Chain-of-Thought Prompting for Automated Evaluation of Revision Patterns in Young Student Writing

Tianwen Li[1], **Michelle Hong**[1],**Lindsay Clare Matsumura**[1], **Elaine Wang**[3]
Diane Litman[1,2], **Zhexiong Liu**[2],**Richard Correnti**[1]

[1]Learning Research and Development Center, University of Pittsburgh
[2]Department of Computer Science, University of Pittsburgh
[3]RAND Corporation
{tianwen.li,mih196,lclare,dlitman,rcorrent}@pitt.edu
zhexiong@cs.pitt.edu, ewang@rand.org

Abstract

This study explores the use of Chain-of-Thought prompting with ChatGPT-4.1 as an approach for identifying revision patterns in young students' argument writing. ChatGPT-4.1 shows substantial agreement with human coders on evidence-related revision patterns and moderate agreement on explanation-related ones. Implications for CoT prompting for writing evaluation are discussed.

1 Introduction

Revision is a crucial component of the writing process (Hayes, 1996; Fitzgerald, 1987). However, young students struggle with effective revision (Graham et al., 1995; Wang et al., 2020). A well-established approach to improving writing skills is to provide formative feedback targeting various aspects of the writing process, as it builds knowledge of the criteria for successful writing (Stein and Matsumura, 2008; Matsumura et al., 2023). Few assessments directly measure revision quality in terms of how successfully students implement feedback to meet particular writing goals. Instead, it is typically inferred from the overall quality of the revised text using holistic or trait-based scoring. However, such measures do not provide direct insight into the specific revision actions taken or into their effectiveness in meeting writing goals. Therefore, there is a need for assessments that directly capture holistic revision patterns (e.g., adding non-text-based evidence) to reveal how effectively students revise between drafts in response to specific writing goals (Correnti et al., 2024). Such assessments can serve two purposes: providing formative information to support the development of students' writing skills and providing data for research. These purposes require different validity investigations, but both involve reliability (Correnti et al., 2022).

One of the major barriers to developing and implementing direct measures of revision quality is the time-consuming nature of the evaluation process, which has been well documented in educational research. In studies that examine students' revision, researchers have relied on manual human coding to track and evaluate changes between drafts (e.g., Cho and MacArthur, 2010; MacArthur and Graham, 1987; Wang et al., 2020). The quality of revisions is often examined across four aspects: the type of revision (surface- or content-level), the type of operation (e.g., addition, deletion), the impact on meaning (meaning-changing or meaning-preserving), and the impact on text quality (increase or decrease). This is a multi-step process that is too labor-intensive for busy teachers, who would also need specialized training. Moreover, the cost of training and hiring raters to assess revision patterns in essays is prohibitively expensive for writing researchers. Therefore, efficient approaches are needed to assess revision patterns.

Recent advancements in large language models (LLMs) have shown significant promise in evaluating student writing quality (Li et al., 2024; Pack et al., 2024; Seßler et al., 2025; Tang et al., 2024; Tate et al., 2024) and can be an efficient approach to assessing revision patterns. The present study explores the potential of one of the most popular LLMs, ChatGPT 4.1, for identifying revision patterns in students' text-based argument writing. We are interested in exploring the integration of the chain-of-thought (CoT) prompting strategy to improve the performance of automatic evaluation of revision patterns. The CoT prompting is frequently used to evaluate student responses to mathematics and science tasks, as specifying the intermediate reasoning steps leading to the final answer can improve LLM performance in automatic scoring within those fields (e.g., Cohn et al., 2024; Lee

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 49–65

et al., 2024). In contrast, prompting strategies for writing evaluation often rely on zero-shot to few-shot prompting by including scoring rubrics and sometimes related essay examples (e.g., Pack et al., 2024; Tate et al., 2024), but they rarely specify intermediate evaluation steps. Akin to prompting strategies employed in math and science evaluation, we aim to extend current research by exploring whether CoT can improve ChatGPT-4.1's performance in assessing revision quality.

To address this goal, this study applies two different prompting strategies: the baseline strategy of few-shot prompting and few-shot CoT prompting. We then examine the reliability of ChatGPT-4.1's predictions from two perspectives: the internal consistency in generating the same output across multiple runs and the accuracy in predicting revision patterns that align with human coding. The reliability of automated scores has most often been evaluated by comparing machine-generated ratings with human ratings, which are often considered the gold standard. We extend the investigation of reliability by examining the consistency of ChatGPT-4.1's ratings. Internal consistency has received less attention in prior research, yet it is important given evidence that LLMs can produce inconsistent ratings (Tang et al., 2024; Tran et al., 2024). Therefore, this paper addresses the following questions:

1. How internally consistent is ChatGPT-4.1 in assessing revision patterns in students' text-based argument writing across different prompting strategies?
2. How accurate is ChatGPT-4.1 in assessing revision patterns in students' text-based argument writing across different prompting strategies?

