Using Large Language Models to Assess Young Students' Writing Revisions

Tianwen Li[1], **Zhexiong Liu**[2], **Lindsay Clare Matsumura**[1], **Elaine Lin Wang**[3]
Diane Litman[1,2], **Richard Correnti**[1]
[1]Learning Research and Development Center, University of Pittsburgh
[2]Department of Computer Science, University of Pittsburgh
[3]RAND Corporation, Pittsburgh, PA 15260 USA
{tianwen.li,lclare,dlitman,rcorrent}@pitt.edu
zhexiong@cs.pitt.edu, ewang@rand.org

Abstract

Although effective revision is a crucial component of writing instruction, few automated writing evaluation (AWE) systems specifically focus on the quality of the revisions students undertake. In this study, we investigate the use of a large language model (GPT-4) with Chain-of-Thought (CoT) prompting for assessing the quality of young students' essay revisions aligned with the automated feedback messages they received. Results indicate that GPT-4 has significant potential for evaluating revision quality, particularly when detailed rubrics are included that describe common revision patterns shown by young writers. However, the addition of CoT prompting did not significantly improve performance. Further examination of GPT-4's scoring performance across various levels of student writing proficiency revealed variable agreement with human ratings. The implications for improving AWE systems focusing on young students are discussed.

1 Introduction

The ability to write is foundational to academic success. Yet, national assessments show that nearly three-quarters of students in the United States are not proficient writers (NCES, 2012). A well-recognized approach for improving students' writing skills is to engage students in cycles of revising their essays in response to formative feedback (Graham and Perin, 2007; Graham and Sandmel, 2011). However, students rarely receive substantive formative feedback on their writing for multiple reasons. First, teachers can be reluctant to assign writing tasks that require students to work across drafts because providing formative feedback is time-consuming (Graham et al., 2014). Second, teachers can feel unsure about how to provide feedback to improve students' essay quality (Brindle et al., 2016). Finally, research shows that teachers are inconsistent in their feedback practices, and tend to focus on surface-level features of students' writing rather than the content of students' ideas and reasoning (Matsumura et al., 2002, 2023).

Automated Writing Evaluation (AWE) systems are gaining prominence as one approach to increasing students' opportunity to receive formative feedback. While research suggests that teachers generally respond positively to AWE systems and can see them as helpful time savers (Grimes and Warschauer, 2010; Palermo and Thomson, 2018), evidence is modest that AWE systems improve the quality of students' writing in the elementary and secondary grades (Graham et al., 2015). One reason why students' writing may not improve in response to automated feedback is that they often lack the skills necessary for effective revision (Roscoe et al., 2013; Wang et al., 2020). Wang et al. (2020) found that only 18% of students successfully implemented the feedback messages they received from an AWE system. For example, when asked to provide more evidence for their claims, students commonly repeated the examples that they had cited before. This highlights the importance of providing students with feedback that builds their revision skills, in addition to feedback that improves their writing quality.

Given that formative assessment fosters writing skill development by establishing and reinforcing clear criteria for successful writing (Matsumura et al., 2023), it is notable that few assessments target students' revision skills. Building on the previous discussion about the necessity of teaching students how to revise, we believe that formative assessments that precisely establish the criteria for effective revision can provide information to students and teachers about the extent to which

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 365–380

revision goals are met and offer guidance for implementing revision feedback. To address this gap, our team developed a rubric for holistically assessing revision quality (Wang et al., 2020). By 'revision quality', we specifically examine whether revisions students made were aligned with the feedback provided, and the extent to which it improved the essay with respect to evidence use. This is in contrast to revisions that may improve essay quality in ways not aligned with the content of feedback messages.

In the context of AWE systems, automatically assessing the revision process is a necessary area of development. Most systems have focused on assessing overall improvement in essay quality. Although these systems can detect revisions, they tend to assign scores or provide feedback based on the overall essay quality, rather than attend to the quality of the revisions undertaken (Foltz and Rosenstein, 2017; Mayfield and Butler, 2018). Recent advancements in large language models (LLMs) show significant promise for analyzing and evaluating student revision quality. GPT-4, standing out among these models, specifically has been shown to generate scores that are comparable to those given by human evaluators (Mizumoto and Eguchi, 2023; Naismith et al., 2023; Tate et al., 2023; Xia et al., 2024; Xiao et al., 2024). While most of these studies have concentrated on GPT-4's ability to assess writing quality, our study extends previous research by investigating the effectiveness of GPT-4 for evaluating revision quality with different prompting strategies. Given that students often find essay revision challenging, it is essential to provide a revision score that reflects diverse revision patterns. This study represents an initial step in exploring GPT-4's capability to score revisions, setting the stage for offering personalized feedback on students' revision practices in future research.

In this study, we specifically explore GPT-4's performance in assessing the revision attempts of young students (ages 10 to 12) who often exhibit less structured and sophisticated writing styles. Given that most existing research concentrates on evaluating essays by adolescents and adults (e.g., Naismith et al., 2023; Xiao et al., 2024), it is of interest to explore how GPT-4 adapts to the writing of younger age groups. In addition, as students may display a wide range of writing proficiency, it is crucial to ensure that GPT-4 does not exhibit systematic biases that could compromise scoring accuracy.

Two research questions are addressed:

1. How accurately can GPT-4 assess the revision quality of students' argumentative writing in comparison with human raters?
2. How does GPT-4's performance in evaluating revisions vary across different levels of students' argumentative writing abilities?

2 Data

In this section we describe the dataset of students' essays, the rubric used for assessing students' revision quality, and the process for evaluating these revisions by human raters.

2.1 RTA space dataset

The corpus for our investigation is drawn from a study of eRevise, an AWE system designed to improve students' argumentative writing in the fifth and sixth grades (Correnti et al., 2022; Zhang et al., 2019). eRevise was designed to score responses and provide feedback to students on the Response-to-Text Assessment (RTA). The RTA aims to assess the quality of students' ability to reason about texts in their writing and to use text evidence to support their claims (Correnti et al., 2012; Correnti et al., 2013). The form of the RTA used in this study is based on a non-fiction article about government funding for space exploration (RTA_{Space}). To administer the RTA, a teacher reads the text aloud to students as they follow along with their copy of the article. The teacher also poses planned questions at certain points in the articles and defines some vocabulary words to ensure that all students comprehend the article in advance of writing. Students respond to the following prompt:

> Consider the reasons given in the article for why we should and should not fund space exploration. Did the author convince you that "space exploration is desirable when there is so much that needs to be done on Earth"? Give reasons for your answer. Support your reasons with 3-4 pieces of evidence from the text.

