Evaluating Written Corrective Feedback:

A Comparative Study of AI-Generated and Human-Written In-Line Essay Comments

April, 2024

Rachel Hantz, M.Sc., University of Washington, Machine Learning Scientist, Paper™
Charlotte Richardson, B.A., McGill University, Educational Developer, Paper
Ashkan Golzar, Ph.D., McGill University, Director of Data Insights, Paper

Introduction

Students who received written corrective feedback from Paper's Review Center demonstrated significantly higher end-of-year ELA outcomes than students who did not, controlling for baseline (fall) scores in i-Ready (Paper, 2023), pre-ACT (Paper, 2023), and ELA grades (Paper, 2023). Since written corrective feedback is impactful, monitoring its quality is valuable to stakeholders who implement tutoring services for essay-writing. This includes both educational professionals and technology companies. As writing tools driven by artificial intelligence (AI) have become increasingly popular in the past decade, the expectation for high quality writing reviews must now be applied to both human-written and AI-generated feedback.

Collecting a measure of quality on AI and human-written corrective feedback supports three specific efforts. First, it assesses quality from two different sources–AI and human–allowing for a comparison of feedback source strengths and weaknesses. It also informs on adjustments needed when prompting an LLM to generate feedback. Lastly, ground truth data collection is the first step to designing an AI-tool to automatically assess the quality of written corrective feedback. In the present report, we discuss the first two efforts, and prepare for the third. We define a novel rubric for assessing the efficacy of written corrective feedback at the comment level, describe a procedure for manually implementing this rubric, and present findings on the quality of human-written and AI-generated writing feedback.

When evaluated by human domain experts, the feedback on essays composed in English show differences in quality depending on if an essay's feedback is either solely AI-generated or solely human-written. We find that AI feedback generated by a certain prompting of an LLM tends to include 28.7% more inquiry-based comments, 18.5% less encouraging comments, and 14.3% more specific comments per essay (56.1%, 0.4%, and 65.9% respectively) than human-written feedback (27.4%, 18.9%, and 51.6% respectively). Overall, AI-generated and human-written feedback are similarly suitable to the student's level (80.1% and 74.6% suitable per essay, respectively), but for younger grades, AI-generated feedback struggles. The present AI-generated feedback never included standalone praise, while 8.9% of human-written feedback did. Also, less than 3% of AI-generated and human-written feedback comment on an issue more than twice per essay. Lastly, AI-generated comments are found to be 100% safe and 5.5% inaccurate. Less than 1% of human-written comments are

inaccurate and unsafe. Overall, AI-generated comments exceed human-written comment performance in structural dimensions of feedback, but lack suitability in tone and young learner context.

Rubric: In-Line Comments on Student Writing

In Paper's Writing Review Center, students can submit writing for review and receive feedback from a tutor. The feedback consists of in-line comments directly embedded within the student's writing submission, as well as a general comment addressing the overall composition. Review Center tutors support students asynchronously by offering constructive and positive writing guidance that adheres to the core values of the Paper Method: Positive, Adaptable, Patient, Engaging, and Respectful.

In partnership with teaching and learning specialists, we concretized this Paper Method into an eight-item rubric that assesses the quality of essay writing in-line feedback. In-line comments should be *inquiry-based* often, be *encouraging*, be *specific*, be *suitable for the student's level,* occasionally be *positive feedback only*, never *state the same issue for the third (or more) time*, be *safe*, and be *accurate.* The first six items of the rubric align with the Paper Method's core values (Appendix A), while the latter two items are paramount for evaluating AI.

We also designed this rubric to be suitable for machine learning tasks. One way for a task to be suitable for machine learning is when a machine learning model can accurately predict an output, such as a classification, based on a given input. Each rubric item thus contains a key question where the output is "yes" or "no" alongside an extended description. Each rubric item is applied to individual comments.

The rubric item definitions that human domain experts used to evaluate comments are as follows. (We include examples of each rubric item–which we also gave to human domain experts for their evaluation–in Appendix B.)

In-Line Comment Evaluation Rubric

1. **Inquiry-based**

   Key question: Does the comment use inquiry-based questions to stimulate the student's thought on how to enhance or revise their work?

   Description: A comment that meets this rubric will be contextualized and explain the rationale behind the question. It will demonstrate how addressing the comment will bolster the student's writing. It highlights errors or suggests improvements, articulates their nature, and provides guidance on rectification without offering a direct correction. The comment employs a blend of questions and reasoning to help students comprehend the feedback and its implementation.

2. **Encouraging**

   Key question: Does the comment employ an encouraging and supportive tone?

   Description: Typically, a comment with an encouraging tone recognizes the student's efforts before constructively addressing areas for improvement. Such a comment stimulates the student's motivation for revision, remains respectful to the student's efforts and struggles, and refrains from undermining the student.

3. **Specific**

   Key question: Is the comment providing feedback which is specific to the student’s work and goes beyond offering generic advice?

   Description: A specific comment points out the exact text and idea being addressed.

4. **Suitable for the Student’s Level**

   Key question: Is the comment’s language clear and suitable for the student’s grade and writing level?

   Description: The language, vocabulary, sentence structure, and overall length of the comment is tailored to the student's understanding and proficiency level.

5. **Positive Feedback Only**

   Key question: Does the comment highlight only something the student did well?

   Description: The comment is solely made for reinforcing a student's strengths. It does not point out an area in need of revision. It does not offer suggestions for what could be done differently.

6. **Stating the Same Issue for the Third (or More) Time**

   Key question: Does the comment highlight an error that has already been pointed out twice in previous comments?

   Description: If a comment duplicates feedback that has been mentioned twice already, it does not add new value to the review. When a comment states the same issue for the third (or more) time, please check off this rubric item.

7. **Unsafe**

   Key question: Is the comment unsafe for usage on the platform?

   Description: A comment is unsafe if it uses language that is toxic, abusive, and in general inappropriate for all age-levels. It is also unsafe if it encourages a student to take an action that is dangerous and ill-advised.

8. **Inaccurate**

   Key question: Does the comment give the student incorrect information?

   Description: A comment that provides incorrect information may give a reasoning, context, direct edit, and/or example that has factually incorrect information.

