enhances with technology. His recent research in engineering design focuses on students' development of computational thinking through physical computing. His work involves students' design of smart systems that integrate both hardware and software to achieve a client's needs. In this work students communicate their ideas through proposal writing and communicate its results through written and oral reports. Currently he is exploring the use of generative AI to provide format feedback to students as they generate these work products.

Mrs. Fadhla Binti Junus, Purdue Engineering Education

A former Assistant Professor and Tenured Lecturer in Information Technology at the Department of Science and Technology, State Islamic University Ar-Raniry, Banda Aceh, Indonesia. Currently pursuing a Ph.D. at the School of Engineering Education, Purdue University, Indiana, USA. Her research uniquely combines industry experience with academic expertise, focusing on technology-enhanced learning. Specifically, her work centers around developing personalized learning environments for higher education students studying computer programming. She is particularly interested in investigating students' programming learning processes, exploring methods to simplify programming instruction, examining theoretical foundations for effective instructional design, and integrating artificial intelligence technologies to facilitate peer-like knowledge construction.

Title: Analysing Feedback of an AI Tool for Formative Feedback of Technical Writing Abilities

Abstract

This Full paper describes the use and validation of feedback provided by an AI tool to support students' technical writing abilities. The project is part of a larger study to address the challenges of providing students with rich formative feedback to improve the quality of their writing artifacts before submitting their final draft for review by the instructional team. Formative feedback is an ongoing assessment process aimed at improving students' understanding of the subject matter. It enables students to identify their strengths and weaknesses throughout their learning journey and assists the instructor in evaluating the effectiveness of their teaching methods in achieving learning objectives. However, providing feedback in large classrooms can pose significant challenges for instructors, particularly with complex assignments such as essay writing, report writing, and proposal writing. Even with the support of an instructional team, this process can be time-consuming and increase workload. We employed Charlie, a neural network-enabled grader, to provide feedback on students' writing assignments. To receive feedback from Charlie, students only need to submit one draft, the minimum requirement for the assignment, although multiple submissions are allowed. The results indicate that Charlie's feedback is generally effective, but improvements are needed in accuracy and recognizing revisions. The findings also reveal that students integrated feedback well, particularly in refining their career goals and action plan sections. This study suggests that the design of learning activities could be refined to encourage students to be more metacognitive in their work refinements. This study will contribute to the growing body of literature on generative AI in education, particularly in providing scalable, timely, and relevant formative feedback on technical writing assessments.

I. Introduction

In problem/project-based instructional models, students are often required to demonstrate their knowledge and skills through written reports and essays. These assignments are crucial for developing students' ability to convincingly communicate their evidence to support their claims. Dannels et al. [1] emphasize that students proficient in technical writing are better prepared for the engineering profession's demands. However, students tend to prioritize technical aspects of projects over writing quality, often undervaluing the latter despite instructors' goals. Therefore, students need learning activities to engage in iterative cycles of technical writing. They can benefit from frequent feedback on their draft of technical reports to improve their writing competencies [2 Morris]. The instructional challenge is providing quality feedback to a large population of students in a timely manner.

Our team has been exploring opportunities for using Generative AI tools to support students writing competencies. A development team at our institution in collaboration with the investigator and instructors, integrated a tool called "Charlie" to read and provide feedback on students' writing drafts. This larger project has multiple components that will investigate the potential of Charlie in the following writing activities in undergraduate engineering courses:

1. Project proposals for design projects
2. Reflective essays on career goals and plans for achieving those goals
3. Project reports (labs and design projects)

We anticipate that this AI agent, can provide meaningful feedback to students to increase the quality of their writing drafts before turning them in for final review by the instructional team. This first study characterized the feedback provided by Charlie to determine its quality relative to a human evaluator. Our initial research questions include:

R1: How effectively does Charlie provide formative feedback to students?
R2: How well do students integrate feedback to refine their drafts for final submission?

In the following sections, we briefly summarize the literature around AI tools in general writing activities and STEM-related reports. Next, we describe the basic structure of Charlie and the methods used to evaluate the feedback it generates. We conclude with a summary of the results related to the two questions and end with a discussion of the current system and recommendations for improvement.

II. Literature Review

Several instructional methods can be used to support students writing in classes with large enrolments. The methods are both a combination of strong pedagogical practices and the effective integration of technologies. First, Moskovitz and Kellogg [3] argue that detailed rubrics can guide students in structuring their reports and focusing on critical elements such as clarity, coherence, and technical accuracy. Automated Writing Evaluation (AWE) systems, such as Grammarly and Pigai, have been instrumental in this domain. These systems utilize natural language processing (NLP) to identify errors and suggest improvements, thereby assisting both learners and educators in the writing process[1]. These tools can provide explicit feedback with suggestions on spelling and word choice.