After students submit their first drafts, the system uses NLP features generated during the automatic scoring of students' initial essays (including the number of pieces of evidence, specificity of evidence, concentration of evidence, and word count) to select formative feedback on evidence. There

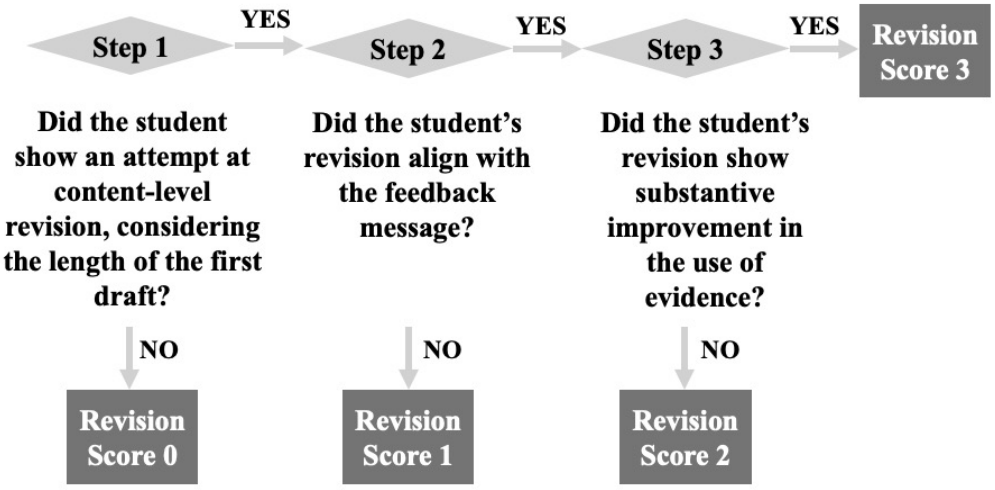

Figure 1: Human rater evaluation steps

are three levels of feedback (Appendix A). Feedback Level 1 focuses on completeness (i.e., guides students to provide more evidence) and guides students to be more specific about the evidence they reference. Feedback Level 2 also directs students to be more specific, in addition to explaining their evidence. Finally, Feedback Level 3 guides students to explain their evidence and connect it to their overall argument (Correnti et al., 2020; Wang et al., 2020). After receiving the tailored feedback, students make revisions to their essays accordingly.

The RTA$_{Space}$ dataset contains a total of 600 essay pairs, which include both initial and revised essay drafts, collected from thirty-four fifth and sixth-grade ELA teachers in Louisiana who participated in the study during the 2018-2019 school year.

2.2 Human assessment of students' revision quality

Our team developed a holistic rubric to assess revision quality based on a detailed qualitative analysis of how fifth and sixth graders applied the automated feedback they received (Wang et al., 2020). We identified four levels of revision: 0 = No attempt at implementing feedback; 1 = Attempted to implement feedback, but no improvement in evidence use; 2 = Slight improvement in evidence use; 3 = Substantive improvement in evidence use. These four levels of revision were further transformed into a sequential flow of reasoning steps that guide human raters' scoring process (Figure 1). In addition, since initial drafts were categorized into three levels, each offering different focuses for revision, the ways in which students attempt to apply the feedback could vary. As a result, beyond the four abstract criteria used to assess the quality of revisions, the rubric was supplemented by specific, frequently observed patterns identified by human raters at each revision score (Appendix B).

For example, if a student receives Feedback Level 1 which focuses on the completeness and specificity of evidence, a successful revision (score 3) involves adding more than one new piece of evidence from the text that was not previously mentioned. A revision score of 2 is assigned when students repeat the same evidence already provided or a score of 1 is given if they fail to align their changes with the Feedback Level 1 messages; for example, instead of introducing new evidence they only provide explanations for the evidence they had used in their first draft. Feedback Level 2 focuses on the specificity and the elaboration of existing evidence; thus, a revision score of 3 is assigned if students add significant detail or explanation to more than one piece of evidence. Conversely, a score of 2 is assigned if students merely paraphrase the existing evidence, and a score of 1 is applied if students, contrary to the focus of the feedback message, add new evidence instead of elaborating on their existing evidence. Feedback Level 3 emphasizes explaining existing evidence and its connection to a claim. A revision score of 3 is assigned if students provide a strengthened explanation for more than one piece of evidence. A less successful revision may result from offering relatively brief or

repetitive explanations (score 2), or from misalignment with the feedback message (score 1). This would be shown, for example, by students merely elaborating on their evidence without effectively connecting it to their claim. These detailed patterns associated with each score thus provide a nuanced guide for humans evaluating revisions.

To evaluate the quality of revisions, human raters began by identifying the changes students made to their essays. Each pair of essays, consisting of the initial and revised versions, was placed in separate Word documents. By using the "Compare Documents" feature in Word, the document highlighted areas where students added, deleted, or modified text. Then, taking into account the feedback level of the initial draft, human raters used the revision rubric (Appendix B) to determine the revision score.

Three human raters engaged in the evaluation process, which was divided into two phases. In the first phase, the primary rater, who played a crucial role in developing the rubric, trained the second rater to score the first 300 essay pairs. Sixty essay pairs were randomly selected from the three feedback levels and were coded by both raters. The interrater agreement for these pairs was 82% for exact matches and a Quadratic Weighted Kappa (QWK) of 0.74, demonstrating substantial consistency. In the second phase, the second rater, now experienced, trained the third rater to assess the remaining 300 essay pairs. This time, 30 essay pairs selected from the three feedback levels were double-coded for calibration. The interrater agreement reached 83% for exact matches and a QWK of 0.75, which again indicated a substantial level of reliability. The distribution of human revision scores at each feedback level is shown in Table 1.