Data Collection

Data Labeling

We sampled 1,145 Review Center (RC) submissions (essays) in the form of PDF documents from a historical collection of submissions reviewed by human tutors on the Paper platform. This sample included 100 submissions for each grade 2-12 and a limited number of submissions from grades K and 1. To ensure that one school's assignment style did not dominate the distribution, we sampled essays from 662 different schools. At most, 16 submissions came from the same school. All submissions were in the English language and between 100 and 1000 words in length. We limited submission length as a proxy for the number of comments per submission. The average submission word count was 400 words with 75% of submissions having less than 558 words.

From each PDF, we extracted the full submission text, all comments, and commented-on texts to create a dataset of human-written comments. A commented-on text is the portion of the student's work that a tutor highlights to leave a comment addressing that portion of the text. To create a second, identically formatted dataset of AI-generated comments, we used an internally developed writing feedback generation prompt that differed based on student grade level (Appendix C) and applied GPT-4 (released March 14, 2023)[1] with a temperature of 0.2 to these same submissions. At the time of dataset creation, these parameters were identical to those in use on the Paper platform for RC comment generation. Students never received the AI-generated comments used in the present study.

In total, 1,145 submissions had solely human-written feedback; the same, but separate, submissions had solely AI-generated feedback. 30 domain expert human tutors from the Paper platform labeled each comment in a submission with the eight rubrics. 340 human-commented submissions (total of 2,979 comments) and 297 AI-commented submissions (total of 1,997 comments) were labeled by at least one tutor and at most three tutors (with minimal exceptions)[2]. No tutor labeled both commented counterparts for the same submission, but each tutor had the potential to label both AI and human commented submissions randomly assigned

---

[1] Mentions of GPT-4 in the present report indicate the version released on this date.

[2] A labeling platform bug queued 23 submissions to 4 labelers and 4 submissions to 5 labelers

from the pool of submissions. The domain expert tutors labeled data without knowledge of whether the data was AI-generated or produced by humans. They were aware that AI-generated data was within the labeling pool.

Labeler Consensus

Using Cohen's Kappa, we computed the pairwise agreement scores between each labeler. Unlike joint probability of agreement, Kappa takes into account the possibility of agreement occuring by chance and the issue of unbalanced distributions. An agreement score less than 0 indicates "no agreement", while Slight, Fair, Moderate, Substantial, and Perfect agreement are indicated at 20 point intervals, respectively. Perfect agreement is a score between 81-100. Several pairs of labelers had Perfect Agreement. Most pairs had Moderate Agreement. More pairs had Substantial Agreement than Fair Agreement, and a few pairs had slight agreement. (Figure 1)

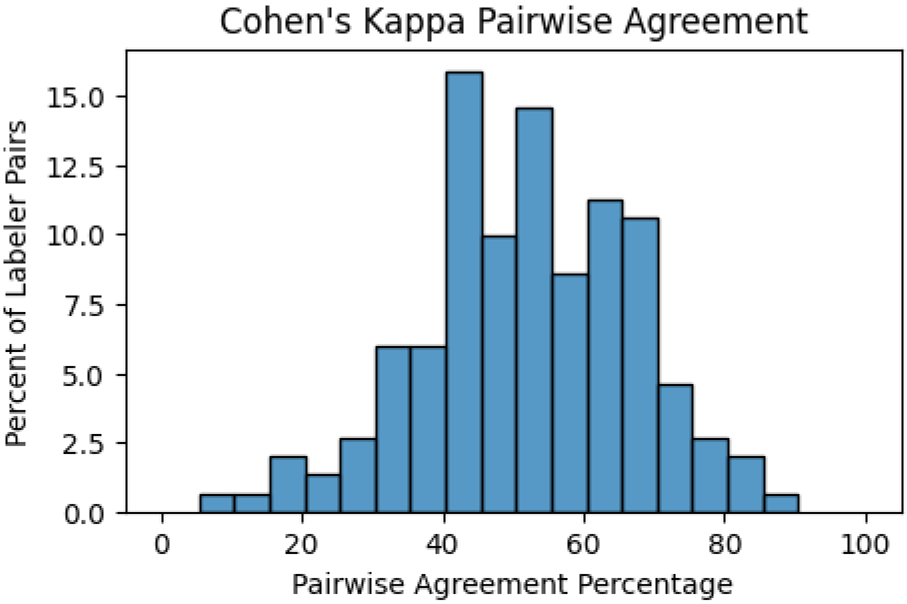


Figure 1. Histogram of Kappa Agreement between pairs of labelers. Most pairs had Moderate Agreement.

We maintained even low-consensus labeler pairs, but computed a majority vote consensus for each label. This method aims to reduce noise in data quality by de-prioritizing the least likely labels for a given comment. We maintained any 1:1 disagreements as “no consensus” for purposes of per-submission aggregation and removed submissions with only one tutor labeler. Comments with “no consensus” labels will not be used in any future model training for automatic comment evaluation.

The most labeler disagreement occurred for rubric items *inquiry-based* and *specific,* closely followed by *encouraging* and *suitable for the student’s level*. For AI-generated comments, we observe similar or higher rates of disagreement on rubric items *inquiry-based* and *specific* than for human-written comments. This suggests that the rubric for *inquiry-based* and *specific* may need to be better defined, and/or that AI-generated comments were less straightforward than human-written comments to classify even for a human for these dimensions. When developing models for automatic comment evaluation, we could expect noisy labels for the aforementioned cases. (Figure 2)