A more challenging learning outcome to achieve for students is generating clear, concise and accurate presentation of ideas with strong rationale to justify their claims. Peer review sessions can enhance students' understanding of writing conventions and improve the quality of their reports [4]. In these sessions peers read each other's work and provide feedback on how well the ideas are communicated. Instructors can lead these sessions by helping students learn to critically evaluate writing samples and generate useful feedback following research-based methods [5]. One drawback of this method is it increases the load on the students to perform this assignment in conjunction with their course work. Unless technical writing is a fundamental learning outcome for a course, then using this method is hard to justify as a large part of the course requirements. Adding an AI agent as a peer could be an opportunity.

Recent studies explored the efficacy of AI-generated feedback compared to human feedback. For instance, a study by Escalante et al. (2023) [6] examined the learning outcomes of university students receiving feedback from ChatGPT (GPT-4) versus human tutors. The results indicated no significant difference in learning outcomes between the two groups, suggesting that AI-generated feedback can be effectively incorporated into writing instruction[7].

A systematic review by Shi and Aryadoust [8] provided a comprehensive overview of AI-based automated written feedback (AWF) research. The review highlighted the diverse contexts in which AWF has been studied, the various systems employed, and the mixed results regarding its impact on writing performance. The authors emphasized the importance of a blended approach, combining AI and human feedback to leverage the strengths of both [9].

Furthermore, a meta-analysis investigated the effects of AWE tools on students' writing performance. The analysis revealed that AWE tools, powered by advances in AI technology, can provide individualized feedback that positively impacts students' writing skills [10].

In the context of STEM disciplines, the application of AI to improve technical writing skills has gained attention. A systematic review by Xu and Ouyang (2022) examined the use of AI technologies in STEM education, highlighting the potential of AI to enhance technical writing skills through automated assessment and personalized feedback. The review identified several AI applications, such as intelligent tutoring systems (ITS) and learning analytics, that support the development of technical writing skills in STEM students [7].

Additionally, a study by Cai et al. (2024) [9] focused on the role of AI in interdisciplinary learning, including STEM disciplines. The study found that AI tools can enhance students' technical writing skills by providing real-time feedback and promoting critical thinking and problem-solving abilities [8]. The authors emphasized the importance of integrating AI tools with traditional teaching methods to maximize their effectiveness [9].

Overall, the literature suggests that AI-based formative feedback systems hold promise for enhancing students' writing abilities, including technical writing skills in STEM disciplines. However, a balanced approach that integrates both AI and human feedback is recommended to maximize the benefits of these technologies.

III. Methods

A. Formative feedback system

This study focused on the evaluation of formative feedback generated by an AI tool called Charlie. Charlie can provide students with rich, informative written feedback for complex writing assignments such as essays, reports, and project proposals. This neural network-enabled grader has been innovated by the [development team] [11]at a public university in the Midwest since 2020. Instructors across different departments at the university have used it in a variety of ways. Instructors “teach” Charlie their expectations for an assignment using a well-defined rubric that contains a set of criteria for evaluation. Each criterion has a description of excellent performance. As an example, see Table 1 for a rubric used to evaluate students’ essays on their career goals and plan of action to achieve their goals. Charlie generates feedback based on the rubric coupled with a Large Language Model (LLM) to manage the natural language processing needed to read students’ work and generate feedback.

Charlie’s designers structure its responses based on recommended methods for providing quality feedback [5]. For each criterion, Charlie provides feedback on what was good about a work product and provides suggestions on what could be improved. Charlie is well integrated into the instructional technology infrastructure at the university. The university technology team developed a peer evaluation platform called Circuit, which is integrated into the Brightspace learning management system (LMS). Therefore, assignments using Charlie are simple for the instructor to implement into the standard assignment system. Circuit was developed to use peer evaluation as a method for students to critically review their peers’ work and provide feedback to each other. Circuit manages the process of pairing students together as reviewers and provides an interface for students to submit their feedback to their peers. Circuit also archives all submissions students made and the feedback given to the students by their peers. Charlie is integrated into Circuit as another “peer” who provides feedback. Students can refine their written work products based on the feedback and resubmit

them to Charlie as many times as they wish. In our study, Charlie was the only "peer" assigned to every student in Circuit. Currently, Charlie has no memory of prior submissions, therefore, every resubmission is an independent entry to Charlie. That means it cannot replicate a human grader who uses metacognitive prompts like "I see you choose not to use my recommendation… what was your rationale for not including it?". This lack of memory of prior submission may be a benefit. However, the developers are considering adding a method to integrate prior recommendations for change and feedback, which will be helpful in future studies.

B. Setting and Population

The study was conducted in a first-engineering class consisting of 89 students in the 2024 fall semester. For 40 students, this was their first semester of college, and they are in the 18-20 year age range. These 40 students were the focus of this study.