	Revision Score 0 N (%)	**Revision Score 1** N (%)	**Revision Score 2** N (%)	**Revision Score 3** N (%)
Feedback Level 1	36 (26.67%)	40 (29.63%)	42 (31.11%)	17 (12.59%)
Feedback Level 2	53 (17.15%)	119 (38.51%)	104 (33.66%)	33 (10.68%)
Feedback Level 3	29 (18.59%)	56 (35.90%)	54 (34.62%)	17 (10.90%)

Table 1: Distribution of human revision scores at each feedback level

3 Experimental design

3.1 Experiment 1: Zero-shot prompt design (Baseline model)

In the initial experiment, we assessed GPT-4's capability in evaluating the quality of students' revisions to their text-based argumentative essays. The prompt was structured in the following order (see Appendix C for the prompt details):

1. Scoring task: This section outlined a clear scoring task for GPT-4. It introduced the stages where students were in their text-based argumentative writing tasks, having completed their first draft and then finished their second draft based on the feedback received. The feedback messages provided to students were incorporated into the prompt.
2. Writing task: This section introduced the text that formed the basis for the students' essays. The writing prompt was also included.
3. Detailed scoring rubric: The aforementioned revision rubric with the concrete revision patterns was included.
4. Student first and second drafts: To assess the quality of revisions, both the first and second drafts of student essays were provided.

3.2 Chain-of-Thought prompt design

We tested two different strategies of Chain-of-Thought (CoT) for improving the performance of GPT-4.

Experiment 2: One-shot CoT with human rater rationale

We provided GPT-4 with one example for each feedback level, all identified as successful revisions (holistic score of 3), accompanied by the human raters' rationale for their ratings (Appendix D). Considering that all essays came from fifth and sixth graders who were in the process of learning how to write argumentative essays, including successful revision examples in the prompt can aid GPT-4 in adjusting its scoring to reflect a more appropriate standard for young learners as opposed to the more advanced revisions that would be expected of adults. By presenting the rationale of human raters, our goal was to instruct GPT-4 to follow intermediate reasoning steps that human raters would apply. We further asked GPT-4 to provide a rationale for scoring before giving its score with the aim of eliciting a chain of reasoning.

Experiment 3: One-shot CoT with intermediate steps

To improve GPT-4's ability to use the rubric effectively, the rubric was transformed into a sequential flow of reasoning steps. This approach aimed to guide GPT-4 through the evaluation process in a step-by-step manner, closely simulating the decision-making pathway used by human raters (Figure 1). In addition, we also provided one example of successful revision for each feedback level in the prompt to support GPT-4 to adjust its scoring to reflect an appropriate evaluation standard for young students. We further asked GPT-4 to provide a rationale before giving its score with the aim of eliciting a chain of reasoning.

4 Results

4.1 Research question 1: How accurately can GPT-4 assess the revision quality of students' argumentative writing in comparison with human raters?

We conducted three experiments employing GPT-4 combined with CoT prompting strategies to assess their effectiveness in predicting the holistic scores for writing revision quality. Our primary evaluation metrics were Quadratic Weighted Kappa (QWK), which are widely used in automated essay scoring (AES) tasks.

	Zero-Shot	**One-Shot CoT** (Human rationales)	**One-Shot CoT** (Intermediate steps)
Exact Agreement	52.00%	54.50%	36.33%
Quadratic Weighted Kappa	0.60	0.60	0.46

Table 2: Overall revision score agreement rate

In the initial zero-shot prompting experiment, which served as our baseline, we observed an exact agreement rate of 52.00% and a QWK of 0.60, which suggested a moderate level of agreement between human raters and GPT-4 (Table 2). In our second experiment, we introduced a single example of a successful revision (revision score 3) along with the human rationale for that score at each feedback level. This approach improved the exact agreement rate to 54.50% while the QWK remained unchanged. Overall, by applying detailed rubrics with specific and concrete revision patterns corresponding to each score, GPT-4 demonstrated notable potential for assessing the quality of student revisions. However, while many studies indicate that including examples with human rating rationales greatly outperforms baseline models (e.g., Xia et al., 2024; Yancey et al., 2023), our second experiment only found a slight improvement in the exact agreement between human raters and GPT-4 when the one-shot CoT was applied.

Furthermore, the rubric used in the baseline and second experiment was developed from observations made by human raters adhering to the scoring procedure. As the rubric only contains the most common revision patterns under each revision score, the rubric may not capture the full depth of our evaluation criteria for student revision quality. Thus, we introduced a structured three-step scoring process as a novel form of Chain-of-Thought to assess whether GPT-4 could mimic the human thinking process during complex tasks. However, this approach yielded a significant decrease in agreement rates. Specifically, as shown in the third column in Table 2, the exact agreement rate decreased to 36.33%, while the QWK dropped to 0.46. The outcomes implied that a rubric with clearly defined patterns for student revisions outperforms the more explicit but abstract scoring process used by human raters.

4.2 Research question 2: How does GPT-4's performance in evaluating revisions vary across different levels of young students' argumentative writing abilities?

We further explored the extent to which the level of agreement between GPT-4 and human raters varied with students' argumentative writing skills. As previously described, we categorized students' initial drafts into three levels based on the number of pieces of evidence, specificity of evidence, concentration of evidence, and word count. Students with Level 1 drafts were advised to improve their writing by adding more evidence, while those with Level 2 and 3 drafts were guided towards more advanced revisions centered on the elaboration and explanation of the evidence provided. From Table 3, it's evident that GPT-4 exhibits a markedly higher level of agreement with human scoring when assessing revisions in Level 1 essays, a pattern that persists across all three prompting strategies. Especially when one-shot CoT prompting is applied, we observed a notable enhancement in the precision of scoring predictions for Level 1 essays in contrast to Level 2 and Level 3, with the exact agreement

	Zero-Shot			One-Shot CoT (human rationales)			One-Shot CoT (intermediate steps)		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Exact Agreement	60.00%	47.90%	53.21%	65.93%	50.16%	53.21%	55.56%	30.74%	30.77%
Quadratic Weighted Kappa	0.73	0.53	0.58	0.77	0.54	0.54	0.71	0.38	0.43