2 Data

In this section, we describe the dataset of students' essays, outline the taxonomy of revision patterns used to code revision quality, and explain the human coding process.

2.1 Dataset of student essays

The corpus for this study is drawn from a larger project of eRevise+RF, an automated writing evaluation system designed to support young students' argument writing and revision (Correnti et al., 2024; Liu et al., 2023; Liu et al., 2025). eRevise+RF is developed 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). To administer the system, the teacher reads the text aloud to students, poses planned questions, and defines selected vocabulary at specific points in the article to ensure that all students comprehend the material before writing. In this study, each student completed one of two RTAs: one task was based on an article about the United Nations Millennium Villages Project (MVP) to fight poverty in Kenya, and another one was based on an article about the benefits and costs of space exploration (SPACE) (Appendix A).

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, the specificity of evidence, the concentration of evidence, and word count) to select appropriate revision goals and related feedback based on the quality of evidence in each draft. Revision goal 1 emphasizes adding additional evidence from the text. Revision goal 2 instructs students to add details to existing evidence to increase specificity. Revision goal 3 guides students to explain their evidence and connect it to the claims (Correnti et al., 2020; Wang et al., 2020). After receiving a tailored revision goal, students revise their essays accordingly.

The dataset was collected from 330 students in grades 4th through 8th in Louisiana and Pennsylvania. It contains a total of 330 essay pairs, including both initial and revised drafts. Among these essay pairs, 172 were written in response to the MVP article, while 158 were written in response to the SPACE article.

2.2 Taxonomy of revision patterns for argument writing

The taxonomy of revision patterns for argument writing is adapted from Wang et al.'s (2020) qualitative study, which examined how students revised their writing in response to the aforementioned revision goals and the feedback generated by eRevise, the earlier version of eRevise+RF. The revision patterns identified by Wang et al. (2020) were reorganized and consolidated around three guiding questions: 1) Do the revisions focus on content? 2) Do the revisions effectively address the targeted goal? and 3) To what extent do the revisions sub-

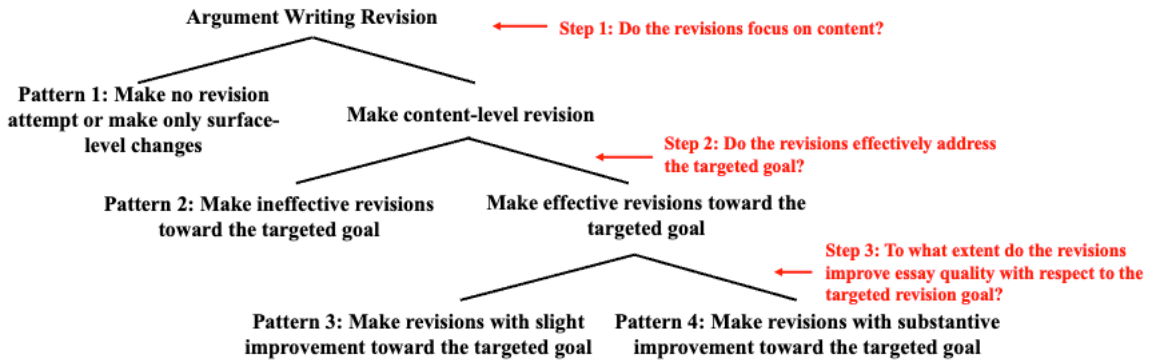


Figure 1: Taxonomy of revision patterns. This taxonomy reflects the general framework for categorizing revision patterns. The manifestations of revision patterns for each goal are presented in Appendix B.

stantially improve essay quality with respect to the targeted revision goal? Based on these questions, four main revision patterns were generated and organized into a taxonomy, as shown in Figure 1. The four revision pattern codes reflect a progression from the least to the most effective type of revision patterns, which represents an ordinal scale. In addition, these revision pattern codes are mutually exclusive, which means coders should assign only one code to each essay pair. As students were assigned different revision goals, the description of each revision pattern for each goal and related examples are presented in detail in Appendix B.

2.3 Human coding of revision patterns

Assessing revision quality is a complex process. To increase interrater reliability between two human coders, we standardized the coding procedure by breaking down the task of assigning revision patterns into a series of manageable steps developed from the three guiding questions (Figure 1). All changes between the first and second drafts were counted as revisions. The procedure was structured as a three-step sequential filtering approach, with each step building on the previous one to progressively focus the analysis on a smaller, more meaningful set of revisions:

Step 1. Examine all revisions to determine whether any content-level changes are presented. If all changes are surface-level (e.g., mechanical issues of writing), the revision pattern is Pattern 1: Make no revision attempt or make only surface-level changes.