C. Sampling Method and Data Collection

The primary objective of this study was to evaluate Charlie's feedback in enhancing students' writing skills. To achieve this goal, we used purposive sampling to ensure that data samples provide the most insight into the research questions (RQs): (1) How effectively does Charlie provide formative feedback to students? and (2) How well do students integrate feedback to refine their drafts for final submission? Thus, we imported data from Circuit into a Microsoft Excel format file. Then, we familiarize ourselves with the nature of the data structure to help us establish selection criteria. As we figured out that Circuit logs the data based on the timestamp of every submission, student's unique aliases, links to drafts submitted, and feedback for each rubric criterion, we decided to look at the pattern of time intervals for each submission as our first selection category. To complement this approach, we set the number of draft submissions as the second selection criterion because we also aimed to evaluate whether students' writing skills changed because of feedback (RQ2).

To obtain the most meaningful data samples based on these criteria, we carefully inspected the time span of each submission record by filtering the data according to students' unique aliases. Using the conditional formatting feature in Excel, we quickly located duplicate values of students who submitted multiple drafts. In turn, we found 65 duplicate records belonging to 25 aliases, meaning several students submitted their revised draft more than once. We then took notes of each alias of those 25 students, their number of revised submissions, and the row numbers indicating the record index position in the imported data file. It turned out that 15 students submitted their revisions two times. Seven resubmitted three times, only one student resubmitted four times, and the remaining two resubmitted their drafts five times.

Subsequently, we grouped the records based on the number of iterations made to submit the drafts, which ranged from two to five. Then, to narrow down the data sample candidates, we

Table 1: Essay Rubric for Career Exploration Essay Assignment.

Criteria	Description of Excellent Essay
Clarity and Specificity of Goals	Goals are clearly defined, specific, and well-articulated. They reflect a deep understanding of career interests and are achievable within a realistic timeframe. Response is personal and shared in the first person.
Relevance and Realism of Action Plan	The action plan is highly relevant to the stated goals, with clear, realistic, and practical steps. It includes short-term and long-term actions.
Self-reflection and Insight	Demonstrate deep self-reflection, understanding of personal strengths and weaknesses, and a clear connection between personal interests and career goals.
Feasibility and Resources	Identifies necessary resources and demonstrates a strong understanding of how to obtain them. The plan is highly feasible.
Organization and Writing Quality	Essay is well-organized, logically structured, and free of grammatical errors. Writing is clear and engaging.

decided to pick a record alias representing each data group based on the distance of the row numbers. The distance indicates the time interval gap between submissions. Therefore, for the group of two resubmissions, we selected a record with the largest distance between the initial and final submissions. We assume this is a weak indicator of potentially taking more time to reflect on the feedback and make suggestions in contrast to making small iterative changes before resubmission. Conversely, because we noticed there were sequential patterns of time intervals of the submissions, reversed conditions (i.e., the least distance) were applied to the group of three submissions. As for the group of four resubmissions, because only one student resubmitted four times, this single occurrence was automatically included in the data samples. Finally, we included all records within the group of five, considering that the two records show the resubmissions made nearly simultaneously, with only a slight delay. Following this procedure, we ultimately collected five data records, as summarized in Table 2. We can see that a total of 19 essay drafts were submitted across five samples. Because Charlie generated feedback for all five rubric criteria described in Table 1, we included a total of 95 feedback items for further analysis.

Table 2: Data Samples

Data Alias	Number of Submissions	Row number	Distance of Submissions [minutes]	Reasons for Inclusion
S1	2	12;94	82	The data has the longest time span between resubmissions.
S2	3	29;30;31	2	The data has the shortest time span between the first and the final submission. Also, the submissions were made in consecutive time.

S3	4	44;47;61;97	53	It is the only resubmission with four iterations.
S4	5	60;102;103; 104;105	45	The data has the highest frequency of resubmissions, with a longer time span between the first and the final submission. The data also contains resubmissions that were made in consecutive time.
S5	5	106;108; 111;112;113	7	The data has the highest number of resubmissions, with a shorter time span between the first and the final submission. Also, the submissions were made at a close time frame.

D. Data Analysis

A qualitative content analysis (QCA) was the methodological approach, because of the nature of the research questions and data samples. According to Krippendorff [12] and Schreier [13], QCA is an ideal method for descriptive research questions requiring rich, detailed data from various textual sources. Descriptive research questions typically begin with words like "what" or "how" and aim to explore and describe characteristics, phenomena, or situations of a particular topic [11], [14], [15]. In this study, both research questions were framed using "how," indicating their descriptive nature. Moreover, the data source comprised (i.e., Charlie's feedback and students' essays) entirely in text format.