Table 3: Revision score agreement rate at each feedback level

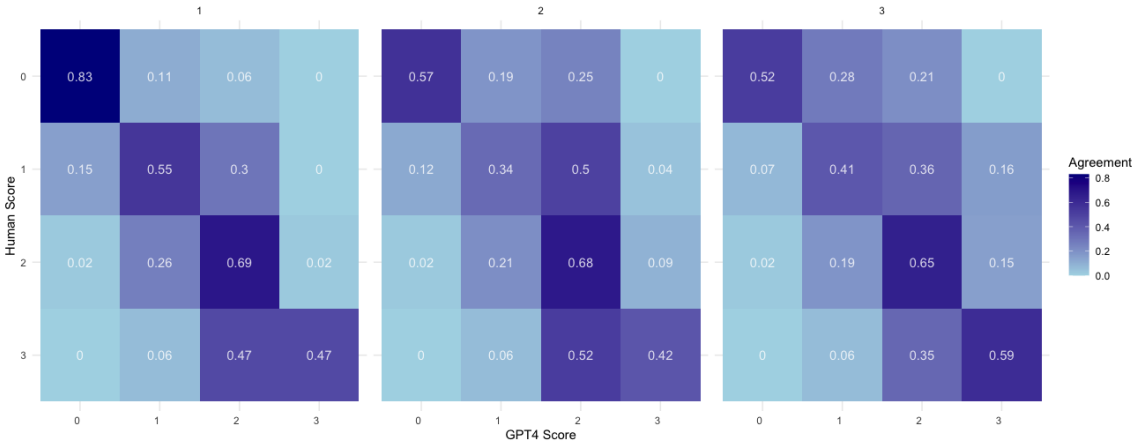

Figure 2: Confusion matrices of one-shot CoT prompting at each feedback level

increase from 60.00% to 65.93%, and the QWK from 0.73 to 0.77. This result suggests that GPT-4 is more likely to accurately evaluate the more concrete and straightforward task of adding evidence compared to evaluating evidence elaboration and explanation.

In contrast, the revision score agreement for Level 2 is lower than for Levels 1 and 3 across all three prompting strategies. Students with Level 1 or Level 3 essays were guided to focus exclusively on one aspect of revision: adding new evidence or adding explanations. Students with Level 2 drafts were in a middle position, as they were instructed not only to elaborate on the evidence but also to offer some explanations. When it comes to assessing the revision quality of draft 2, GPT-4 needs to examine revisions from two aspects, and this complexity may result in its inaccuracy. This result reemphasizes the potential limitations of GPT-4's accuracy in evaluating multifaceted tasks than simpler ones.

As the second experiment that applied one-shot CoT prompting demonstrated a relatively higher agreement among all three strategies, we focused on this condition for error analysis. Confusion matrices in Figure 2 reveal a strong consensus among humans and GPT-4 on the assignment of score 0 across all three levels, indicating no attempt at revision in the students' first drafts. Although the prediction of score 0 is highly accurate at Level 1, at Levels 2 and 3, despite being moderate, the accuracy of predicting score 0 diminished as GPT-4 tended to assign higher scores. A key factor could be that human raters might take into account the length of the student's initial draft when judging the revision effort, a nuance that GPT-4 might not effectively adjust for based on the student's writing proficiency.

Another noticeable trend is that GPT-4 tended to assign lower scores when human raters assigned a score of 3, consistent across all three levels (Figure 2). This discrepancy could stem from GPT-4's higher criteria for defining "a substantive improvement" in revisions. Table 4 provides an example from a Level 2 essay where the human rater assigned a score of 3 and GPT-4 assigned a score of 2. In the second draft, the student first improved one piece of evidence by adding a more relevant explanation of how providing money can contribute to better health. They also introduced new evidence regarding pollution issues, along with an explanation of how this supports their argument. Despite the repetition of ideas and less clear reasoning, the effort demonstrates a significant attempt at revision, as well as improvement in elaboration and explanation of existing evidence, from the perspective

First Draft	Second Draft (Student's additions to their essay are indicated in red font)
We readers should fund money to space exploratons . . . one reason for the readers wanting to give money to the people is so that they can have food and shelter for their family.according to the text it states(1)" nearly half of all americans also have difficulty paying for housing,food,and medicine at some point in their lives." " in other countries, people are dying because they do not have access to clean water,medical care,or so simple solutions that prevent the spread of diseases." (3)" for example,malaria, a disease spread by mosquito bites,kills many people in africa every year." this quotation shows how we readers should donate money to the people who are living an unhappy life.this quotation makes it clear that we readers feel that the people who are homeless feels more important than the space exploration because they are poor. another reason is that people should get	We readers should fund money to space exploratons. . . one reason for the readers wanting to give money to the space exploration is because they want the people to be healthier and have a better and successful life to raise their children.According to pharagraph 2,it states" nearly half of all americans also have difficulty paying for housing, (2)" people are dying because they have no food or clean water to drink, also,it states" people needed medical instruments to keep the diseases from spreading and learn and develop body's reaction area's." this quotation shows how my evidence makes it seem important that you should give money to people who are homeless and need to learn about medica instrument so they can clear their diseases and sickness.this quotation makes it clear that people would stay healthy by using medical instruments to cure their sickness and disease. another reason is that people need money so that they can clean and help earth stay healthy. according to pharagraph 3,it states "(1) many scientist believe that pollution from burning fossils fuels is harming our air and oceans." " we need new,cleaner forms of energy to power cars,homes, and factories." " a program to develop clean energy could be viewed as a worthy investment." this quotation shows how my evidence explains why space explorations also should still donate money to people so they can help earth get cleaned and to power factories and cars and also homes. this makes it clear that my evidence supports my reasoning state and also supports my claim.

Table 4: Example of student revision at feedback level 2

of a fifth or sixth-grader at least. In other words, humans appear more likely to consider students' developmental level when scoring, a consideration that GPT-4 may overlook.

5 Discussion and conclusions

Revising is a very difficult skill to master, and many young students struggle to implement the feedback they receive (Roscoe et al., 2013). To foster the development of students' revision skills, assessing revision quality and identifying revision patterns across various levels of writing proficiency is essential for providing targeted feedback to students on their revision efforts. With this aim, this study explored the potential of using a large language model, specifically GPT-4, to evaluate the quality of essay revisions aligned with the feedback messages students received from an AWE system.