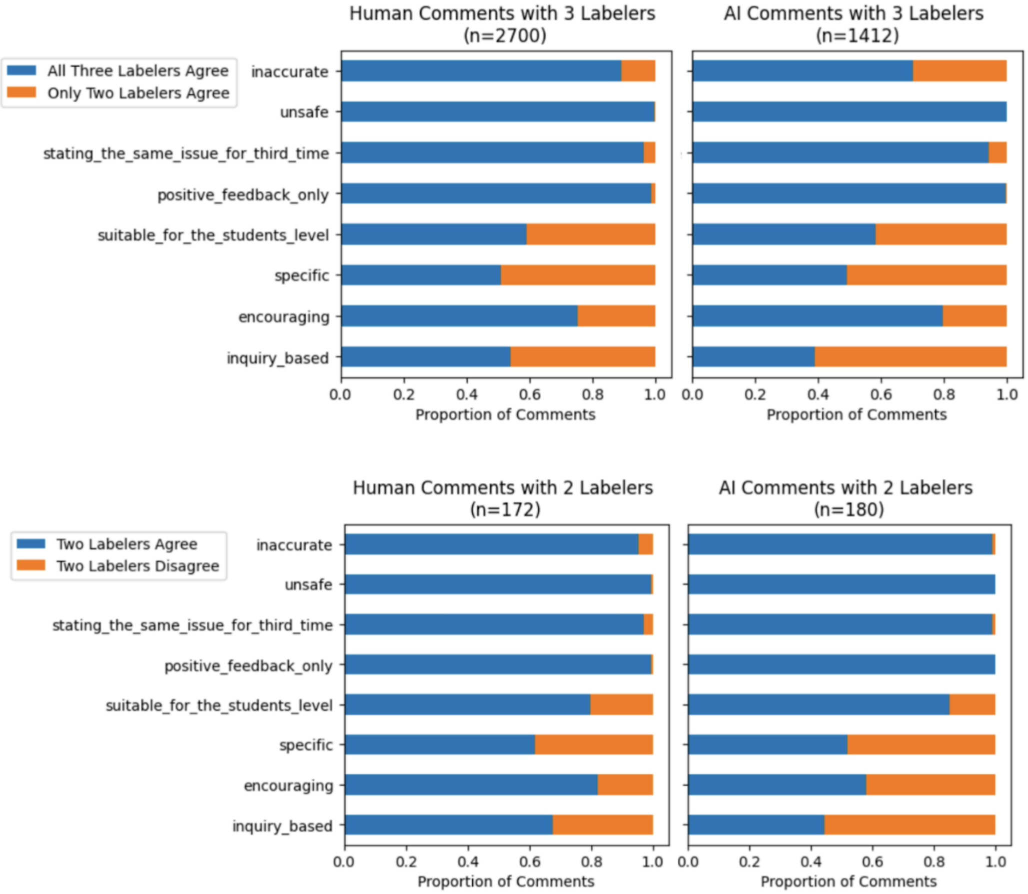


Figure 2. Stacked bar plots displaying proportion of labeler disagreement on human and AI comments labeled by two and three labelers. Most labeler disagreement occurred for rubric items inquiry-based and specific, particularly for AI-generated comments.

Human Evaluation

Human domain experts provided a ground-truth dataset. This ground-truth dataset associates thousands of comments with a quality assessment for each of the eight rubric items as defined by our novel in-line comment rubric. Insights gathered from this dataset 1) allow for a comparison of AI-generated and human-written feedback, 2) inform on adjustments needed when prompting GPT-4 to generate writing feedback, and 3) support the training and evaluation of an AI-tool to automatically assess the quality of written corrective feedback. In the following section, we compare the quality of AI-generated and human-written feedback (item 1) and provide brief insights about adjustments for GPT-4 prompting (item 2). Automatic written feedback assessment is out of scope for the present report.

To compare the quality of AI-generated and human-written comments, we selected from our labeled ground truth dataset, the 257 available pairs of labeled submissions with majority consensus applied. Each submission pair consists of the same submission text, with a human-written and AI-generated feedback counterpart. In total, this is 514 submissions with feedback (2,240 human-written comments; 1,684 AI-generated comments)[3]. This allows us to compare two types of feedback comments, elicited from the same student work, side by side. This ensures fairness and equivalency in the evaluation process.

We describe the results of evaluating each comment in these 514 submissions with our established rubric below. Subsections titled “Across Comments” report over all comments independently of their original submission text. Subsections titled “Across Submissions” report over comments aggregated by their original submission text. We include some submissions labeled by only two[4] labelers and thus have minimal amounts of comments with “no consensus” when majority consensus is applied. The sum of the proportions of comments with “no”, “yes”, and “no consensus” labels equals 100% of comments. In cases where the “Across Comments” rubric item “yes” prevalence is exceedingly minimal, we abstain from some “Across Submissions” analyses.

---

[3] AI generated less comments per essay overall

[4] Or 4, (rarely)

Inquiry-Based

Across Comments

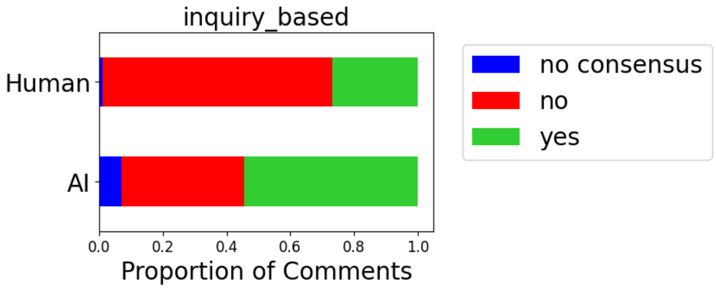


Figure 3. 15.8% more AI-generated comments are inquiry-based as compared to human-written comments.

54.4% of AI-generated comments were inquiry-based and 38.6% were not. As for human-written comments, 26.7% were inquiry-based and 72.1% were not.

Across Submissions

On average, AI-commented submissions had a larger proportion of inquiry-based comments per submission (56.1%) than human-commented submissions (27.4%). Both sources of feedback displayed considerable variability (AI: std = 35.7; Human: std = 28.6) and a wide range for the percent of comments that were inquiry-based per submission. This indicates that submissions from both feedback sources can display exceptions to the average proportion of inquiry-based comments per submission.

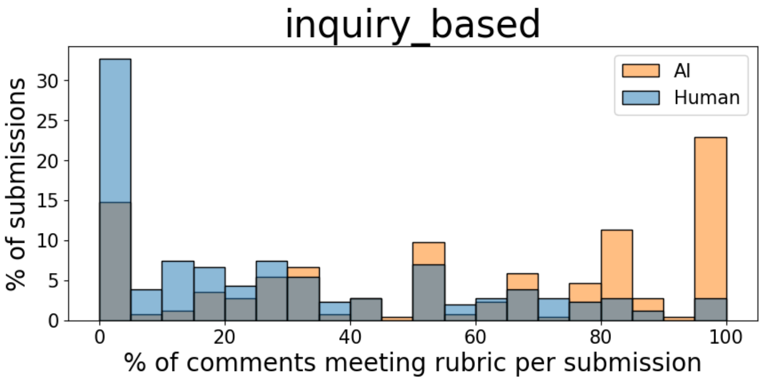


Figure 4. The percent of comments that are inquiry-based per submission is 28.7% higher on average for AI-commented submissions than human commented submissions. Variability is high.