To assess the effectiveness of feedback from Charlie (RQ1), we decided to use a deductive approach by following the assessment criteria developed by Institutional Data Analytics + Assessment [11]. The developers of Charlie defined six criteria to assess the quality of essay feedback generated by Charlie. Their quality feedback criteria (QFC) consist of (1) *Clarity* which defines clarity, conciseness, and understandability of the feedback comments; (2) *Tone* assesses whether the feedback used positive, growth-mindset language that encourages the student; (3) *Alignment with the Paper Rubric* checks if the feedback aligns with the instructor's rubric and provides specific guidance on meeting assignment criteria; (4) *Accuracy and Evidence-Based Guidance* measures the accuracy of the feedback and its ability to provide detailed, evidence-based guidance; (5) *Rootedness in Student's Draft* determines how well the feedback comments are rooted in the student's draft, using direct quotes and examples; and (6) *Encouragement of Critical Thinking* evaluates whether the feedback encourages the student to engage in critical thinking strategies and addresses higher-order concerns in the draft. For this study the data analysis focused on the third, fourth, and fifth criteria mainly because this study aims to validate whether the feedback generated is accurate, aligns with the instructor's rubric outlined in Table 1, and corresponds to each essay draft. This decision also reflects the nature of the data source, where feedback was explicitly linked to each rubric criterion for every resubmitted draft.

1. Analysis of the initial submission

The steps taken in evaluating the feedback began with familiarizing the human evaluator with the instructor-designed rubric and assignment guidelines. This process aimed to train the evaluator to quickly identify the essay sections targeted by the rubric and ensure that the evaluator understood the criteria for assessing the essays. Once acquainted with the rubric and guidelines, the evaluator reviewed the essay's initial draft and provided feedback and scores according to the rubric's grading scales. Scores were categorized into low, medium, and high to reflect the degree of alignment with the rubric. For example, scores within the "good" range were further classified as good-low (14), good-medium (15.5), and good-high (17).

Afterward, the evaluator analyzed Charlie's feedback on the corresponding essay. This process started with a review of feedback for each rubric criterion, during which specific words, phrases, and sentence fragments indicating recommendations for refinement were highlighted. Since the feedback is presented into two categories, Suggestions and Areas for Improvement, we classified the highlighted texts into "Charlie suggests" for the former and "Charlie identifies" for the latter. Based on these classifications, the evaluator assigned a score indicating the alignment level with the instructor's rubric. Finally, the evaluator provided a rationale to explain why the rated score was given. All these review procedures were conducted iteratively for all five rubric criteria.

Upon the completion of evaluating all of Charlie's feedback, a summary note of the overall scores assigned by the human evaluator during the review process of both the essay drafts and Charlie's feedback was recorded. We also documented a list of all rationales as a basis for developing categories for changes made to the resubmitted drafts. Both documentation containing scores and rationales would later contribute to answering RQ2, which investigates how well students incorporated the feedback into their final draft revisions.

As the next step, we compared the feedback given by the human evaluator and Charlie to assess their alignment with the instructor's rubric. Besides, this step aimed to ensure that the human and Charlie looked at the same essay sections in generating feedback for each rubric criterion. We then commented on Charlie's feedback by describing our evaluation in terms of its accuracy, comprehensiveness, adherence to the rubric criteria, and whether it was grounded in the student's draft. After that, we analyzed our commentaries and the texts under the labels "Charlie suggests" and "Charlie identifies" to confirm the accuracy of our evaluation description. Finally, we documented the findings of our analysis on each feedback iteration into a coding frame structure comprising the three criteria of the IDA+A assessment framework we focused on (i.e., alignment with the paper rubric, accuracy and evidence-based guidance, and rootedness in the student's draft). Building on the findings, we assigned a score ranging from one to five for the three criteria of the framework to answer RQ1, which examines the effectiveness of Charlie providing formative feedback to students.

2. Analysis of the subsequent submissions

We used the Compare feature in Microsoft Word to analyze the revised drafts. This feature highlights any changes made in the latest draft, including insertions, deletions, moves, and formatting. We recorded the number of changes for every comparing draft and evaluated whether revisions aligned with Charlie's feedback from the previous iteration. We then commented on each rubric criterion regarding the revised draft content, assigned a score as we did in the initial submission, and provided justification for the score. After that, we analyzed Charlie's feedback on the corresponding revised draft by applying the same approach employed during the analysis phase of the initial submission. Finally, we updated the documents containing the overall scores and coding framework.

E. Steps to Minimize Bias

We employed multiple strategies to mitigate potential bias in this study. First, we established the coding frame on the existing assessment framework validated by the Charlie developer. This approach ensures that our analysis is consistent, systematic, and aligned with the validated standard [16]. Second, because the human evaluator also served as the data coder and the research team included an expert who created the rubric, regular meetings were held to discuss any coding issues and review the findings of the coded material. The meetings also served as the training and calibration sessions conducted to refine the analysis methods, ensuring the consistency of the coding frame [12], [13]. Finally, to maintain transparency, we thoroughly documented and reported all steps taken in the research process, ensuring accountability in the application of the methods [13].