Step 2. Further examine the revisions identified as content-level to determine whether those revisions effectively address the targeted goal. If none of the revisions are effective, the revision pattern is Pattern 2: Make ineffective revisions toward the targeted goal.

Step 3. Focus on the revisions that effectively address the revision goal, and consider both the quantity of these revisions and the overall quality of the first draft to determine whether there is a substantive improvement in overall essay quality. This step aims to select whether the revision pattern should be Pattern 3 of making revisions with slight improvement or Pattern 4 of making revisions with substantive improvement. This three-step coding process was developed into a decision-making flowchart presented in Appendix C.

All the essay pairs were double-coded by two human coders. Discrepancies were discussed and resolved between coders to establish the benchmark for the comparison with coding by ChatGPT-4.1. We assessed interrater reliability with two metrics: exact agreement, calculated using confusion matrices, and quadratic weighted kappa (QWK). The interpretation of Kappa follows the guideline proposed by Landis & Koch (1977): values below 0 indicate poor agreement; 0.01–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1, almost perfect agreement. Human coders engaged in identifying revision patterns for each of the three goals (i.e., evidence, details, and explanations). For the revision goal of adding evidence, the exact agreement between the two coders was 87.5%, with a QWK of 0.87, indicating almost perfect agreement. For the revision goal of adding details, the exact agreement was 84.91%, and the QWK was 0.91, indicating almost perfect agreement. For the final revision goal of adding explanations, the exact agreement was 79.67%, and the QWK was 0.77, indicating

	Goal 1: Evidence	Goal 2: Details	Goal 3: Explanations
Pattern 1: Make no revision attempt or make only surface-level changes	2	3	19
Pattern 2: Make ineffective revisions toward the targeted goal	14	22	60
Pattern 3: Make revisions with slight improvement toward the targeted goal	44	17	32
Pattern 4: Make revisions with substantive improvement toward the targeted goal	74	17	26
Total	134	59	137

Table 1: Distribution of revision patterns for each revision goal

substantial agreement. The distribution of human-coded revision patterns for each revision goal is shown in Table 1.

3 Methods

We used ChatGPT-4.1 to assign revision pattern codes to each essay pair (i.e., first and second draft). To evaluate the model's internal consistency in assessing revision quality, each essay pair was processed three times using the same prompt. To minimize randomness in the output, the temperature was set to 0. Two types of prompts were tested in this study: few-shot prompting and few-shot CoT prompting. In the following section, we provide an overview of these prompting methods (see Appendix D for details).

3.1 Few-shot prompting

Few-shot prompting served as the baseline, in which ChatGPT-4.1 was directly asked to assign one of the revision pattern codes to each essay pair. The prompt consisted of three parts. First, it introduced the RTA by providing the source text and the related writing prompt. Second, it specified the student's revision goal and presented the list of revision pattern codes associated with that goal. We also include the descriptions of each code and an essay pair to exemplify the pattern. Finally, the student's first and second drafts were provided, and the model was prompted to output one revision pattern code.

3.2 Few-shot chain-of-thought prompting

The CoT prompting approach was developed based on the human coding process described in the previous section. While most of the information included in the baseline prompt remained the same, the only change was substituting the list of revision patterns with stepwise guidance for selecting a pattern. Given the three revision goals, more detailed instructions tailored to each goal were developed.

3.3 Evaluation metrics

To address RQ1, we ran each type of prompt three times to evaluate ChatGPT-4.1's consistency in generating the same code. We then calculated the QWK for each pairwise comparison and averaged the results to determine internal consistency. To address RQ2, we used a majority vote approach to determine the final code assigned by ChatGPT-4.1 across the three runs and computed exact agreement and QWK to evaluate the agreement between ChatGPT-4.1 coding and human coding. We further computed per-class F1 scores for each revision goal, which enabled a direct comparison between baseline prompting and CoT prompting in their ability to identify different revision patterns.

4 Results & Discussion

4.1 RQ1: Consistency of revision pattern predictions

Both the baseline prompting strategy and the CoT prompting strategy exhibited almost perfect consistency across the three runs, with QWK values exceeding 0.90 for each revision goal (Table 2). This finding demonstrates that ChatGPT-4.1 is highly reliable in assigning the same revision patterns to essays when prompted multiple times. Moreover, consistency remained stable across prompting strategies, suggesting that the internal consistency of ChatGPT-4.1 in assessing revision quality is not affected by different prompting strategies.