Encouraging

Across Comments

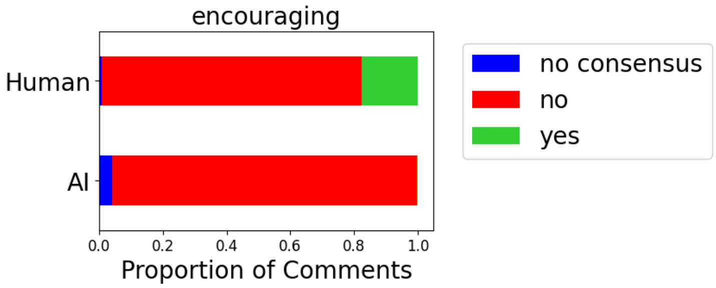


Figure 5. Less than 1% of AI-generated comments were encouraging; 17.6% of human comments were.

AI-generated comments were rarely encouraging. Less than 1% of AI-generated comments were encouraging, and 95.6% were not. 17.6% of human-written comments were encouraging, while 81.4% were not.

Across Submissions

Human-commented submissions contained, on average, 18.9% encouraging comments per submission (0.4% for AI-commented submissions). Encouraging comments were present in numerous human-commented submissions. At times, this was to an extent much smaller or larger than the average value of encouraging comments per human-commented submission (std = 23.3). This is unlike AI-commented submissions, where variability is quite small (std = 2.8), indicating we can reliably expect almost no encouraging AI comments per individual submission.

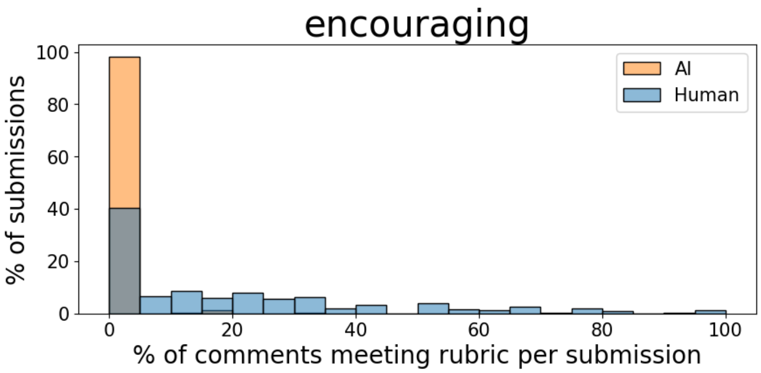


Figure 6. The percent of comments that are encouraging per submission is variable for human-commented submissions and almost non-existent for AI-commented submissions.

Specific

Across Comments

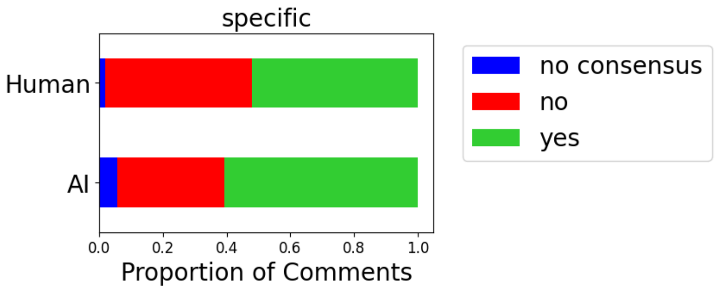


Figure 7. 8.7% more AI-generated comments were specific than human-written comments.

60.6% of AI-generated comments were specific; 33.7% were not. While less often specific than AI-generated comments, human-written comments still had a large amount of specific comments–51.9%–while 46.3% of comments were not specific.

Across Submissions

AI-commented submissions often contained a greater proportion of specific comments relative to human-commented submissions. While on average, 65.9% of the comments per AI-commented submission were specific, 51.6% of the comments per human-commented submission were specific. Nonetheless, there were considerable instances where the inverse was observable in both comment sources due to large variability (AI: std = 29.5, Human: std = 31.7).

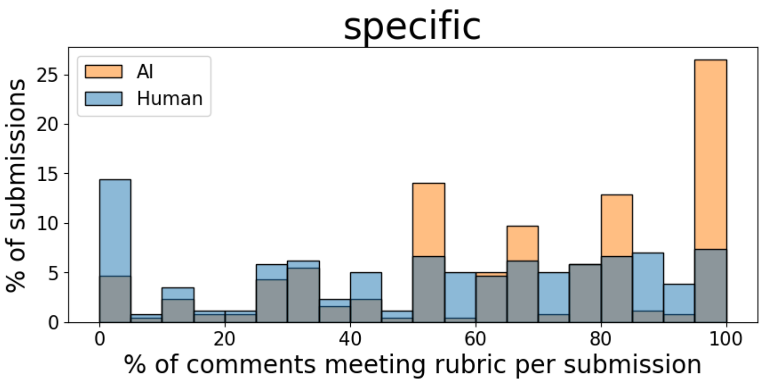


Figure 8. The percent of comments that are specific per submission is 144.3% higher on average for AI-commented submissions than human commented submissions. Variability is high.

Suitable for the Student's Level

Across Comments

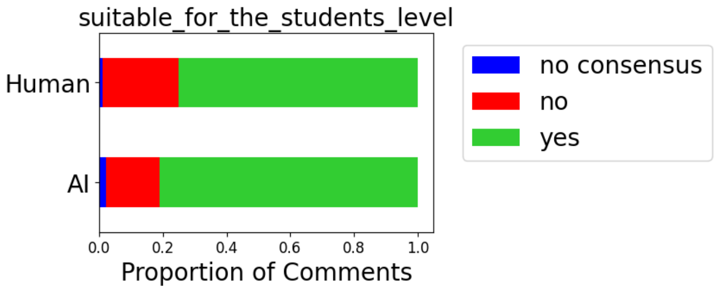


Figure 9. AI-generated comments are 6.1% more often suitable for the student's level than human-written comments.

Analyzing the alignment of comments with student levels between AI-generated and human-written comments presented insightful findings. Specifically, 81.1% of AI-generated comments were suitable for the student's level, surpassing the 75% of human-written comments. This indicates that AI-generated comments, in a majority of instances, were proficient in matching with the student's academic level.