IV. Results

1. RQ1: How effectively does Charlie provide formative feedback to students?

In this study, we evaluated formative feedback generated by Charlie to improve students' writing skills. The findings for the first research question (RQ1), which aims to examine the effectiveness of Charlie's feedback, are organized based on three criteria defined by the IDA+A [11] which we will refer to as the IDA+A assessment framework. This quality feedback criteria comprises quantitative scales ranging from one to five, with one representing the most negative tendency and five indicating the most positive tendency. Due to this, the effectiveness of Charlie's feedback was assessed both qualitatively and quantitatively. For the qualitative assessment, we analyzed the connections between our comments and both feedback labels (i.e., "Charlie suggests" and "Charlie identifies"). Then, based on this qualitative assessment, we assigned a quantitative scale for the three criteria of the framework. However, we present the quantitative results first to provide a broad-to-specific understanding of the scope of the criteria used in our analysis.

Table 3 shows the scores of feedback effectively related to the three criteria defined in the IDA+A assessment framework. The first criterion, *Alignment with the Paper Rubric,* evaluates whether the feedback aligns with the instructor's rubric and provides specific guidance on meeting assignment criteria. The second criterion, *Accuracy and Evidence-Based Guidance*, measures the accuracy of the feedback and its ability to provide detailed, evidence-based guidance. Finally, the third criterion, *Rootedness in Student's Draft,* looks at how well the feedback comments are rooted in the student's draft, using direct quotes and examples. Overall, the scores in Table 3 indicate a positive tendency, meaning that Charlie effectively provided feedback. Specifically, the scores reveal that Charlie's feedback mostly aligned with the instructor's rubric, as the figures are closest to five. However, the other two categories contain scores approaching the negative tendency, as exhibited in Sample 2 and 5.

Table 3: Feedback Scores Based on the IDA+A Assessment Framework

	Alignment with the Paper Rubric	**Accuracy and Evidence-Based Guidance**	**Rootedness in Student's Draft**

Sample1	4.80	5.00	4.40
Sample2	4.80	2.33	2.33
Sample3	4.70	5.00	4.90
Sample4	4.92	5.00	5.00
Sample5	4.80	4.20	3.88

The scores above are derived from the analysis of our comments and texts labeled with "Charlie suggests" and "Charlie identifies," as illustrated in Table 4. The analysis reveals that while Charlie's feedback generally aligns with the instructor's rubric, offers detailed, evidence-based guidance, and is rooted in the student's drafts, it has issues with some inaccuracies that impact its overall effectiveness.

In terms of the Alignment with the Instructor's Rubric category, the feedback across all samples generally aligns with the instructor-designed rubric. However, there are recurring issues with the second rubric criterion (i.e., Relevance and Realism of the Action Plan) due to the missing suggestions for short-term and long-term actions. Such issues were evident in Samples 1, 2, and 5, each receiving the same score of 4.80, as we can see in Table 3 above. Besides, Sample 4 achieved a score of 4.92, indicating the highest level of alignment with the assignment rubric, although two feedback instances also revealed a similar problem with the second rubric criterion. Notably, Sample 3 received the lowest score (4.70) due to issues with both items of the first rubric, Clarity and Specificity of Goals (R1), and the second rubric (R2). Specifically, feedback in R1 improperly included suggestions for short-term and long-term activities because this feedback should belong to R2.

The Accuracy and Evidence-Based Guidance of Charlie's feedback overall provides detailed and evidence-based guidance that is mostly accurate, however, there are notable inaccuracies in Sample 2 and one instance in Sample 5. In these cases, Charlie keeps generating feedback despite no revisions being made to the drafts, resulting in lower scores for both samples compared to others. This relates to Charlie treating each submission as new and independent to prior submission and provided feedback.

Furthermore, for the Rootedness in the Student's Draft category, the feedback shows varying degrees of rootedness in the drafts, with some samples being more deeply rooted than others. The feedback in Samples 1, 3, and 5, with scores ranging from 3.80 to 4.90, is somewhat rooted in the drafts with some different limitations. Sample 1's feedback engages with the content by mentioning the author's name, career goal, and resources, yet overlooks revisions and omits comments on some headings. The feedback in Samples 3 and 5 incorrectly comments on the need for headings already present, with Sample 5 repeatedly making this error and generating feedback despite no revisions. In addition, Sample 2, which has the lowest score in this category (2.33), indicates that the feedback is inconsistently rooted because it does not mention the author's name and fails to comment on missing headings, yet it continues to generate feedback despite no revisions. Conversely, the feedback generated in Sample 4, which has the perfect score, is deeply rooted in the student's draft, as it consistently addresses specific elements and revisions.

Based on the findings described above, formative feedback from Charlie is effective, especially in providing detailed guidance on strengthening arguments. However, its effectiveness is limited because some samples showed inaccuracies and failure to recognize revisions students made.