	Baseline		CoT	
	Exact Agreement	QWK	Exact Agreement	QWK
Goal 1: Evidence	90.30%	0.94	90.30%	0.94
Goal 2: Details	89.83%	0.96	88.14%	0.94
Goal 3: Explanations	89.78%	0.94	92.70%	0.96

Table 2: Internal consistency across two prompting strategies

	Baseline		CoT	
	Exact Agreement	QWK	Exact Agreement	QWK
Goal 1: Evidence	64.18%	0.56	64.93%	0.63
Goal 2: Details	67.80%	0.76	61.02%	0.80
Goal 3: Explanations	40.88%	0.39	52.55%	0.54

Table 3: Prediction accuracy across two prompting strategies

4.2 RQ2: Accuracy of revision patterns predictions

As shown in Table 3, with the baseline prompting strategy, we observed an exact agreement of 64.18% and a QWK of 0.56 for revision goal 1, indicating moderate agreement with human coding. For revision goal 2, the exact agreement was 67.80% with a QWK of 0.76, indicating substantial agreement with human coding. For revision goal 3, the exact agreement was 40.88% with a QWK of 0.39, indicating only fair agreement with human coding.

We further tested the CoT prompt, and the results showed that it improved accuracy in predicting revision patterns. For goal 3, the exact agreement rose from 40.88% to 52.55%, and the QWK increased from 0.39 to 0.54, representing a moderate level of agreement. For goal 1, the QWK slightly increased from 0.56 to 0.63, indicating substantial agreement with human coding, while the agreement remains substantial for revision goal 2.

Across both prompting strategies, revision goal 3 (adding explanations) consistently showed the lowest accuracy in predicting revision patterns. Similarly, in human coding, interrater reliability was lowest for revision goal 3 (QWK=0.77). Assessing the quality of newly added explanations (goal 3) is more subjective than assessing the quality of added evidence (goal 1) or related details (goal 2). The quality of evidence can be directly checked against the source text. By contrast, explanations of how evidence supports claims vary widely in length, clarity, logic, and persuasiveness, which requires more nuanced judgment. Simply including the definition of revision patterns in the prompt does not capture the judgments made by humans during coding the quality of explanations. As a result, ChatGPT-4.1 struggled to assign revision patterns that aligned with human coding when working on goal 3. This finding suggests that the performance of LLMs co-varies with the level of human agreement in coding educational materials (Cohn et al., 2024; Wang et al., 2023). In other words, when human coders demonstrated higher interrater reliability, ChatGPT-4.1 also achieved higher accuracy in predicting revision patterns.

We further computed F1 scores to gain insight into how baseline prompting and CoT prompting performed differently on identifying revision patterns, with particular attention to revision goal 3. As shown in Table 4, the CoT prompt increased the weighted-average F1 score from 0.40 to 0.51. When examining the per-class F1 scores for each revision pattern, the CoT improved performance in predicting ineffective revisions of explanation (pattern 2), achieving an F1 score of 0.63, nearly double that of the baseline model (F1 = 0.34). This higher score reflects CoT's ability to capture more true instances of ineffective revisions while reducing misclassifications of other revision types as ineffective. This is an important improvement in prediction accuracy, as the evaluation of ineffective

	Goal 1: Evidence		Goal 2: Details		Goal 3: Explanations	
	Baseline	CoT	Baseline	CoT	Baseline	CoT
Pattern 1	0.40	0.29	0.67	0.75	0.52	0.62
Pattern 2	0.46	0.46	0.63	0.73	0.34	0.63
Pattern 3	0.52	0.58	0.63	0.57	0.42	0.24
Pattern 4	0.74	0.74	0.78	0.77	0.44	0.48
Weighted average F1 score	0.63	0.65	0.68	0.70	0.40	0.51

Table 4: F1 scores across two prompting strategies

explanation revision is the most complex in the human coding process. This complexity arises primarily from the pedagogical knowledge required to recognize the diverse forms of ineffective explanation attempts. Instead of adding explanations that clearly connect evidence to claims, young students often insert personal comments, empty explanations, summaries of the evidence, or elaborations that do not strengthen the argument. Moreover, students' ineffective explanations are not always presented in a single pattern; rather, they frequently appear as a mix of multiple inadequate attempts in their revision, sometimes even accompanied by partial but effective explanations. With the baseline prompt, when an essay contained both effective and ineffective revisions of exlanations, the selection of a revision pattern often appeared arbitrary, as no clear major pattern emerged. By contrast, with the CoT prompt, ChatGPT-4.1 was instructed to evaluate the quality of explanation revisions first at the sentence level and then transition to the essay level by considering the quantity of effective explanations shown within the revision. The inclusion of standardized evaluation steps in the CoT prompting, which makes explicit the considerations human coders apply during coding, likely contributed to the accuracy of identifying the revision pattern that applied ineffective explanations.