First, our results suggest that GPT-4 has a great deal of potential for effectively evaluating writing revision quality. We used a detailed rubric providing specific revision patterns in the zero-shot (baseline experiment) prompting and one-shot CoT prompting and both approaches showed a moderate level of agreement between human raters and GPT-4. However, both CoT prompting strategies implemented in the study did not improve GPT-4 baseline performance. It is not altogether clear why this was the case as other researchers have found that CoT prompting tends to improve the accuracy of writing quality scores (Xia et al., 2024; Yancey et al., 2023). We note, however, that evaluating the quality of revisions in younger students' essays may be a more complex task than assessing overall quality. It contains a series of evaluative steps beyond simply identifying revision patterns with a rubric. This includes interpreting feedback messages, identifying what was added in second drafts, and evaluating the alignment of those additions to the feedback. We recommend that future research explore additional prompting strategies to better address this complexity. For example, Tree-of-Thoughts prompting, which encourages LLMs to explore various ideas and assess intermediate steps in order to provide an optimal response (Yao et al., 2024), could be a useful way forward for generating more accurate assessments of complex writing processes.

Secondly, unlike studies that focus on adult writers such as college students, our research provides insight into the capabilities of LLMs to assess the writing produced by young students. We observed that GPT-4 tended to assign lower scores to revisions than human raters. One reason for this might

be that fifth and sixth graders are still in the midst of developing their language as well as reasoning skills. The changes they make to their essays are constrained then by their overall ability to elaborate and explain their thinking in writing. Human raters took into account the age of students, and what they deemed reasonable to expect for revision at that age, and gave credit for effort (incremental changes) rather than only the quality of students' final product. Unlike human raters then, GPT-4 may lack knowledge of developmentally appropriate expectations for student writing which potentially affects its scoring accuracy. Therefore, LLMs would benefit from tailored training to adjust their criteria for "good" writing to be calibrated for different-aged students.

Limitations

Future research should consider the reliability of human ratings when evaluating GPT-4 scoring quality. While human raters remain the "gold standard" of writing evaluation, they are not always particularly consistent with one another (Brown, 2009; Cohen et al., 2018). In this study, we calculated only the overall reliability across three feedback levels among human raters, without specifically assessing the reliability at each feedback level. Further research is necessary to explore how human raters' scoring accuracy may vary across different levels of writing proficiency and within various scoring tasks, as well as how the reliability of human raters may influence the accuracy of automated scoring systems.

Moreover, this study focuses solely on exploring the potential of GPT-4, using it as an example among LLMs, for evaluating the quality of student revisions. Although GPT-4 has demonstrated impressive capabilities in various writing assessment tasks, alternative large language models, such as those outside the GPT family, may yield different results. Future research should investigate other LLMs, which would offer a more comprehensive understanding of the effectiveness of LLMs in assessing writing revisions.

Ethics statement

In our research, we assert that the dataset applied poses minimal risk regarding potential harm to individuals. The collection of writing essays from fifth and sixth-grade ELA teachers in Louisiana received approval from the Institutional Review Board at our institution. The intentions, procedures, and methods for collecting and using student essays are thoroughly detailed in the required consent forms provided to our teacher participants. Additionally, all collaborators adhere to stringent data privacy policies, ensuring further protection of participant information. Furthermore, our team is dedicated to enhancing participants' awareness and understanding of AWE systems. We meticulously developed interview questions aimed at uncovering their perceptions and concerns regarding the use of AWE technology. We also respect and value their suggestions on improving the design of the AWE systems to ensure they are intuitive, easy, and safe to use.

Acknowledgements

The research was supported by the National Science Foundation Award #2202347. The opinions expressed are those of the authors and do not represent the views of the foundation.

References

Mary Brindle, Steve Graham, Karen R Harris, and Michael Hebert. 2016. Third and fourth grade teacher's classroom practices in writing: A national survey. *Reading and Writing*, 29:929–954.

Gavin Thomas Lumsden Brown. 2009. The reliability of essay scores: The necessity of rubrics and moderation. *Tertiary assessment and higher education student outcomes: Policy, practice and research*, pages 40–48.

Yoav Cohen, Effi Levi, and Anat Ben-Simon. 2018. Validating human and automated scoring of essays against "true" scores. *Applied Measurement in Education*, 31(3):241–250.

Richard Correnti, Lindsay Clare Matsumura, Laura Hamilton, and Elaine Wang. 2013. Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, 114(2):142–177.

Richard Correnti, Lindsay Clare Matsumura, Laura S Hamilton, and Elaine Wang. 2012. Combining multiple measures of students' opportunities to develop analytic, text-based writing skills. *Educational Assessment*, 17(2-3):132–161.

Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, Diane Litman, Zahra Rahimi, and Zahid Kisa. 2020. Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, 55(3):493–520.

Richard Correnti, Lindsay Clare Matsumura, Elaine Lin Wang, Diane Litman, and Haoran Zhang. 2022. Building a validity argument for an automated writing evaluation system (erevise) as a formative assessment. *Grantee Submission*, 3.

Peter W Foltz and Mark Rosenstein. 2017. Data mining large-scale formative writing. *Handbook of learning analytics*, 199.

Steve Graham, Andrea Capizzi, Karen R Harris, Michael Hebert, and Paul Morphy. 2014. Teaching writing to middle school students: A national survey. *Reading and Writing*, 27:1015–1042.

Steve Graham, Michael Hebert, and Karen R Harris. 2015. Formative assessment and writing: A meta-analysis. *The elementary school journal*, 115(4):523–547.

Steve Graham and Dolores Perin. 2007. Writing next-effective strategies to improve writing of adolescents in middle and high schools.

Steve Graham and Karin Sandmel. 2011. The process writing approach: A meta-analysis. *The Journal of Educational Research*, 104(6):396–407.

Douglas Grimes and Mark Warschauer. 2010. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment*, 8(6).

Lindsay Clare Matsumura, G Genevieve Patthey-Chavez, Rosa Valdés, and Helen Garnier. 2002. Teacher feedback, writing assignment quality, and third-grade students' revision in lower-and higher-achieving urban schools. *The Elementary School Journal*, 103(1):3–25.

Lindsay Clare Matsumura, Elaine Lin Wang, Richard Correnti, and Diane Litman. 2023. Tasks and feedback: An exploration of students' opportunity to develop adaptive expertise for analytic text-based writing. *Assessing Writing*, 55:100689.