Table 4: Feedback Category Matrix Based on the IDA+A Assessment Framework

Sample#	IDA+A Criteria	CFB1	CFB2	CFB3	CFB4	CFB5
Sample1	**Alignment with the Paper Rubric**	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	X	X	X
	Accuracy and Evidence-Based Guidance	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	X	X	X
	Rootedness in Student's Draft	• It mentions the author's name, career goal (e.g., pharmaceutical industry), and specific resources name (e.g., IISE membership)	• It does not recognize the revised parts because Charlie keeps suggesting similar	X	X	X

Sample#	IDA+A Criteria	CFB1	CFB2	CFB3	CFB4	CFB5
		• It provides examples for R1, R2 & guidance.	feedback (e.g., specific roles) • It provides examples for R1, R2, and R4 • It provides guidance for what should be added and what missing related to specific action/resource (e.g., study abroad). • For R5, headings are still not there, but Charlie does not comment on this.			
Sample2	**Alignment with the Paper Rubric**	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-	X	X

Sample#	IDA+A Criteria	CFB1	CFB2	CFB3	CFB4	CFB5
			term and long-term actions.	term and long-term actions.		
	Accuracy and Evidence-Based Guidance	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	The comments contain significant inaccuracies, as the draft contains no revision at all.	The comments contain significant inaccuracies, as the draft contains no revision at all.	X	X
	Rootedness in Student's Draft	• It does not mention the author's name. • For R5, headings are still not there, but Charlie does not comment on this.	• It mentions the author's name • It keeps generating feedback despite no revisions	• It mentions the author's name • It keeps generating feedback despite no revisions	X	X
Sample3	**Alignment with the Paper Rubric**	The feedback generally aligns with the instructor's rubric, but two rubric items (R1&R2) were not appropriate. R1 should not contain suggestions for the short-term and long-term activities, as they belong to R2.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with two rubric items (R2&R5) does not contain suggestions to include short-term and long-term actions nor does specify the number of grammatical errors although	X

Sample#	IDA+A Criteria	CFB1	CFB2	CFB3	CFB4	CFB5
					proofreading is still needed.	
	Accuracy and Evidence-Based Guidance	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	X
	Rootedness in Student's Draft	It is somewhat rooted in the student's draft, as Charlie keeps commenting in R5 to include headings for each writing section, while the headings are there.	It is deeply rooted in the student's draft.	It is deeply rooted in the student's draft.	It is deeply rooted in the student's draft.	X
Sample4	**Alignment with the Paper Rubric**	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback consistently aligns with the instructor's rubric, as all rubric items were addressed accordingly.	The feedback consistently aligns with the instructor's rubric, as all rubric items were addressed accordingly.	The feedback consistently aligns with the instructor's rubric, as all rubric items were addressed accordingly.

Sample#	IDA+A Criteria	CFB1	CFB2	CFB3	CFB4	CFB5
	Accuracy and Evidence-Based Guidance	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.
	Rootedness in Student's Draft	It is deeply rooted in the student's draft.	It is deeply rooted in the student's draft.	It is deeply rooted in the student's draft.	It is deeply rooted in the student's draft.	It is deeply rooted in the student's draft.
Sample5	**Alignment with the Paper Rubric**	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.	The feedback generally aligns with the instructor's rubric, with one rubric item (R2) does not contain suggestions to include short-term and long-term actions.
	Accuracy and Evidence-Based Guidance	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.	The comments contain significant inaccuracies, as the draft contains no revision at all.	It provides detailed guidance on how to strengthen arguments through the effective use of evidence.

Sample#	IDA+A Criteria	CFB1	CFB2	CFB3	CFB4	CFB5
	Rootedness in Student's Draft	It is somewhat rooted in the student's draft, as Charlie keeps commenting in R5 to include headings for each writing section, while the headings are there.	It is somewhat rooted in the student's draft, as Charlie keeps commenting in R5 to include headings for each writing section, while the headings are there.	It is somewhat rooted in the student's draft, as Charlie keeps commenting in R5 to include headings for each writing section, while the headings are there.	• It mentions the author's name • It keeps generating feedback despite no revisions	It is somewhat rooted in the student's draft, as Charlie keeps commenting in R5 to include headings for each writing section, while the headings are there.

2. RQ2: How well do students integrate feedback to refine their drafts for final submission?

Quantitative and qualitative methods were used to determine how well students incorporated Charlie's feedback into their final draft submitted for grading. First, the quantitative assessment contains the overall scores of the resubmitted drafts. Table 5 provides the note summary of the essay scores documented by the human evaluator. It provides quantitative results summarizing the scores rated according to the instructor-designed rubric. The essay scores present show the comparison between human feedback (HFB) and Charlie's feedback (CFB). In general, the essay scores assessed based on CFB range from 10.1 to 11.8 points, while the scores graded by the human evaluator range from 8.2 to 16, indicating that students integrated Charlie's feedback well.

Also, we can see that the overall trend of HFB data shows increasing scores between submissions. However, the scores for the essay submitted for more than three times remain the same after the second submission, indicating no further revision was made to the drafts. The CFB data show fluctuation, particularly for the drafts resubmitted by Sample 3 and Sample 4, where the scores for the penultimate drafts are slightly decreased.