5 Conclusions

Revision is a crucial component in writing development, yet many young students struggle to revise effectively (Wang et al., 2020). Accurately evaluating the revision quality (e.g., identifying revision patterns) is a key step in providing targeted feedback that supports the growth of their revision skills. However, such evaluation is time-consuming for human coders. Therefore, this study investigates the potential of ChatGPT-4.1 as an alternative tool for identifying revision patterns across various writing goals. Our findings demonstrate that ChatGPT-4.1 is highly consistent in predicting the same revision patterns across multiple runs and shows strong potential for effectively identifying patterns that align with human coders. Similar to studies that explore CoT prompting in the automated scoring of math and science tasks (e.g., Lee et al., 2024), we also found that including intermediate evaluation steps improves the accuracy of predicting revision patterns, particularly those under the goal of adding explanations. Specifying evaluation steps makes the nuanced judgments of human coders more explicit, which likely contributed to this improvement. Moving forward, we suggest that researchers and teachers carefully reflect on and document their writing evaluation processes, standardize these steps, and transform them into a sequence of manageable subtasks or decision points. Such practices may better support collaboration with LLMs in scoring tasks more broadly.

6 Limitations

First, our study focused on a specific writing evaluation task of assessing revision quality among young students. Future research should apply CoT prompting strategies across diverse writing evaluation tasks, such as holistic scoring or trait-based scoring in different writing genres to examine whether CoT can outperform baseline models. Second, the assessment of revision introduced in this study is designed primarily for the purpose of providing feedback by teachers in the classroom, and we only focused on testing the reliability of ChatGPT-4.1 scoring. Although we demonstrated its potential in identifying revision patterns, future research should investigate the validity of the assessment to ensure that it captures the meaningful dimensions of student revision or develop a more comprehensive format based on it.

Acknowledgments

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

Kwangsu Cho and Charles MacArthur. 2010. Student revision with peer and expert reviewing. *Learning and instruction*, 20(4):328–338.

Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. 2024. A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 23182–23190.

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. *Computers and Education Open*, 3:100084.

Rip Correnti, Elaine Lin Wang, Lindsay Clare Matsumura, Diane Litman, Zhexiong Liu, and Tianwen Li. 2024. Supporting students' text-based evidence use via formative automated writing and revision assessment. In *The Routledge international handbook of automated essay evaluation*, pages 221–243. Routledge.

Jill Fitzgerald. 1987. Research on revision in writing. *Review of educational research*, 57(4):481–506.

Steve Graham, Charles MacArthur, and Shirley Schwartz. 1995. Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology*, 87(2):230.

Andrew F Hayes. 1996. Permutation test is not distribution-free: Testing h: ρ= 0. *Psychological Methods*, 1(2):184.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Gyeong-Geon Lee, Ehsan Latif, Xuansheng Wu, Ninghao Liu, and Xiaoming Zhai. 2024. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.

Tianwen Li, Zhexiong Liu, Lindsay Matsumura, Elaine Wang, Diane Litman, and Richard Correnti. 2024. Using large language models to assess young students' writing revisions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 365–380, Mexico City, Mexico. Association for Computational Linguistics.

Zhexiong Liu, Diane Litman, Elaine Wang, Tianwen Li, Mason Gobat, Lindsay Clare Matsumura, and Richard Correnti. 2025. erevise+ rf: A writing evaluation system for assessing student essay revisions and providing formative feedback. *arXiv preprint arXiv:2501.00715*.

Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura, and Richard Correnti. 2023. Predicting the quality of revisions in argumentative writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 275–287, Toronto, Canada. Association for Computational Linguistics.

Charles A MacArthur and Steve Graham. 1987. Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *The Journal of special education*, 21(3):22–42.

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.

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 462–472.

Mary Kay Stein and Lindsay Clare Matsumura. 2008. Measuring learning for teacher instruction. *Measurement Issues and Assessment for Teaching Quality*, page 179.

Xiaoyi Tang, Hongwei Chen, Daoyu Lin, and Kexin Li. 2024. Harnessing llms for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14).

Tamara P Tate, Jacob Steiss, Drew Bailey, Steve Graham, Youngsun Moon, Daniel Ritchie, Waverly Tseng, and Mark Warschauer. 2024. Can ai provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255.

Nhat Tran, Benjamin Pierce, Diane Litman, Richard Correnti, Lindsay Clare Matsumura, et al. 2024. Multi-dimensional performance analysis of large language models for classroom discussion assessment. *Journal of Educational Data Mining*, 16(2):304–335.

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.

Rose E Wang, Pawan Wirawarn, Noah Goodman, and Dorottya Demszky. 2023. Sight: A large annotated dataset on student insights gathered from higher education transcripts. *arXiv preprint arXiv:2306.09343*.

A RTA writing task

MVP writing task: The author described how the quality of life can be improved by the Millennium Villages project in Sauri, Kenya. Based on the article, did the author convince you that "winning the fight against poverty is achievable in our lifetime"? Explain why or why not with 3 to 4 examples from the text to support your answer.