Elijah Mayfield and Stephanie Butler. 2018. Districtwide implementations outperform isolated use of automated feedback in high school writing. In *International Conference of the Learning Sciences*, volume 2128.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403.

NCES. 2012. The nation's report card: Writing 2011.

Corey Palermo and Margareta Maria Thomson. 2018. Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, 54:255–270.

Rod D Roscoe, Erica L Snow, and Danielle S McNamara. 2013. Feedback and revising in an intelligent tutoring system for writing strategies. In *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*, pages 259–268. Springer.

Tamara P Tate, Jacob Steissa, Drew Baileya, Steve Grahamb, Daniel Ritchiea, Waverly Tsenga, Youngsun Moona, and Mark Warschauera. 2023. Can ai provide useful holistic essay scoring? *OSF Preprints*.

Elaine Lin Wang, Lindsay Clare Matsumura, Richard Correnti, Diane Litman, Haoran Zhang, Emily Howe, Ahmed Magooda, and Rafael Quintana. 2020. erevis (ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, 44:100449.

Wei Xia, Shaoguang Mao, and Chanjing Zheng. 2024. Empirical study of large language models as automated essay scoring tools in english composition_taking toefl independent writing task for example. *arXiv preprint arXiv:2401.03401*.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv preprint arXiv:2401.06431*.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, LC Matsmura, Emily Howe, and Rafael Quintana. 2019. erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9619–9625.

A Feedback focus corresponding to each feedback level

Feedback Level 1 (Completeness & Specificity):

- Use more evidence from the article (Completeness)
- Provide more details for each piece of evidence you use (Specificity)

Feedback Level 2 (Specificity & Explanation):

- Provide more details for each piece of evidence you use (Specificity)
- Explain the evidence (Explanation)

Feedback Level 3 (Explanation & Connection):

- Explain the evidence (Explanation)
- Explain how the evidence connects to the main idea and elaborate (Connection)

B Rubric for assessing revision quality aligned with feedback message

Essay Level	**0—No Attempt** No content revision attempted	**1—Attempted, Not Aligned** Content revision attempted but not aligned with feedback message	**2—Aligned, Not Improved** Content revision aligned with feedback message but no/slight improvement in evidence use	**3—Aligned, Improved** Content revision improved evidence use in line with feedback message
Level 1	• No edits at all • Revision focused solely on writing mechanics. • Only several words added or changed.	• Student added evidence that is not directly related to the argument or text • Student provided explanation for evidence provided • Student elaborated on explanation they already attempted to provide. • Student connected evidence to argument	• Student added one relevant piece of evidence • Student added general discussion (without a specific quote or paraphrase) that supports the argument and is generally based in the text • Student added direct quotes to support paraphrases that were already there.	• Student added at least two relevant piece of evidence that are on the correct side of the argument
Level 2	• No edits at all • Revision focused solely on writing mechanics • Only a short line or two changed without significant content added.	• Student added evidence or details that are not directly related to the argument or text • Student added evidence, but did not add specificity (more details to evidence already provided) without any explanation • Student added empty explanation (i.e., "I included this evidence because it supports my point") • Student added explanations that did not connect to the argument or that contradict the argument • Student made minimal content-based edits of any sort considering the length of the entire essay	• Student added small details (at least 2 small instances) • Student added brief explanations of evidence (at least 2 small instances) • Student paraphrased existing evidence	• Student added relevant and solid details of evidence or explanations to at least two existing evidence
Level 3	• No edits at all • Revision focused solely on writing mechanics • Only a short line or two changed without significant content added	• Student added evidence or details that are not directly related to the argument or the text • Student added evidence or added more details to evidence without any explanation • Student added empty explanation (i.e., "I included this evidence because it supports my point") • Student added explanations that do not connect to the argument or that contradict the argument • Student made minimal content-based edits of any sort considering the length of the entire essay	• Student recycled same explanation for each piece of evidence • Student paraphrased existing evidence • Student only added one strong explanation for only one piece of evidence • Student added a decent explanation only at the end of the essay, not after each piece of evidence • Student added personal commentary, not explanation of evidence that connects to argument	• Student strengthened explanation for at least two pieces of existing evidence • Student provided strong connection between evidence presented to the overall argument

C GPT-4 prompt

Scoring task. 5th and 6th graders are learning how to write and revise text-based argumentative essays, particularly focusing on the use of evidence from the text. After they submit their first drafts, each student's work is assessed and categorized into levels—Level 1, Level 2, or Level 3—reflecting the quality of their writing. Based on the level their drafts are assigned, students receive corresponding feedback for Level 1, Level 2, or Level 3, which helps guide their revisions.

Level 1 feedback message concentrates on "Using more evidence from the article" and "Providing more details for each piece of evidence you use". Level 2 feedback message concentrates on "Providing more details for each piece of evidence you use" and "Explain the evidence". Level 3 feedback message concentrated on "Explain the evidence" and "Explain how the evidence connects to the main idea and elaborate".

Your role is to score the quality of revision from the first draft to the second draft based on a rubric that will be provided to you. The rubric comprises four ratings (0,1,2,3), focusing on evaluating whether students' revisions align with the feedback provided and if there is an improvement in their essays.

Writing task. This is the text the student needs to read before writing: A Question to Consider: Is space exploration really desirable when so much needs to be done on Earth? This is a question that has been asked for several decades and requires serious consideration. The arguments against space exploration stem from a belief that the money spent could be used differently – to improve people's lives. In 1953, President Eisenhower captured this viewpoint. He opposed the space program, saying that each rocket fired was a theft from citizens that suffered from hunger and poverty. Indeed, over 46.2 million Americans (15%) live in poverty. Nearly half of all Americans also have difficulty paying for housing, food, and medicine at some point in their lives. In other countries, people are dying because they do not have access to clean water, medical care, or simple solutions that prevent the spread of diseases. For example, malaria, a disease spread by mosquito bites, kills many people in Africa every year. It is possible to lower the spread of this disease by hanging large nets over beds that protect people from being bitten as they sleep. These nets cost only $5; however, most people affected by malaria cannot afford these nets. It is not just people that need help. The Earth is suffering also. Many scientists believe that pollution from burning fossil fuels (gasoline and oil) is harming our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. A program to develop clean energy could be viewed as a worthy investment. Maybe exploring space should not be a priority when there is so much that needs to be done on Earth. Right now, the government spends 19 billion dollars a year for space exploration. Some people think that this money should be spent instead to help heal the people and the Earth.