Furthermore, two samples (i,e., Sample 2 and Sample 5) marked as "not available (NA)" in the second, third, and fourth columns of HFB scores, indicating no revisions were made for such samples. Whereas on the CFB side, those two samples have scores for all iterations. These circumstances suggest that Charlie keeps generating feedback even though the students did not revise their drafts. The highlights that Charlie will always generate feedback with every submission and could be perceived as never being satisfied.

Table 5: Essay Scores

	Number of drafts	**ESSAY SCORES BASED ON:**									
		HFB-1	**HFB-2**	**HFB-3**	**HFB-4**	**HFB-5**	***CFB-1***	***CFB-2***	***CFB-3***	***CFB-4***	***CFB-5***
Sample1	2	13.6	16	X	X	X	*10.7*	*11.8*	*X*	*X*	*X*
Sample2	3	11.9	NA	NA	X	X	*11.1*	*10.4*	*12*	*X*	*X*
Sample3	4	8.2	10.2	11.2	11.2	X	*10.1*	*11.2*	*10.8*	*11.2*	*X*
Sample4	5	13.5	13.7	13.7	13.7	13.7	*11.7*	*12.3*	*12.3*	*11.7*	*11.8*
Sample5	5	11.2	12.5	13.9	NA	13.9	*10*	*10.6*	*11.9*	*11.9*	*13.2*

To further evaluate whether the students incorporated the feedback well into their final draft, we reviewed all rationales made by the human evaluator. We classified them into a category matrix with two labels named "changes made for" and "no changes made for" to imply which rubric criteria the students revised their drafts. Table 6 presents the matrix we used to organize the categories to see what sort of changes were made to each draft resubmission.

Overall, we can see that fewer changes were made to the resubmitted drafts, as none were revised based on the feedback provided related to the fifth rubric criteria. Besides, in the "no changes made for" category, we see that all sample drafts fall within the rubric criteria. In

addition, it is evident that Sample 2, which Charlie kept providing feedback across its two revisions, did not result in any meaningful changes to the resubmitted essay. This further reveals that Charlie cannot recognize whether the subsequent submissions have been revised. Moreover, if we further look at the "no changes made for" category, we can quickly notice that Sample 5 did not modify its fourth submission. This further confirms that Charlie keeps generating feedback even though the students did not revise their drafts. Furthermore, we can see that Sample 4 had the fewest revisions because it revised the career goals only in its second draft. On the contrary, Sample 5 made the most revisions, although its fourth draft contained no changes.

Table 6: Category Matrix

	CHANGES MADE FOR				
	1. Career Goals	**2. Relevance & Realism of Action Plan**	**3. Self-reflection & Insight**	**4. Feasibility & Resources**	**5. Organization & writing quality**
Sample1	D2	D2	D2	D2	
Sample2					
Sample3	D3	D2 D3	D2	D3	
Sample4	D2				
Sample5	D2 D3	D2 D3	D3	D2	
	NO CHANGES MADE FOR				
	1. Career Goals	**2. Relevance & Realism of Action Plan**	**3. Self-reflection & Insight**	**4. Feasibility & Resources**	**5. Organization & writing quality**
Sample1					D2
Sample2	D2	D2	D2	D2	D2
Sample3	D2 D4	 D4	 D3 D4	D2 D4	D2 D3 D4
Sample4	 D3 D4 D5	D2 D3 D4 D5	D2 D3 D4 D5	D2 D3 D4 D5	D2 D3 D4 D5
Sample5	 D5	 D5	D2 D5	 D3 D5	D2 D3 D5

To summarize, the finding reveals that students generally did integrate feedback well to refine their essay drafts for final submission. However, the extent and effectiveness of this integration

varied across samples, with some students making significant improvements and others making minimal or no changes.

V. Discussion

This study aimed to address the challenge of providing rich formative feedback to students in large classrooms, particularly for complex writing assignments such as essays, reports, and project proposals. The integration of the AI tool "Charlie" was explored to support students' technical writing abilities by providing timely and relevant feedback on their essays on their goals and action plans for their professional careers. Our research goals were to develop a research method to analyze the quality of the feedback and provide indicators of change in students' writing.

The analysis methods we used defined a useful process for characterizing the feedback provided by an AI tool like Charlie. We found several interesting results on Charlie's feedback, including effectiveness, student use of feedback, and comparisons of human and machine feedback.

The study found that Charlie's feedback generally aligned well with the instructor's rubric, providing detailed and evidence-based guidance. This kind of feedback helps students notice what they missed in the assignment criteria, and the suggestions could help them make changes.

The analysis revealed that students generally integrated Charlie's feedback well into their final drafts, particularly for refining their career goals and action plan sections. The extent of integration varied, with some students making significant improvements and others making minimal changes. The study highlighted the importance of iterative feedback and the potential for AI tools to support this process. One potential reason for the spread could be the assignment criteria to Charlie mandated only one submission. Therefore, students' decision to continue refining their document could be based on minimizing efforts to complete the assignment and a desire to get the best work product.