SPACE writing task: 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 leads to long-term benefits" that justify the cost? Give reasons for your answer. Support your reasons with 3 to 4 pieces of evidence from the text.

B Revision Pattern Codes

Revision Goal 1: Adding more text-based evidence

1. Make no revision attempt or make only surface-level changes: The revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions.

Example: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

2. Make ineffective revisions toward the targeted goal: The revision is at content-level but uses ineffective revision strategies that fail to improve the quality of evidence used in the essay. These strategies include adding explanations instead of adding new evidence and adding new evidence that is not based on the text.

Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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...". Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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...". Therefore, that new technique can be used to more people and help them to be healthy.

3. Make revisions with slight improvement toward the targeted goal: The revision involves adding only one piece of evidence or adds a list of keywords of various evidence in the second draft; however, the draft as a whole still contains only two or fewer pieces of evidence.

Example: Draft 1: 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 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. Draft 2: 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 space exploration is desirable. Space exploration can even solve problems. It can also monitor land, track corps, stop wars, provide motivations to nations.

4. Make revisions with substantive improvement toward the targeted goal: The revision involves adding more than one piece of evidence from different parts of the text to support the main claim.

Example: Draft 1: 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 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. Draft 2: 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 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. Another reason why space exploration is desirable is how scientist use monitors to check astronauts health. 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. Finally, Over 46.2 million Americans (15%) live in poverty. The investment in space exploration argue

that 19 billion dollars is not to much.

Revision Goal 2: Adding details to the pieces of evidence used in the essay

1. Make no revision attempt or make only surface-level changes: The revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions.

Example: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

2. Make ineffective revisions toward the targeted goal: The revision is at content-level but uses ineffective revision strategies that fail to improve the specificity of evidence used in the essay. These strategies include adding explanations instead of adding new evidence, and adding new evidence and details that are not based on the text.

Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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...". Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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...". Therefore, that new technique can be used to more people and help them to be healthy.

3. Make revisions with slight improvement toward the targeted goal: The revision involves adding details to only one piece of evidence in the second draft or introduces a new piece of evidence with limited detail. However, the rest of the evidence in the second draft continues to lack specificity and details.

Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have solve hunger and poverty. These include better exercise machines, better airplanes, and better weather forecasting." That is why I believe that we should find space exploration more desirable. Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". Before NASA allowed astronauts to go on missions, scientists had to find ways to monitor their health..." and, "... innovations that have solve hunger and poverty." For example, the text states that Satellites that circle Earth can monitor land and the atmosphere. They can track and measure the conditions of crops, soil, and rainfall. We can use this information to improve the way we produce and distribute food. That is why I believe that we should find space exploration more desirable.

4. Make revisions with substantive improvement toward the targeted goal: The revision involves adding details to more than one piece of evidence in the second draft, so the evidence becomes more specific.

Example: Draft 1: 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 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. Draft 2: 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 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. Another reason why space exploration is desirable is how scientist use monitors to check astronauts health. 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. Finally, Over 46.2 million Americans (15

Revision Goal 3: Explain the evidence and connect to the claims 1. Make no revision attempt or make only surface-level changes: The revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions.

Example: Draft 1: There is alot of people that are nerds and they wont all the money to go to

space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

2. Make ineffective revisions toward the targeted goal: The revision is at content-level; however, the changes do not improve the quality of the explanation of how the evidence supports the claim. Ineffective revision strategies include adding new evidence but not addressing the revision goal of adding the explanation, adding personal comments instead of explaining, providing empty explanations, paraphrasing existing evidence without explaining how the evidence supports the claim, or elaborating on the evidence without explaining how the evidence supports the claim.

Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting. Malaria is common in Africa. Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting." Malaria is common in Africa." this is a preventable illness, just need people to donate some money, and children can live.

3. Make revisions with slight improvement toward the targeted goal: The revision involves adding a brief explanation to one piece of evidence to show how it supports the main claim or reuses the same explanation for multiple pieces of evidence. However, how each distinct piece of evidence supports the claim remains unclear, and some evidence may be left unaddressed.