Tangible Benefits of Space Exploration: People in favor of space exploration argue that 19 billion dollars is not too much. It is only 1.2% of the total national budget. Compare this to the 670 billion dollars the US spends for national defense (26.3% of the national budget), or the 70 billion dollars spent on education (4.8% of the budget), or the 6.3 billion dollars spent on renewable (clean) energy. The investment in space exploration is especially worthwhile because it has led to many tangible benefits, for example, in the area of medicine. Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health under stressful conditions. This was to ensure the safety of the astronauts under harsh conditions, like those they would experience on launch and return. In doing this, medical instruments were developed and doctors learned about the human body's reaction to stress. In rising to meet the challenges of space exploration, NASA scientists have developed other innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting. All these resulted from technologies that NASA engineers developed to make space travel possible. Even the problems of hunger and poverty can be tackled by space exploration. Satellites that circle Earth can monitor lots of land at once. They can track and measure the condition of crops, soil, rainfall, drought, etc. People on Earth can use this information to improve the way we produce and distribute food. So, when we fund space exploration, we are also helping to solve some serious problems on Earth.

The Spirit of Exploration: Beyond providing us with inventions, space exploration is important for the challenge it provides and the motivation to bring out the best in ourselves. Space exploration

helps us remain a creative society. It makes us strive for better technologies and more scientific knowledge. Often, we make progress in solving difficult problems by first setting challenging goals, which inspire innovative work. Finally, space exploration is important because it can motivate beneficial competition among nations. Imagine how much human suffering can be avoided if nations competed with planet-exploring spaceships instead of bomb-dropping airplanes. We saw an example of this in the 1960's. During what is called the Cold War, the United States and Russia competed to prove their greatness in a race to explore space. They each wanted to be the first to land a spacecraft on the moon and visit other planets. This was achieved. It also resulted in many of the technologies and advancements already mentioned. In addition, the 'space race' led to significant investment and progress in American education, especially in math and science. This shows that by looking outward into space, we have also improved life here on Earth.

Returning to the Question All this brings us back to the question: Should we explore space when there is so much that needs to be done on Earth? It is true that we have many serious problems to deal with on Earth, but space exploration is not at odds with solving human problems. In fact, it may even help find solutions. Space exploration will lead to long-term benefits to society that more than justify the immediate cost.

This is the writing prompt: Consider the reasons given in the article for why we should and should not fund space exploration. Did the author convince you that "space exploration is desirable when there is so much that needs to be done on earth"? Give reasons for your answer. Support your reasons with 3-4 pieces of evidence from the text.

Scoring rubric with intermediate steps. We developed two types of rubric. The detailed rubric with concrete revision patterns would be introduced in Appendix C. The scoring rubric with intermediate steps was presented here:

Feedback Level 1. Step 1: Please compare the first draft and second draft, did the student show an attempt at content-level revision, considering the length of the first draft? If answer is no attempt or minimal attempt (including no edits at all, or only few words, revision focused solely on writing mechanics), please output score 0. Step 2: If yes, did the student's revision align with the feedback message, considering the text content? If answer is no (including that student provided explanation or elaborate on evidence for evidence provided), please output score 1. Step 3: If yes, did the student's revision show substantive improvement in the use of evidence? If answer is no improvement or slight improvement (including that student added one relevant piece of evidence, or student added direct quotes to support paraphrases that were already there), please output score 2. If yes (substantive improvement is that student added at least two solid and relevant piece of evidence that are on the correct side of the argument), please output score 3.

Feedback Level 2. Step 1: Please compare the first draft and second draft, did the student show an attempt at content-level revision, considering the length of the first draft? If answer is no attempt or minimal attempt (including no edits at all, or revision focused solely on writing mechanics, or only a short line or two changed without significant content added), please output score 0. Step 2: If yes, did the student's revision align with the feedback message? If the answer is no (including that student added new evidence but did not add more details to evidence already provided, or student added empty explanation, or student added explanations that did not connect to the argument or that contradict the argument, or student added personal commentary or non-text-based evidence), please output score 1. Step 3: If yes, did the student's revision show substantive improvement in the use of evidence, ? If answer is no improvement or slight improvement (including student added at least two small details, or student added at least two brief explanations of existing evidence, or student paraphrased existing evidence), please output score 2. If yes (substantive improvement is that student added relevant and solid details of evidence or explanations to at least two existing evidence), please output score 3.

Feedback Level 3. Step 1: Please compare the first draft and second draft, did the student show an attempt at content-level revision, considering the length of the first draft? If answer is no attempt or minimal attempt (including no edits at all, or revision focused solely on writing mechanics, or only a short line or two changed without significant content added), please output score 0. Step 2: If yes, did the student's revision align with the feed-

back? If the answer is no (including that student added evidence or added more details to evidence without any explanation, or student added empty explanation, or student added personal commentary, not explanation of evidence), please output score 1. Step 3: If yes, did the student's revision show substantive improvement in the use of evidence? If answer is no improvement or slight improvement (including that student recycled same explanation for each piece of evidence, or student paraphrased more than 1 existing evidence, or student only added one strong explanation for one piece of evidence, or student added at least two brief explanations of existing evidence, or student added a decent explanation only at the end of the essay, not after each piece of evidence), please output score 2. If yes (substantive improvement is that student strengthened explanation for at least two pieces of existing evidence, or student provided at least two pieces of strong connection between evidence presented to the overall argument), please output score 3.

D Examples of score 3 with the human rater rationale at each feedback level

Feedback Level 1:

- First draft: I am convinced that space exploration is desirable because space exploration helps us remain a creative society.It makes us strive for better technologies and scientific knowledge. This shows that people need more on earth than space. Another example is that space exploration will lead to long term benefits to society that more than justify the immediate costs. This shows that space exploration is desirable .This is why I am convinced that space exploration is desirable when so much needs to be done on space and earth.