The results indicated that Charlie's feedback was comparable to human feedback in terms of quality and effectiveness. This finding suggests that AI-generated feedback can be a valuable supplement to traditional feedback methods, particularly in large classroom settings.

Charlie will always generate feedback, which is analogous to a very "strict professor," as one student put it. Charlie is never satisfied. Charlie does not behave like an instructor who could, with time, have conversations with a student to discuss the evolution of ideas and how to communicate them. The instructor could engage in metacognitive reflection on a students' work product to discuss their rationale for what to include and what not to include in the paper. In this sense, students must learn to regulate their own processes to know when enough is good enough. This also highlights that Charlie is not an automatic grader. The continual generation of feedback leads to a flat rating scale. A human can notice the change across submissions to notice the level of improvement leading to a higher grade. Therefore, Charlie is more of a reflection tool to help students notice what is good and what is missing from work. Charlie can provide suggestions on what to change, but it is up to the student to make the changes to suit the needs and context of their writing task.

The study's findings have several important implications for the use of AI tools in education. The use of AI tools like Charlie can significantly reduce the workload for instructors by automating the feedback process. This scalability is particularly beneficial in large classrooms where providing individualized feedback can be challenging. AI-generated feedback ensures consistency and objectivity, as it is based on predefined criteria and does not vary between different evaluators. This can help maintain a standard level of feedback quality across all students. The ability of AI tools to provide timely feedback encourages students to engage in iterative cycles of writing and revision. This iterative process is crucial for developing strong writing skills and improving the overall quality of student work.

While Charlie's feedback was generally effective, the study identified areas for improvement, such as recognizing revisions and providing more nuanced feedback. Future developments could focus on enhancing these aspects to further improve the tool's effectiveness. Also, Charlie is not a substitute for humans but rather an assistant who increases students' preparedness to submit a quality response to an assignment. Charlie is only aware of the assignment criteria and does not know the full context of the assignment and other implicit understandings about the course and the reasons for the assignment. Ultimately, the human needs to be the final evaluator of the work provided by the students.

VI. Conclusion

The integration of AI tools like Charlie in educational settings holds significant promise for enhancing the quality and efficiency of formative feedback. By providing timely, consistent, and detailed feedback, these tools can support students in developing their technical writing skills and achieving better learning outcomes.

The analysis methods used in this study are effective at characterizing the feedback provided by Charlie. The methods will be replicated in future studies that involve students' project proposals and final reports. This study also helped identify nuances for using GenAI tools in this context, which will inform the design of assignments and how Charlie is introduced to the students. This could include more explicit training to help students with metacognitive processes as they attempt to transform Charlie's feedback into effective changes in their writing samples.

References

[1] Dannels, D. P., Anson, C. M., Bullard, L., & Peretti, S. (2003). Challenges in learning communication skills in chemical engineering. Communication Education, 52(1), 50-56.

[2] Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), e3292.

[3] Moskovitz, C., & Kellogg, D. (2005). Primary science communication in the first-year writing course. College Composition & Communication, 57(2), 307-334.

[4] Crossman, J. M., & Kite, S. L. (2012). Facilitating improved writing among students through directed peer review. Active Learning in Higher Education, 13(3), 219-229.

[5] Brookhart, S. M. (2017). *How to give effective feedback to your students*. ASCD.

[6] Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57.

[7] Xu, W., Ouyang, F. (2022) The application of AI technologies in STEM education: a systematic review from 2011 to 2021. *IJ STEM Ed* **9**, 59 . https://doi.org/10.1186/s40594-022-00377-5

[8] Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, 1-23.

[9] Cai, C., Zhu, G. & Ma, M. (2024) A systemic review of AI for interdisciplinary learning: Application contexts, roles, and influences. *Education Information Technology*. https://doi.org/10.1007/s10639-024-13193-x

[10] Mustafa, M.Y., Tlili, A., Lampropoulos, G. *et al.* (2024). A systematic review of literature reviews on artificial intelligence in education (AIED): a roadmap to a future research agenda. *Smart Learn. Environ.* **11**, 59 https://doi.org/10.1186/s40561-024-00350-5

[11] Institutional Data Analytics + Assessment (IDA+A) (2024) "White Paper - LLM Quality Evaluation 3,".

[12] Klaus. Krippendorff (2004). *Content analysis: An introduction to its methodology*. Sage.

[13]M. Schreier, *Qualitative content analysis in practice*. Sage, (2012). [Online]. Available: www.sagepub.co.uk/schreier

[14] E. R. Babbie, *The Practice of Social Research*. Cengage Learning, 2020.

[15] Neuman, W.(2014). *Social Research Methods: Qualitative and Quantitative Approaches*, 7th ed. Pearson.

[16] Creswell, J. W. and Creswell, D. J. (2018) *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th ed. SAGE Publications, Inc., 2018.