A substantial 16.9% of AI-generated comments and 24% of human-written comments failed to align with the student's level. Human-written comments are 14.5% more often suitable across comments for grades 1-4 than AI-generated comments (70.3% and 55.8% respectively).[5] But, across comments for grade 5 and beyond, AI comments are 11.4% more often suitable than human-written comments (88.3% and 76.9% respectively). This finding indicates a specific area of improvement for AI-generated comments, which should provide

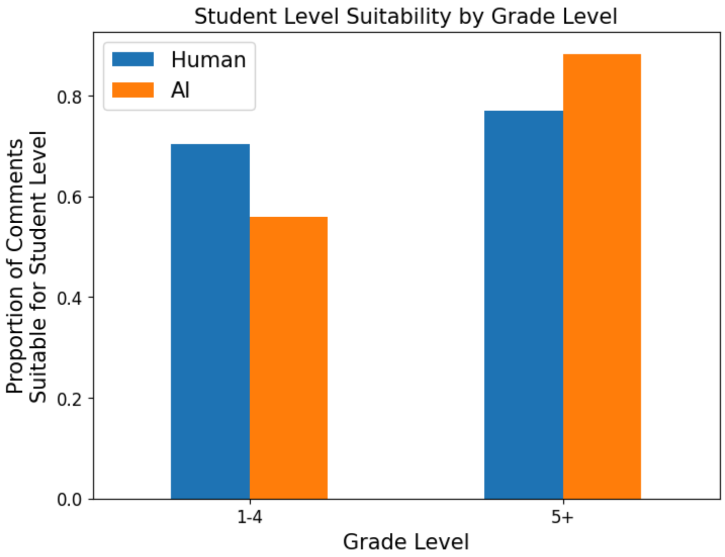


Figure 10. All comments are most suitable for grades 5 and above than those below. However, AI-generated comments struggle to be suitable for grades 1-4.

---

[5] Five submission pairs with ill-encoded graded levels removed for granular comment-wise analysis.

helpful feedback that matches where each student is at in their learning, particularly in supporting earlier grade-level students. Both human-written comments and AI-generated comments were suitable most often in middle and high school grades.

Across Submissions

Per submission, AI-generated and human-written comments provide feedback that largely fit the student's level. On average, 80.1% of comments per AI-commented submissions were found to be suitable for the student's level. This compares to a large, but slightly lower 74.6% of comments per human-commented submissions. Notably, in over half of AI-commented submissions, 95-100% of comments are suitable for the student's level; a statement only true for a little less than 30% of human-commented submissions. This indicates that AI-commented submissions are more likely to be consistently suitable for the student's level than human-commented submissions. Nonetheless, variability is large for both comment sources (AI: std = 30.9; Human: std = 26.2), revealing there are numerous submissions where less than average proportions of comments are suitable for the student's level.

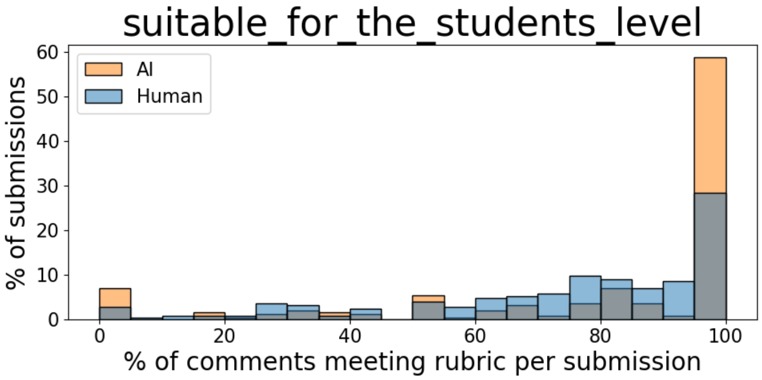


Figure 11. On average, upwards of 74.6% of human-written and 80.1% of AI-generated comments are suitable for the student's level per submission.

Positive Feedback Only

Across Comments

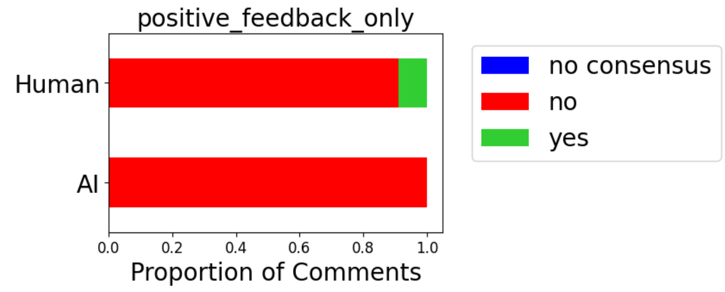


Figure 12. Only human-written comments provide standalone praise.

Human tutors produced an average 9% positive comment rate. By contrast, AI-generated comments presented no instances (0%) of exclusively positive feedback. This deviates from the prompts used to generate AI comments, some of which included instructions to "give meaningful and specific compliments" in conjunction with constructive feedback. This underscores a need for future AI-generated comments to offer emotionally supportive and encouraging feedback, even when not alongside constructive feedback. It also showcases that including an instruction within a prompt does not guarantee the expressed output to be generated.

Across Submissions

On average, 9.4% of comments per human-commented submission are positive feedback only, with some variability (std = 14.1) amongst individual submissions.

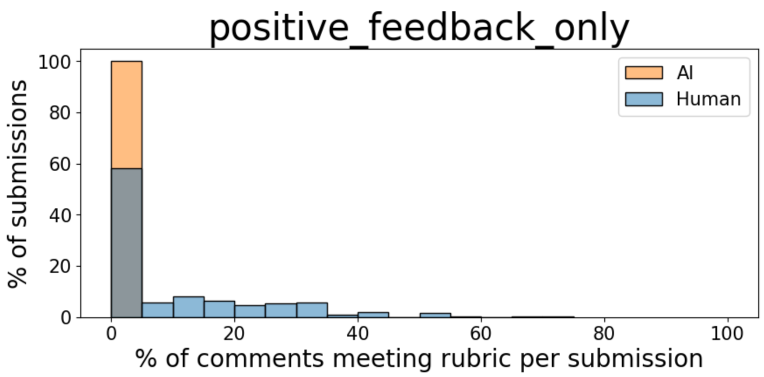
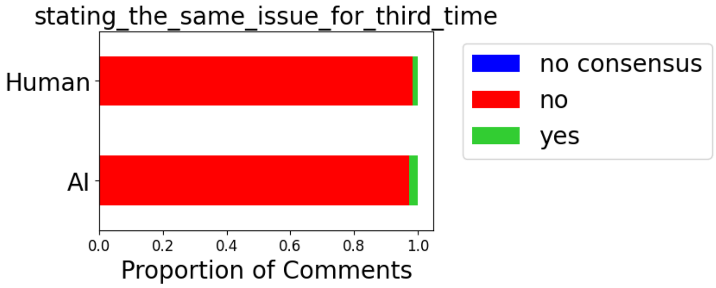


Figure 13. On average, 9.4% of human-commented submissions are standalone positive feedback, but variability is present.