- Second draft: I am convinced that space exploration is desirable because space exploration helps us remain a creative society.It makes us strive for better technologies and scientific knowledge. This shows that people need more on earth than space. Another example is that space exploration will lead to long term benefits to society that more than justify the immediate costs. This shows that space exploration is desirable .This is why I am convinced that space exploration is desirable when so much needs to be done. Another reason why space exploration is desirable is how scientist use monitors to check astronauts health before they go on an mission. This is another reason why space exploration is desirable. My next reason is, in addition ,the race led to significant investment and progress in american education ,especially in math and science. this shows that by looking outward into space ,we also improved life here on earth. This is why I am convinced that space exploration is desirable.

- Human rationale for scoring: This is Level 1 feedback, requiring "Using more evidence from the article" or "Providing more details for each piece of evidence used." The student attempted a content-level revision. The student added "Another reasons . . . we also improve life here on earth.", which seems to be an effort to add three text-based evidence to support their argument. Thus, the revision aligns with the feedback message and also results in a substantive improvement of the essay's evidence use. Therefore, the revision score is 3.

Feedback Level 2:

- First draft: Space exploration is desirable when there is so much that needs to be done on the earth. The space exploration can help solve some of the worlds problems. serious problem accrue on earth but the space exploration can fix some of them. Hunger problems, soil,crops,rainfall,droughts etc, can be solved by space exploration like the satellites that are around earth that monitor lots of land for the way food is produced and distributed. The text states "people on Earth can use this information to improve the way we produce and distribute food." This shows that the production of food and the way its distributed is going to be better if the scientist do the space explo- ration. The text also states "In rising to meet the challenges of space exploration, NASA sci- entist have developed other innovations that have improved our lives." Space exploration is desirable when there is so much that needs to be done on the earth. Earth has problems on it but scientist can solve them with space exploration. So space exploration is desirable to solve the needs of earth.

- Second draft: Space exploration is desirable when there is so much that needs to be done on the earth. The space exploration can help solve some of the worlds problems. serious problem accrue on earth but the space exploration can fix some of them. Hunger problems, soil,crops,rainfall,droughts etc, can be solved by space exploration like the satellites that are around earth that monitor lots of land for the way food is produced and distributed. The text states "people on Earth can use this information to improve the way we produce and distribute food." This shows that the production of food and the way its distributed is going to be better if the scientist do the space exploration.The way we distribute our food is important we have to make sure we have the right amount for everyone.The text also states "In rising to meet the challenges of space exploration, NASA scientist have developed other innovations that have improved our lives." This piece of evidence explains the way we face challenges on Earth,but that we can improve our lives a little better with the space exploration. Space exploration is desirable when there is so much that needs to be done on the earth. Earth has problems on it but scientist can solve them with space exploration. So space exploration is desirable to solve the needs of earth.The text states"Beyond providing us with inventions, space exploration is important for challenges it provides and the motivation to bring out the best in ourselves. Space exploration helps us remain a creative society." This shows that the space exploration helps in more ways than we thought, like we stay creative and it brings out our best side. Space exploration is desirable when there is so much that needs to be done on the earth. This shows how much we need space exploration.

- Human rationale for scoring: This is Level 2 feedback, requiring “Providing more details for each piece of evidence you use” or “Explain the evidence”. The student attempted a content-level revision. The student first added, "This piece of evidence explains the ...," which appears to be an attempt to provide an explanation for existing evidence. Additionally, the student added "the text states ’Beyond providing us ...’," which seems to be an effort to introduce detailed evidence along with an explanation for the argument. Thus, the revision aligns with the feedback message and also results in a substantive improvement of the essay’s evidence use. Therefore, the revision score is 3.

Feedback Level 3:

- First draft: They should get paid because 19 billion dollars a year for exploration. Most people think that this money should be spent instead of heal the people and the earth. Then 70 billion dollars spent on education (4.8% of the budget), or the 6.3 billion dollars spent on renewable (clean) energy. Before NASA allowed astronauts to go on the missions, scientists had to figure out how to monitor there health under any stress- ful conditions. They did this for the safety of the astronauts. NASA scientists have de- veloped other innovations that have improved our lives. NASA engineers developed to make space travel so they can do there mission. it is not just the people that need help. The Earth is suffering also. Many scientists believe that pollution from burning fossil fuels (Gasoline and oil) is harming our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories.

- Second draft: They should get paid because 19 billion dollars a year for exploration. Most people think that this money should be spent instead of to heal the people and the earth. Then 70 billion dollars spent on education (4.8% of the budget, or the 6.3 billion dollars spent on renewable (clean) energy. Before NASA allowed astronauts to go on the missions, scientists had to figure out how to monitor there health under any stressful conditions. They did this for the safety of the astronauts. NASA scientists have developed other innovations that have improved our lives. NASA engineers developed to make space travel so they can do there mission. it is not just the people that need help. the Earth is suffering also so that means that they need money to have the stuff to look and see what is going to happen in the future and there is a machine in space to see what the weather is going to be so they need money for that. It is important because like what is there is a tornado unexpected so they will not know how cold or what is going to happen there might be snow

coming and we do not know. Many scientists believe that pollution from burning fossil fuels (Gasoline and oil) is haring our air and oceans. We need new, cleaner forms of energy to power cars, homes, and factories. They also need money to have satellite see if we did not have a satellite we would not know when a tornado would come so that means we would not be prepared for a tornado we would not be able to evacuate or not get water food for a flood we would know have anything if we were not prepared it would come unexpected that is why they need money for all the things like satellite so we can be prepared for any storm. I think we should keep giving them money because they are keeping us safe by making a satellite and telling us on the news so we can get the info so we should keep giving they money so we can stay safe the money is a reward for keeping us safe so they should get money.

- Human rationale for scoring: This is level 3 feedback, requiring “Explain the evidence” or “Explain how the evidence connects to the main idea and elaborate”. The student attempted a content-level revision. The student first added, "so that means that they need ..." which appears to be an attempt to provide an explanation for why innovation can improve life on the the earth, such as weather. Additionally, the student added "they also need moeny to have satellite..." which seems to be an effort to introduce detailed evidence of satellite along with an explanation for how satellite can prepare for storm. Thus, the revision aligns with the feedback message and also results in a substantive improvement of the essay’s evidence use. Therefore, the revision score is 3.