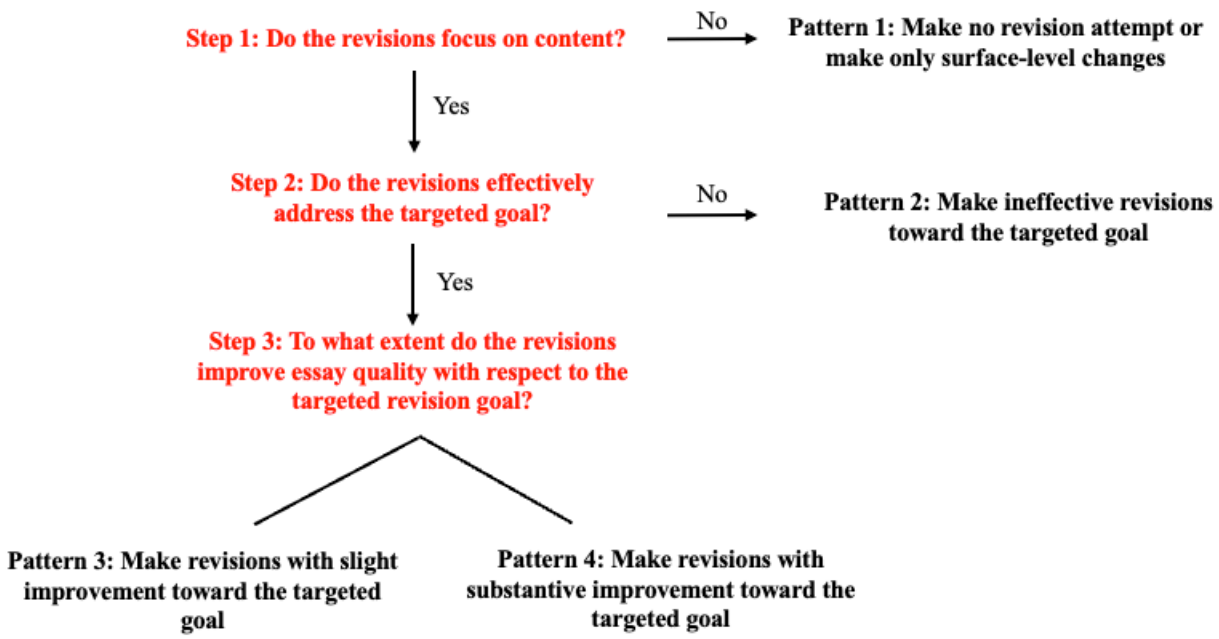
Example: Draft 1: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. Malaria kills over 3,000 African children every day. On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy. People need money. Draft 2: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. That is showing how if we did have to pay the fund than some people would not be able too pay it. If you do so much on earth to make it better it will be better to live in (an explanation to one piece of evidence). Malaria kills over 3,000 African children every day On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy.

4. Make revisions with substantive improvement toward the targeted goal: The revision involves adding multiple explanations to existing evidence, clearly showing how the various pieces of evidence support the claim, thereby making the argument more persuasive.

Example: Draft 1: They should get paid because 19 billion dollars a year for exploration. 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. 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. Draft 2: They should get paid because 19 billion dollars a year for exploration. 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. 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. (a strong explanation) 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. 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.

C Human Coding Flowchart



D Prompt Designs for ChatGPT-4.1 (Using Goal 3 as an Example)

Baseline prompt Students from 4th grade to 8th grade are currently working on text-based argument essays. After submitting their first draft, students received feedback and revised their second draft. Here is the text and writing task:

<Text + Writing Prompt>

You are a writing teacher who works with students from 4th grade to 8th grade. Students are working on the revision goal of explaining how the evidence supports the main claim. Your task is to compare the first and second drafts to identify the major revision patterns in the students' work in response to the revision goal. Below are four revision patterns; select the one that best represents the student's revision.

1. Make no revision attempt or make surface-level revision: The revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions. If changes in the second draft exceed 35 words, it is not considered a surface-level revision. Example: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

2. Make revisions with ineffective strategies toward the targeted revision goal: The student attempts a content-level revision; however, the changes do not improve the quality of the explanation of how the evidence supports the claim. Ineffective revision strategies include adding new evidence but not addressing the revision goal of adding the explanation, adding personal comments instead of explaining, providing empty explanations, paraphrasing existing evidence without explaining how the evidence supports the claim, or elaborating on the evidence without explaining how the evidence supports the claim. Example: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting. Malaria is common in Africa. Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting." Malaria is common in Africa." this is a preventable illness, just need people to donate some money, and children can live.(personal comments instead of explaining how existing evidence supports the claims).

3. Make slight improvement toward the targeted revision goal: The student adds a brief explanation to one piece of evidence to show how it supports the main claim or reuses the same explanation for multiple pieces of evidence. However, how each distinct piece of evidence supports the claim remains unclear, and some evidence may be left unaddressed. Example: Draft 1: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. Malaria kills over 3,000 African children every day. On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy. People need money. Draft 2: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. That is showing how if we did have to pay the fund than some people would not be able too pay it. If you do so much on earth to make it better it will be better to live in (an explanation to one piece of evidence). Malaria kills over 3,000 African children every day On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy.

4. Make substantive improvement toward the targeted revision goal: The student adds multiple explanations to existing evidence, clearly showing how the various pieces of evidence support the claim, thereby making the argument more persuasive. Example: Draft 1: They should get paid because 19 billion dollars a year for exploration. 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. 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. Draft 2: They should get paid because 19 billion dollars a year for exploration. 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. 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. (a strong explanation) 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. 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 (a strong explanation).