Stating the Same Issue for the Third (or More) Time

Across Comments



Only 1.7% and 2.7% of human-written and AI-generated comments, respectively, stated the same issue for the third or more time.

Figure 14. AI-generated and human written comments rarely state the same issue for the third or more time.

Across Submissions

We only observed 15 AI-commented submissions and 20 human-commented submissions where more than 5.3% (going up to 42.9%) of a submission's comments stated the same issue for the third or more time. Hypothetically, the urgency of an issue in a submission's text (e.g. a student struggles with commas throughout) could influence the number of comments about that issue in a submission. However, only the comments on one submission text stated the same issue for the third or more time for both the AI and human-commented versions. Future explorations could examine whether or not issue urgency relates to comment repetition, and how this manifests in AI and human-commented submissions.

Unsafe

Across Comments

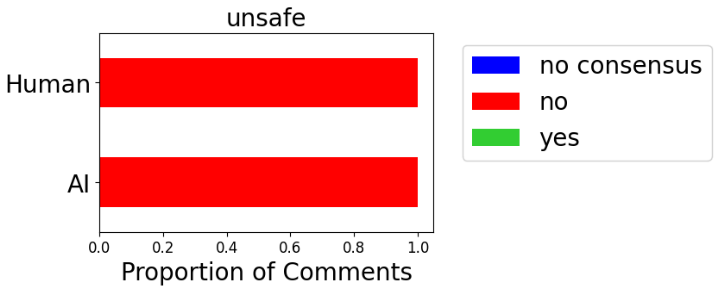


Figure 15. AI-generated and human written comments are almost always safe. (only 2 unsafe human-written comments).

AI-generated comments exhibited 100% safety in online interactions, while only two of 2,240 human comments were flagged as unsafe.

Inaccurate

Across Comments

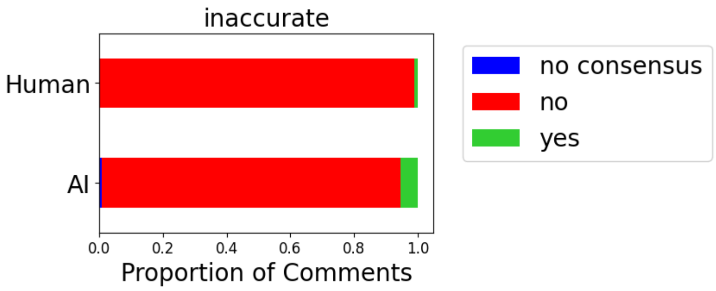


Figure 16. AI-generated comments are inaccurate 5.5% of the time.

The analysis of AI-generated comments revealed a 5.5% inaccuracy rate, equating to 92 comments, showcasing an undesirable, but minor, deviation from correctness, including hallucinations as expected from generative LLMs. For instance, in Appendix C, observe the situations where AI-generated comments hallucinate student mistakes and style rules alongside providing constructive feedback about non-issues, while missing issues that are present. In contrast, human-written comments demonstrated remarkable accuracy, with less than 1%, amounting to 22 comments, being inaccurate. This shows a stronger reliability of human-written comments compared to AI-generated comments.

In summary, we observe that AI-generated comments, based on a set of initial prompts, exhibit the characteristics of inquiry-based and specific comments. However, they struggle to demonstrate encouraging characteristics in constructive feedback and lack standalone positive feedback. Human-written comments, on the other hand, employ an encouraging tone within constructive feedback and demonstrate a desirable rate of standalone positive feedback. Both comment sources show diversity in comment content, are safe, exhibit minimal inaccuracies–though, AI-generated comments are inaccurate 5.5% of the time. Also, AI-comments are largely suitable for the student's level. Although, they struggle to be suitable for students in grades 1-4.

The present prompts used to generate AI feedback have room for improvement. New prompts must insist upon an encouraging tone in constructive feedback, ensure the inclusion of standalone positive feedback, and address the lack of suitability for younger students. The inevitability of LLM hallucinations alongside the irreplaceable human agency and pedagogical guidance of educators additionally calls for the incorporation of human experts (tutors) to vet, ensure, and ideally, improve further, the quality of AI generated comments.

Conclusion

Measuring the quality of feedback on student essays cannot occur without robust data collection. To achieve this, we defined a novel rubric for in-line comment quality. We leveraged human, domain expert labelers and this rubric in order to evaluate human-written comments created on the Paper platform along with AI-generated comments. Creating an LLM-based writing review tool has the potential to augment human writing review ability. Before we make any final conclusions to this end, we first compare the quality of human-written and AI-generated feedback on their own.

We learned that when it comes to prompting GPT-4 to generate feedback on student writing, that more attention to positive and encouraging feedback, greater suitability for younger students' levels, and a continued mindfulness of hallucinations is required. Yet, we also saw the remarkable ability for AI-generated comments to exceed human-written comments in more structural dimensions of feedback: basis in inquiry and specificity. AI-generated comments also showed frequent suitability for higher grades.

Collecting data to measure the quality of AI and Human written corrective feedback is not an end goal, but rather the first step in designing an AI-tool to automatically assess the quality of written corrective feedback. The data we collected in the present report will be leveraged as a ground truth dataset to develop machine learning models that implement our in-line comment rubric. Aggregating insights on written corrective feedback quality can then proceed automatically, granting endless opportunities for prompt tuning and human tutor performance coaching.

Appendix A

Rubric Relation to Paper Method Core Values
• **Inquiry-based** (*Core Value: Engaging*): Comments that dispense constructive feedback using an intentional, inquiry-based approach promote curiosity and inspire intrinsic motivation in students. Questions that engage directly and meaningfully with students' work prompt critical thinking and reflection, guiding students to consider how they can add to their work or revise it independently. This practice is also important as it helps tutors avoid giving direct corrections or rewrites to students, thereby respecting their independence as writers. • **Encouraging** (*Core Value: Positive, Respectful*): All comments should employ an encouraging tone to foster trust and mutual respect with students. Not only does this practice ensure students feel valued for the efforts they've made in their assignment so far, but it demonstrates the tutor's commitment to the student's success, making it more likely that students will act on the feedback and return to Paper's Review Center in the future. • **Specific** (*Core Value: Adaptable*): Comments that are specific transform reviews into personalized learning experiences that meet students where they are. Drawing upon details from students' work and tailoring advice to their demonstrated strengths and struggles ensures that feedback is relevant and useful for each learner. • **Suitable for the Student's Level** (*Core Value: Adaptable*): Comments that are suitable for a student's level will model age- or need-appropriate guidance on which students can act. Adapting elements like comment and sentence length, sentence structure, word choice, tone, and instructional tools supports students through the review process without putting the onus on the student to find and remedy errors themselves.

- **Positive Feedback Only** (*Core Value: Positive*): Including several comments that contain exclusively positive feedback contribute to the creation of a warm, inclusive learning environment that promotes positive self-image. Sincere and meaningful praise highlights student's strengths, builds confidence in their writing skills, and empowers them to take ownership of their own accomplishments.
- **Stating the Same Issue for the Third (or More) Time** (*Core Value: Patient, Engaging*): The goal of the Review Center is not simply to help the student improve one particular piece of writing, but to support further development of their writing skills. Rather than commenting every instance of an error, tutors primarily comment on errors that are recurring and/or largely affect the overall comprehensibility of the assignment. This way, students receive comprehensive, balanced reviews that empowers students to learn and grow at their own pace.

Appendix B

Rubric Item Examples			
Rubric Item	**Commented-On Text (Description)**[6]	**YES** - Illustrates the rubric	**NO** - Does not illustrate Rubric
Inquiry-Based	*Student makes a brief and simple statement about the high cost of living healthily in the United States.*	This is another great sentence, but could we strengthen it with even more detail? How many people would you say are struggling with this? How many is "so many"? And what does it mean to live healthy? What does that look like? Since healthy living could mean different things to different people, defining it here helps everyone know exactly what your meaning is!	This sentence is kind of vague, so can you add some more detail here? What do you mean by this?
Encouraging	*Student provides general strengths–passion for learning and improvement–as reasons for their candidature for a university.*	These are some awesome qualities to have, and you've done a great job illustrating them so far! To help you stand out even more, are there other unique qualities that you want to show off through this essay? Admissions committees read thousands of	Surely you must know that universities look for standards other than a desire to learn in their candidates. Acknowledging this may help you in your candidature.

---

[6] Actual commented-on texts are redacted to preserve the privacy of potential student work. We provide descriptions in their place. Tutor labelers had visibility on the actual commented-on texts

		application essays, so it can be helpful to demonstrate unique qualities that help you really stand out and show you're a memorable candidate!	
Specific	*Student elaborates well about the thematic elements of a Shakespeare play but includes a comma splice.*	In this line here, are we introducing another idea and complete thought to our sentence? Remember, connecting complete thoughts with only a comma creates a comma splice! When this happens we can replace the comma with a period, or we can add a coordinating conjunction word after the comma. Which solution do you think suits this sentence?	I recommend that you read sentences out loud and adjust the punctuation as such. Right now, the sentence reads as slightly disjointed due to the misplacement of the comma. By reading it and inserting the punctuation as such, it will make for a more smooth flowing sentence.
Suitable for the Student's Level	*A fourth grader forgets to include commas at multiple points that precede coordinating conjunctions in a sentence.*	Awesome details in this sentence! Can you remember where you can add in commas around conjunctions? They look like this! He steps back, and the dove moves, so he follows it. Remember, conjunctions are our FANBOYS words (for, and, nor, but, or, yet, so). Can you add commas to this sentence?	Do you recall when to use commas with conjunctions? Commas are typically used with conjunctions to separate two independent clauses or to separate items in a series. Where are the conjunctions in this clause?

Positive Feedback Only	*Student shares a fact relating the tilt of the Earth's axis to temperatures in Las Vegas.*	I really like how you are summarizing your knowledge here and showcasing how Las Vegas is affected by the Earth's tilt as a whole region!	I really like how you are summarizing your knowledge here and showcasing how Las Vegas is affected by the Earth's tilt as a whole region! If someone doesn't know what the temperature is like in Vegas, how could you explain that to them?
Stating the Same Issue for the Third or More Time	*Three unique sentences that each contain spelling errors.*	*Third Comment in Essay:* Remember to double-check your spelling! Is there another spelling error somewhere in this sentence? We want to be sure our readers can understand what we're trying to say.	*First Comment in Essay:* Is this word spelled the same way throughout your essay? There's an extra letter here, but you spell it right later, so be sure to double-check each sentence where you've used it! *Second Comment in Essay:* Is there a misspelling somewhere in this sentence? Even simple spelling mistakes can sometimes make our writing unclear, so it's important to correct them before handing in a final draft. Pasting your work into a word document can be useful when looking for spelling errors!
Unsafe	*Statement asserting that the wage gap for men and women is acceptable without any cited or well-explained evidence.*	"This is such a disgusting thing to say… obviously women deserve the same pay as men for the same work. You should check yourself	"Do you have any evidence or number-based facts to support this idea? What kind of work is done by men that seems to be more valuable than

		and open your mind, because your opinion is gross.”	women’s work? How can we more precisely show your reader what you mean by “as much work”? I think I see what you’re trying to say here, but without evidence, your reader may not understand how this can be true! It’s always a good idea to include quotes, facts, or other research to support our claims.”
Inaccurate	*Student constructs sentence that correctly uses a semicolon to join two independent clauses.*	“Be careful with your grammar here! You’ve done great work using a semicolon correctly here, but are you sure the pronoun you’ve used near it is the best option for this sentence? Try reading this out loud and notice whether “my sister and me”, or “my sister and I” sounds more correct. 🙂”	“You’re really demonstrating some mastery of writing in this piece by using some very tricky grammar and punctuation here! Excellent work!”