Output one of the following category numbers: **1; 2; 3; 4**

CoT prompt Students from 4th grade to 8th grade are currently working on text-based argument essays. After submitting their first draft, students received feedback and revised their second draft. Here is the text and writing task:

<Text + Writing Prompt>

You are a writing teacher who works with students from 4th grade to 8th grade. Students are working on the revision goal of explaining how the evidence supports the main claim. Your task is to compare the first and second drafts to identify the major revision patterns in the students' work in response to the revision goal. To determine the appropriate revision pattern, follow these steps:

Step 1: Determine whether the revision is surface-level. Surface-level revision involves only mechanical changes, such as correcting spelling, grammar, or making minor word substitutions. If changes in the second draft exceed 35 words, it is not considered a surface-level revision. If the revision pattern is surface-level, output pattern number "1." If not, proceed to step 2. Example of Pattern 1: Draft 1: There is alot of people that are nerds and they wont all the money to go to space and not on earth. The arguments stem from a belif that money spent could be used differently- to improve people's lives. Draft 2: There are a lot of people that are nerds and they wont all the money to go to space and not on earth. The arguments against space explortion stem from a belief that money spent could be used differently- to improve people's lives.

Step 2: Assess the quality of the newly added content after each piece of evidence. Assign one of the following codes to indicate the effectiveness of each explanation attempt: 1. Personal Commentary: An elaboration that is about personal reactions, evaluations, or feelings. For example, after presenting evidence that Malaria is a disease common in Africa, the student added, "it is pretty cool I never knew that and I am afraid of getting it." 2. Empty explanation: This type of explanation attempt is overly broad, vague, and does not have content, though it may look like an explanation in form. For example, the student added, "Based on what I provided, this is why I agree." 3. Paraphrase: A revision attempt where the explanation merely rewords the evidence rather than explains it. For example, after presenting the evidence " Those in favor of space exploration argue that 19 billion dollars is not too much and satellites can solve the problem of poverty," students added, "we should fund space exploration because 19 billion dollars is not too much and provide more food for people in poverty." 4. Elaboration of Evidence but no

Connection to the claim: Explain the evidence by discussing the implications or providing more information, but it lacks reasoning of why the evidence supports the claim. For example, after presenting the evidence that " Malaria is common in Africa," the student added, “this is a preventable illness, just need people to donate some money, and children can live.” This is an elaboration of evidence, but no explanation of how this example supports the claim. 5. Linked Claim-Evidence: A revision that explains why the provided evidence supports the main claim. For example, to support the claim that we should support space exploration, the student added the explanation, “Malaria is common in Africa. And space exploration can develop new medicine and equipment. These are good evidence because space exploration can save people lives and prevent Malaria”

Step 3: Determine the quality and quantity of explanations added in the second draft If the second draft only contains explanations coded as “personal comments”, “empty explanation”, “paraphrase”, and “elaboration of the evidence”, output 2.

Example of Pattern 2: Draft 1: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting. Malaria is common in Africa. Draft 2: Yes the author did convince me that "space exploration is desirable when there is so much that needs to be done on earth". The text states, "... 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..." and, "... innovations that have improved our lives. These include better exercise machines, better airplanes, and better weather forecasting." Malaria is common in Africa." this is a preventable illness, just need people to donate some money, and children can live.(personal comments instead of explaining how existing evidence supports the claims).

If the second draft contains one piece of explanation category as Linked Claim-Evidence, or reuses the same Linked Claim-Evidence for multiple pieces of evidence, output 3.

Example of Pattern 3: Draft 1: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. Malaria kills over 3,000 African children every day. On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy. People need money. Draft 2: The space exploration does not help our city/town to be the best. In paragraph 3 it says over 46.2 million Americans live in poverty and nearly half of all Americans also have difficulty paying for housing,food,and medicine at some points of their lives. That is showing how if we did have to pay the fund than some people would not be able too pay it. If you do so much on earth to make it better it will be better to live in (an explanation to one piece of evidence). Malaria kills over 3,000 African children every day On paragraph 6 it is saying how much they are spending like 670 billion the US spends for national defense than they spend 70 billion dollars spent on education and like 6.3 billion dollars on renewable energy. If the second draft contains at least two Linked Claim-Evidence, output pattern number “4.” Example of Pattern 4: Draft 1: They should get paid because 19 billion dollars a year for exploration. 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. 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. Draft 2: They should get paid because 19 billion dollars a year for exploration. 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. 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. (a strong explanation) 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. 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 (a strong explanation).

Apply the aforementioned evaluation steps and reason step by step. Output one of the following category numbers: **1; 2; 3; 4**