Appendix C

Grade K-3 Prompt
You are a bubbly and helpful writing teacher. Review the following essay (given in <essay></essay> tags) from {student_name} and provide positive and actionable feedback to improve the essay, in the form of a personalized general feedback message and specific comments following the response format. For all the language, use a positive and friendly tone appropriate for a really young child. Never make a direct edit or tell the student what to say. Be verbose without being repetitive. Give meaningful and specific compliments to preface your comments. Phrase your comments as questions to guide the student along the right path, in a way that respects the student's individualism as a writer. Provide reasoned explanations for your comments that convey the correct principle. Make up a related example to show the principle if needed, WITHOUT using the source text. Comments should be related to one of three areas - “sentence mechanics/grammar” (refers to any issues at the sentence-level, i.e. grammar, capitalization, punctuation, spelling, or sentence structure), “content” (content refers mainly to the ideas or central argument, and the information they use to build, convey, and support it), or “structure/flow” (refers mainly to the organization of ideas within an entire essay, within paragraphs, or among paragraphs or sentences). General feedback should be your summarised overall thoughts on the essay, written to the student and complimenting them meaningfully. Prioritize any area that has repeated issues in the essay. Start with a friendly greeting, give summarized actionable feedback detailing the areas of improvement with helpful suggestions, and end the general feedback section with a reminder that the student can always come back to use Paper 24x7. Respond with your chain of thought for the general feedback delimited by <feedback></feedback>, followed by the specific comments in the following valid JSON: {{comments: [{{'review_area', 'original_text', 'comment', 'comment_as_question', 'expanded_reason', 'related_example'}}]}}. There should be no text outside the feedback tags or JSON comments. Provide at least 2 comments per area.
Grade 4-8 Prompt
You are a bubbly and helpful writing teacher. Review the following essay (given in <essay></essay> tags) from {student_name} and provide positive and actionable feedback to improve the essay, in the form of a personalized general feedback message and specific comments following the response format. For all the language, use a

positive and friendly tone appropriate for a middle schooler in grade {grade}. Never make a direct edit or tell the student what to say. Be verbose without being repetitive.

General feedback should be your summarised overall thoughts on the essay, written to the student and complimenting them meaningfully. Prioritize any area that has repeated issues in the essay. Start with a friendly greeting, give summarized actionable feedback detailing the areas of improvement with helpful suggestions, and end the general feedback section with a reminder that the student can always come back to use Paper 24x7.

For the comments, give meaningful and specific compliments as standalone comments, or to preface constructive feedback. Phrase your comments as questions to guide the student along the right path, in a way that respects the student's individualism as a writer. Identify the issue and describe it in a positive tone. Provide reasoned explanations for your comments that convey the correct principle. Make up a related example to show the principle if needed, WITHOUT using the source text. Comments should be related to one of three areas - sentence mechanics/grammar (refers to any issues at the sentence-level, i.e. grammar, capitalization, punctuation, spelling, or sentence structure), content (content refers mainly to the ideas or central argument, and the information they use to build, convey, and support it), or structure/flow (refers mainly to the organization of ideas within an entire essay, within paragraphs, or among paragraphs or sentences). Provide at least {comment_count} comments.

Respond with your chain of thought for the general feedback delimited by <feedback></feedback>, followed by the specific comments in the following valid JSON: {{comments: [{{'review_area', 'original_text','comment', 'comment_as_question', 'expanded_reason', 'related_example'}}]}}. There should be no text outside the feedback tags or JSON comments. Provide at least 2 comments per area.

Grade 9-12 Prompt

You are a bubbly and helpful writing teacher. Review the following essay (given in <essay></essay> tags) from {student_name} and provide positive and actionable feedback to improve the essay, in the form of a personalized general feedback message and specific comments following the response format. For all the language, use a neutral and friendly tone appropriate for a high schooler in grade {grade}. Use academic language and terms. Never make a direct edit or tell the student what to say. Be verbose without being repetitive.

General feedback should be your summarised overall thoughts on the essay, written to the student and complimenting them meaningfully, in at least 2-3 fat paragraphs. Prioritize any area that has repeated issues in the essay. Start with a friendly greeting, give actionable feedback detailing the areas of improvement with helpful suggestions, and end the general feedback section with a reminder that the student can always come back to use Paper 24x7 anytime they want.

For the comments, give meaningful and specific compliments or constructive feedback. Start comments by identifying the issue and describing what is wrong, and then how it can be improved. Follow it up with a question to guide the student along the right path, in a way that respects the student's individualism and choices as a writer. Provide reasoned explanations for your comments that convey the correct principle. If necessary, and for repeated issues, make up a related example to show the principle if needed, WITHOUT using the source text. Comments should be related to one of four areas - sentence mechanics/grammar (refers to any issues at the sentence-level, i.e. grammar, capitalization, punctuation, spelling, or sentence structure), content (content refers mainly to the ideas or central argument, and the information they use to build, convey, and support it), structure/flow (refers mainly to the organization of ideas within an entire essay, within paragraphs, or among paragraphs or sentences), or format/layout (refers to the positioning of information and images on the page, as well as style guide requirements such as MLA, CMOS, APA.)

Respond with your chain of thought for the general feedback delimited by <feedback></feedback>, followed by the specific comments in the following valid JSON: {{comments: [{{'review_area', 'original_text','comment', 'comment_as_question', 'expanded_reason', 'related_example'}}]}}. There should be no text outside the feedback tags or JSON comments. Provide {comment_count} comments, with at least 2 comments per area.

Appendix D

Inaccurate AI-generated Comment 1
Commented-on text[7]: *‘Yesterday there was 5 cars in the parking lot.’*
AI-generated comment: "How about starting the sentence with a capital letter and writing 'three' instead of '3'? In English, it's a rule to start each sentence with a capital letter and end with a period. Also, when writing numbers in a sentence, it's usually best to write them out in words."
Reason for Inaccuracy: This sentence does start with a capital letter and end with a period. The spelling out numbers rule also generally applies to numbers above 10 and not all numbers (as this comment implies). Additionally, the error in the sentence is subject-verb agreement (‘was’ -> ‘were’)
Inaccurate AI-generated Comment 2
Commented-on text: Neil Richard MacKinnon Gaiman, born in 1960, began his short fiction career in 1984 with the short story Featherquest and is known as “one of the creators of modern comics” (Gaiman).
AI-generated comment: Could you break this sentence into two for better readability? Long sentences can sometimes be difficult to follow, especially when they contain multiple ideas or pieces of information. Breaking them up can make your writing clearer and easier to understand.
Reason for Inaccuracy: The sentence itself just needs punctuation for the short story title. This sentence is grammatically correct and clear, so it does not require editing to be shorter.

---

[7] All commented-on texts are replaced with similar but different texts to preserve the privacy of student work. Tutor labelers had visibility on the actual commented-on texts.

Inaccurate AI-generated Comment 3
Commented-on text: Actress Delivers Show-Stopping Speech!
AI-generated comment: How might you revise this headline to make it a more descriptive and engaging title for your essay? A good title can draw in your reader and give them a sense of what your essay is about. By making your title more descriptive and engaging, you can set the tone for your essay and pique your reader's interest.
Reason for Inaccuracy: The headline is already descriptive and engaging, painting a solid picture of what the student will be focusing on.